Models and Estimation for Phylogenetic Trees

Faisal Ababneh

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

School of Mathematics and Statistics
The University of Sydney

May 2006

This thesis contains no material which has been accepted for the award of any other degree. To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor John Robinson for his advice, incredible support, insightful comment and encouragement. I also would like to thank my associate supervisor Dr. Lars Jermiin for his help, understanding, encouragement and invaluable discussion.

I would like to also acknowledge the Al-Hussain Bin Talal University for the financial support during my Ph.D candidature.

I should never forget to thank all of my friends and teachers who help me to conduct this thesis.

Finally, I dedicate this thesis to my mother, the memory of my father, to my brothers, sisters and all of my nephews and nieces for their support, love, understanding. Without them, I wouldn't have had the courage to complete this thesis.

Abstract

In this thesis, we consider Markov models for matched sequences. Define $f_{ij}(t) = P(X(t) = i, Y(t) = j|X(0) = Y(0))$, where f_{ij} is the joint probability that, for a given site, the first and second sequences have the values i and j at a given site, given that they were the same at time 0. This can generalized to several sequences. The sequences (taxa) are then arranged in an evolutionary tree (phylogenetic tree) depicting how taxa diverge from their common ancestors. We develop tests and estimation methods for the parameters of different models.

Standard phylogenetic methods assume stationarity, homogeneity and reversibility for the Markov processes, and often impose further restrictions on the parameters. The parameters in these cases are estimated using many popular packages, including PHYLIP and PAUP*. We describe a new and more general method for calculating the joint probability distribution under stationary and homogeneous models for the more general models with some weakening of the stationarity and homogeneity assumptions. We describe the method for a two edged tree and then extend it to the case for a K tipped tree. We discuss the case of a five edged tree for a set of bacterial sequences for which stationarity and homogeneity are not present. This data set is very similar to that of Galtier and Gouy (1995), and the search for methods appropriate for its analysis has provided the raison d'etre for this work. The extension we propose is to allow non-stationarity, so that from the root of the tree we permit different Markov processes to operate along different descendant lineages; furthermore, we permit non-homogeneous Markov processes to operate across the tree. We obtain methods that give the joint distribution of the leaves of a K-edged tree under the new models.

Next we derive a number of methods for simulating the evolution of DNA for these general models using a multinomial distribution, direct simulation via Markov processes and an approximation approach. We derive a number of tests, which can be used to check the assumptions of stationarity and homogeneity for the phylogenetic data. We review the distance methods and develop functions, which use the paralinear distance method to find the topology of the tree to be used in the estimation method.

Finally we describe the maximum-likelihood approach and use it to estimate the parameters of different models. We discuss and apply our methods of testing and estimation on simulated data models, which have the same structure as our models, such that the stationarity and homogeneity are not represented in these cases. Then we apply our methods of testing and estimation to real data.

Contents