SYNOPSIS

OF THE THESIS ENTITLED

BAYESIAN METHODS IN GENETICS-
A GRAPH THEORETIC APPROACH

To be submitted to

Kannur University in partial fulfilment of the

DEGREE OF DOCTOR OF PHILOSOPHY IN STATISTICS
UNDER THE FACULTY OF SCIENCE

By

GEETHA ANTONY PULLEN

Supervised by

Dr. M. KUMARAN

RESEARCH CENTRE, DEPARTMENT OF STATISTICS
NEHRU ARTS AND SCIENCE COLLEGE, KANHANGAD, KERALA,
INDIA
MAY  2011
0.1 Introduction

The discovery of the structure of Deoxyribose Nucleic Acid (DNA) by James Watson and Francis Crick stands out as one of the most important scientific findings of the last millennium. Since, DNA is so critical to our existence and since small mutations in one or more genetic sequences can have devastating effects on the affected individual, it is essential that we fully understand how DNA controls life if we want to cure individuals who suffer from a genetic disease. In human genetic research, sophisticated statistical methods are increasingly being used to analyze data. The most widely used one being the Bayesian inference method. A new technique that is becoming increasingly used in genetic studies is the application of an area in mathematics called graph theory to map networks of biological interactions. The methods of graph theory have been applied to build individual interactions into pathway, enabling a more holistic analysis of the cell interactions. Graph Theory was originated by Euler in the eighteenth century, and over the last forty years, influenced by the work of two mathematicians Paul Erdos and Alfred Renyi, who initiated the study of the mathematical properties of random networks, the use of graph theory flourished. The behavior of complex systems emerges from the sum-total activity of many components that inter-
act with each other through pair-wise interactions. These components are reduced to a series of nodes that are connected to each other by links (or edges) and together form a network or graph. Depending on the nature of their interactions, networks can be directed or undirected. Bayesian Networks are gaining an increasing popularity as modeling tools for complex problems involving probabilistic reasoning under uncertainty. A Bayesian Network (BN) is a Graphical model in which the underlying graph contains only directed edges. A dynamic Bayesian network (DBN) is a Bayesian network that represents sequences of variables. These sequences are often time-series (for example, in speech recognition) or sequences of symbols (for example, protein sequences). The hidden Markov model and the Kalman filter can be considered as the most simple DBNs. Imprecise probability is the study of generalizations of the mathematical theory of probability to allow partial or interval specifications. Within the applied use of imprecise probabilities the adoption of a Bayesian position is particularly common. The use of imprecise probability in Bayesian theory is both a recognition of the validity of some of the criticisms of Bayesian theory and an attempt to address them. A mixture of trees is a set of trees along with a probability distribution over them.

0.2 Summary of the Thesis

Monte Carlo methods when applied to samples from Markov chains via the weak law of large numbers for the number of passes through
a recurrent state in an ergodic Markov chain (Feller 1957) are called MCMC methods. MCMC methods are probably about 50 years old, and has been both developed and extensively used in Physics for the last four decades. It has now come to be an all pervading technique in statistical computation, in particular for Bayesian inference, and especially in complex stochastic systems. In this study we made an attempt to highlight the effectiveness of one of the very powerful statistical procedures - the Bayesian inference - for modeling and analysis of various problems in Genetics. In Chapter 2, we applied Bayesian posterior density estimation through MCMC method on a population genetics data.

The states of a Markov chain may be represented by the vertices of a graph, one step transitions between states by directed arcs; if \( p_{ij} \neq 0 \), then vertices \( i \) and \( j \) are joined by directed arc with arrow towards \( j \), the value of \( p_{ij} \) which corresponds to the arc weight, may be indicated in the directed arc. If \( S = \{1, 2, \ldots, m\} \) is the set of vertices corresponding to the state space of the chain and \( A \) is the set of directed arcs between these vertices, then the graph \( D = \{S, A\} \) is the directed graph or digraph or transition graph of the chain. A digraph such that its arc weights are positive and the sum of the arc weight of the arcs from each node is unity, is called a stochastic graph; the digraph of a \( MC \) is a stochastic graph. A transition graph is of great aid in visualizing an \( MC \). It is a useful tool in studying the properties of the chain.
Definition 0.2.1. A Bayesian Network for $X = \{X_1, X_2, \ldots, X_n\}$ consists of

1. a network structure - a directed acyclic graph (DAG) where the vertex set is $X$ and,

2. a set of conditional probabilities - $P = \{p(X_i \mid pa(X_i))\}$ - conditional probability distributions of variables $X_i$ given its parents in the graph.

In a Bayesian network (BN), conditional independence properties can be used to simplify the general factorization formula for the joint probability and the result is a factorization that can be expressed by the use of DAGs. In Chapter 3, we applied the algorithm Probability Propagation in Trees Of Clusters (PPTC) on a BN to estimate inheritance probabilities of pedigrees. Our study revealed that PPTC applied on BNs could be a more efficient method than Markov Chain transition probability methods, in population genetics.

Chow and Liu (1968) have used unrestricted Bayesian tree models to approximate multivariate discrete probability distributions. They show that solving the maximum weight spanning tree problem in the complete graph with edges between events weighted by their mutual information provides a maximum likelihood tree estimate. Molecular biologists and statisticians started modeling evolution as a Markov process proceeding through a tree structure. Cavendar (1978) and Farris (1973) proposed a very simplified model, in which the evolu-
tion of a binary value (the presence or absence of a genetic element) is followed through the set of species, known as phylogenetic trees or Cavendar-Farris (CF) trees. Desper et al (1999) uses tree structures called mutagenetic trees to model mutations in genetics, who call these tree models in the context of Oncogenesis, as oncogenetic trees. Oncogenetic graph models on gene level would be a powerful tool to get a deeper understanding of cancer progression. In Chapter 4, we prepared an algorithm based on the creedal set theory, Bayesian imprecise probability concepts and set-based weighted graph theory concepts to construct efficient pathways of mutations.

BNs are effective when there are no cyclic dependencies among data. Also, several networks with the same undirected graph structure, but different directions of some edges may represent the same distribution. And the dynamics of the system under study, could not be taken into account when using a BN to model an evolutionary pathway. Signaling pathways are dynamic events that take place over a given period of time. The limitations of BN may be over come by Dynamic Bayesian Networks (DBN), which model the stochastic evolution of a set of random variables over time. In Chapter 5, we constructed Dynamic Bayesian Networks that allow cycles, and used the concepts of Bayesian imprecise probability and set-based weighted graphs.

Bayesian multinets further generalize BNs and can further reduce
computation. A multinet can be thought of as a network where edges can appear or disappear depending on the values of certain nodes in the graph, a notion called asymmetric independence assertions. The tree models pioneered by Desper et al. (1999) in the context of oncogenesis fits only certain subgroups of the data. In many situations the underlying structure of observed samples cannot be described sufficiently by a single oncogenetic tree. To overcome this limitation, Beerenwinkel et al. (2005), proposed mixture models of trees known as K-mutagenetic trees mixture model. Mixtures of trees are a subclass of multinets. Rtreemix is a package for estimating mutagenetic/oncogenetic trees mixture models and genetic progression scores. Beerenwinkel et al. (2005) introduced the Mtreemix package implemented in C/C++ that provides an efficient code for estimating the mutagenetic trees mixture models from cross-sectional data. Similar to Mtreemix, the Rtreemix package provides functions for learning the mixture model from given data, simulation, likelihood computation and estimation of the GPS values. In chapter 6, we used the Rtreemix package to fit a 2-trees mixture model to a clonal HIV sequence data and could establish its stability through bootstrap resampling method. In all the different models we studied, it is observed that the use of Bayesian procedures either provide results much better than or at least the same as that of alternate procedures.
0.3 References


