Chapter 1

Introduction

1.1 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.

Genetics is the branch of science that studies the ways in which hereditary information is passed on from parents to offsprings and that is why a genetic component is considered to be evidence in the aetiology of most human diseases. The Human Genomic Project (HGP) which was started in the late 1980’s is an international effort to map and sequence all the genes in the human body. Its mission is to identify the full set of genetic instructions contained in our cells and to
read the text written in the language of the hereditary chemical DNA. Detailed standardized maps will allow scientists to decipher the genetic instructions encoded in the human DNA and enable them to find and study the genes involved in human diseases in a more efficient and rapid way. The knowledge and analysis of this information will provide new strategies to diagnose, treat and possibly prevent human diseases. The identification of genes responsible for a particular genetic disease might enable the defective gene to be replaced with a normal gene by gene therapy or gene transfer, thereby offering a possible cure for patients, who suffer from such disorders.

In human genetic research, sophisticated statistical methods are increasingly being used to analyze data. The most widely used one being the Bayesian inference method. In simple data analysis, the data are analyzed on their own terms, essentially without extraneous assumptions.

In classical inference methods, observations are postulated to be the values taken on by random variables (rvs) which are assumed to follow a joint probability distribution $P$, belonging to some known class $\mathcal{P}$. Frequently, the distribution is indexed by a parameter $\theta$ taking values in a set $\Theta$ so that $P = \{P_\theta; \theta \in \Theta\}$. The aim of the analysis is then to specify a plausible value for $\theta$ or at least to determine a subset of $\Theta$ which we can plausibly assert that it does or does not contain $\theta$. Such a statement about $\theta$ can be viewed as a summary of the information provided by the data and may be used as a guide to action.
In Bayesian approach, it is assumed in addition that $\theta$ is itself a random variable (rv) with a known distribution. This prior distribution is modified in light of the data to determine a posterior distribution. That is, the conditional distribution of $\theta$ given the data, which summarizes what can be said about $\theta$ on the basis of the assumptions made and the data.

Historically, Bayesian analysis have been of limited use for many practical applications. The posterior distributions are very difficult to calculate unless quite simple models are specified, and in part this complexity had led to concerns about the need to make assumptions about the prior distributions of the parameters. But, now, the situation is quite different. The Markov Chain Monte Carlo (MCMC) procedures allow us to avoid the difficulties of calculating posterior distributions by using simulations based approximation.

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 interact 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.

Genetic information may determine for individuals and their families whether they have a certain disease or may have a propensity for disease, or may be carrier for a genetic disease. A way of processing this kind of information, is using Bayesian Network Model (BNM). BNs create a very efficient language for building models of domains with inherent uncertainty.

BNs rely on inference algorithms to compute probabilities in the context of observed evidence. One established method for exact inference on BNs is the *Probability Propagation in Trees of Clusters* (PPTC) algorithm. Many complex genetic computations can only be performed
approximately and involve repeated random sampling techniques, typically in the form of MCMC methods. PPTC algorithms are of crucial importance to many of the steps within any efficient MCMC algorithm.

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. Perhaps the most straightforward generalization is to replace a single probability specification with an interval specification. Lower and upper probabilities, or more generally, lower and upper expectations (previsions), aim to fill this gap. The theory of Bayesian imprecise probabilities when used to estimate DBN and oncogenetic trees are capable of giving efficient and robust results.

A mixture of trees is a set of trees along with a probability distribution over them, and RtreeMix is a package that offers an environment for estimating the mutagenetic trees mixture models from cross-sectional data and using them for various predictions. It includes functions for
fitting the trees mixture models, likelihood computations, model comparisons, waiting time estimations, stability analysis, etc.

1.2 Objectives of the Study

Probabilistic graphical models are useful for extracting meaningful biological insights from the resulting data sets. These models provide a concise representation of complex cellular networks by composing simpler sub models. Procedures based on well-understood principles for inferring such models from data facilitate a model-based methodology for analysis and discovery.

The objectives of this thesis work could be briefed as,

1 Application of MCMC methods to estimate oncogenetic characteristics.

2 Application of Bayesian networks to model Genetic inheritance patterns.

3 Application of Probability Propagation in Trees of Clusters (PPTC) algorithm on Genetic inheritance models and comparison with usual Markov chain method.

4 Construction of oncogenetic trees using concepts of imprecise probability theory and set-based weighted graph theory.

5 Estimation of Genetic characteristics using dynamic Bayesian networks.
6 Application of mixture of oncogenetic trees and Rtreemix package in Genetics.

1.3 Review of Literature

The term Monte Carlo was first used by Ulam and von Neumann as a Los Alamas code word for the stochastic simulations they applied to building better atomic bombs. Monte Carlo is about invoking laws of large numbers to approximate expectations. This applies when the simulated variables are independent of one another. Monte Carlo methods are reviewed thoroughly in Neal(1993), Gilks et. al(1996) and Tanner(1996). These methods could also be 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). MCMC methods are probably about 50 years old, and has been both developed and extensively used in Physics for the last four decades. However, the most spectacular increase in its impact and influence in Statistics and Probability has come since the late 80’s. It has now come to be an all pervading technique in statistical computation, in particular for Bayesian inference, and especially in complex stochastic systems. Many authors(Thompson 2000, Jensen, Kong 1995, Sheehan 2000, Fernandez et. al 2001, Heath 2003) adopt MCMC methods in complex genetic computations. Pearl(1988) is one of the pioneers who helped Bayesian methods for uncertain reasoning become popular in the artificial intelligence community. Spiegelhalter et. al(1993) reviews
these probabilistic expert systems.

Dawid (1979) introduced the axiomatic basis for treating conditional independence. A probabilistic approach to dealing with uncertainty in expert systems began with the realization that calculations on seemingly intractable high-dimensional problems can be efficiently performed when a set of simplifying conditional independent assumptions is imposed (Pearl 1988, Lauritzen and Spiegelhalter 1988). These assumptions essentially split the problem into small manageable components. The immediate advantage is that a complex problem can be represented in a graphical form which can then inform the development of efficient computational algorithms for performing calculations.

More recent accounts of conditional independence with emphasis on graphical models and their Markov properties are given by Whittaker (1990) and Lauritzen (1996). The most common of these graphical models are the BNs (Pearl 1986, Jensen 1996). A BN is a graphical model for probabilistic relationships among a set of variables. Over the last decade, the BN has become a popular representation for encoding uncertain expert knowledge in expert systems (Heckerman et al 1995a). Graphical models and methods for learning them are provided in Buntine (1996). One established method for probabilistic inference on these BN’s is the PPTC method, developed by Lauritzen and Spiegelhalter (1988) and refined by Jensen et al (1990). The set of familial relationships among a group of individuals forms a pedigree and a variety of
graphical representations have been developed for handling pedigrees in a precise and consistent manner. As several authors including Kong (1991) and Heath (2003) have acknowledged, such representations logically tempt an exploitation of graphical models (Lauritzen 1996) for the description and analysis of genetic problems associated with pedigrees. Many pedigree analysis can be formulated using BN representation.

The problem of classification is one of the main concerns in the design of intelligent information systems such as pattern recognition, inductive learning and expert systems. In many of these applications the essential task is to estimate the underlying $k$-dimensional probability distributions from a finite set of samples. Because of the dimensionality concerns, the probability distribution function is often approximated by some simplifying assumptions, such as statistical independence. The independence assumption is simple but may be unrealistic in certain applications. It was suggested by Lewis (1973) that the optimal product approximation can be obtained by minimizing the divergence measure between the true and approximate distributions. Chow and Liu (1968) introduced the notion of tree dependence to approximate a $k$-th order probability distribution by a product of $(k-1)$ second order component distributions. One can then reduce the problem to finding a dependence tree with maximum total branch weight of mutual information. Wong and Wang (1989) suggested another product approximation by minimizing an upper bound of the Bayes’ error rate. This method is referred to as Error Probability Minimax. As only approximations can
be obtained it is necessary to have a tool capable to quantify the goodness of an approximation by adopting a measure of distance between the true and estimated probability distributions. One such measure is the Kullback-Leibler divergence measure. In order to minimize this theoretical information measure the well known mutual information measure (MIM) was used by Chow and Liu (1968), as a practical tool for measuring the degree of dependence between all pairs of features and consequently, for selecting the most dominant dependence, which determine the dependence tree. Datcu et al. (2002) uses this method in multisource data classification.


Modeling sequential data is important in many areas of science and engineering. Hidden Markov models (HMMs) and Kalman filter models (KFM)s are popular for this because they are simple and flexible. For example, HMMs have been used for speech recognition and bio-sequence analysis, and KFMs have been used for problems ranging
from tracking planes and missiles to predicting the economy. However, HMMs and KFMs are limited in their "expressive power". DBNs generalize HMMs by allowing the state space to be represented in factored form, instead of as a single discrete random variable. DBNs generalize KFMs by allowing arbitrary probability distributions, not just (unimodal) linear-Gaussian (Murphy 2002). Kim et al (2003) applies DBNs to time series micro array gene network data. Perrin et al (2003) applies stochastic differential equation method with Kalman Filter method to form a DBN to represent evolution of gene networks.


1.4 Organization of the Thesis

The thesis is divided into six chapters where each chapter is further divided into sections and subsections. Each chapter begins with a brief introduction of the problem described in it.
In the first chapter, a brief introduction of the relevance of probabilistic graph models and Bayesian methods is presented. The objectives of the study and a brief survey of the related literature on Bayesian methods and probabilistic graph models are also included.

In chapter 2, basic concepts of Markov chains (section 2.1), Monte Carlo methods (section 2.2), MCMC methods (section 2.3) and Bayesian methods (section 2.4) are explained. In section 2.5, we apply MCMC method to a breast cancer study.

In chapter 3, we introduce the basic concept of graph theory (section 3.1), the association between Markov chains and directed graphs (section 3.2) and Markov properties of graph models (section 3.3). Markov chains in Genetics are explained with a view to the well known Hardy-Weinberg law in Genetics in section 3.4. Section 3.5 defines Bayesian Networks and applies the algorithm PPTC to a genetic inheritance model.

Chapter 4 describes various methods of estimation of probabilistic trees. Sections 4.1 and 4.2 define trees and oncogenetic trees. In section 4.3 we present existing methods of tree estimation. Section 4.4 develops our proposed method of tree construction based on imprecise Dirichlet model.

In chapter 5, various methods of estimation of dynamic Bayesian
networks are explained. We define DBN and methods of its construction in Genetics in section 5.1. Imprecise Beta Model DBN is proposed in section 5.2 and algorithm for its construction in section 5.3. The section 5.4 is about construction of proposed DBN.

In chapter 6 we explain the concept of mixtures of trees in the context of cancer studies. The general class of Dynamic Bayesian Multinets is defined in section 6.1 and mixtures of trees in section 6.2. Beerenwinkel’s method of estimation of mixtures of K-mutagenetic trees is explained in section 6.3 and the relevance of Genetic Progression Score in section 6.4. Rtreemix package is used to construct mixtures of onco-genetic trees in section 6.6.

The program codes used for performing the numerical computations are given in Appendix I to Appendix III.