uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Amir R. Raji-Kermany

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.A.Sc. (Electrical Engineering)

GRADE / DEGREE

School of Information Technology and Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Stochastic Modeling of Multi-Locus Recombination in Mendelian Populations

TITRE DE LA THÈSE / TITLE OF THESIS

D. Hickey

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

M. Turcotte

X. Xia

Gary W. Slater

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES /
DEAN OF THE FACULTY OF GRADUATE AND POSTDOCORAL STUDIES

Stochastic Modeling of Multi-Locus Recombination in Mendelian Populations

Amir R. Raji-Kermany

School of Information Technology and Engineering (SITE)
University of Ottawa
Ottawa, Canada

January 2005

A Thesis submitted to the Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements for the degree of
Master of Applied Science

Your file Votre référence
ISBN: 0-494-11389-8
Our file Notre référence
ISBN: 0-494-11389-8

Abstract

Several mathematical models have been developed to describe the genetic structure of populations. Most of these models focus on only one, or few genetic loci. This thesis focuses on the models, which describe the dynamics of populations with respect to a large number of loci simultaneously. Such models are appropriate for recent large-scale studies of single nucleotide polymorphisms (SNPs). A random set representation of haplotypes in the population under study is proposed and the recursion equation for change of haplotype distributions in discrete time model is derived. A continuous time model for change of haplotype distributions is discussed and the solution to the differential equation describing change of probability measure is derived using the concept of blocks identical by descent (BID). The behavior of the solution is studied and a graph theoretical interpretation of the solution is investigated.

Acknowledgment

I would like to express my sincere gratitude to my supervisor Professor Charalambos Charalambous from School of Information Technology and Engineering (SITE) and my co-supervisor Professor Donal Hickey from department of Biology of University of Ottawa, for their indispensable guidance and discussions, technical expertise as well as their supports in helping me completing this thesis.

I am thankful to Professor NasirUddin Ahmed, from SITE for his valuable theoretical advices and his guidance and support during my first year of graduate study. I would like to recognize Professor Yongyi Mao from SITE for his advices on the topic of graphical models.

I would like to thank University of Ottawa, for providing financial support for this research by "Interfaculty Collaborative Research Initiatives" grant.

I am thankful to SITE and Department of Biology of University of Ottawa, and my colleagues in SITE, especially Mr. Stojan Denic for our fruitful discussions.

I am grateful to my girlfriend, Christine, for her constant support in helping me write this thesis.

Finally, my deep gratitude goes to my father as my first teacher, and my mother for her patience, everlasting encouragement and love throughout my studies.

Contents

List of Figures

List of Tables

List of Terms and Abbreviations

cdf	Cumulative Distribution Function
BID	Blocks Identical by Descent
BN	Bayesian Network
DAG	Directed Acyclic Graph
DNA	Deoxyribonucleic acid
HapMap	Haplotype mapping for human genome
HW	Hardy-Weinberg Principle
LD	Linkage Disequilibrium
ODE	Ordinary Differential Equation
MRF	Markov Random Field
mRNA	Messenger RNA
pdf	Probability Density Function
pmf	Probability Mass Function
RNA	Ribonucleic acid
RV	Random Variable
SNP	Single Nucleotide Polymorphism

Chapter 1

Introduction

Recent advances in molecular biology have provided efficient methods for sequencing DNA from many different living organisms. With the easy access to these sequences for researchers all around the world, biology in general, and more specifically genetics, has entered into a new era. Available data from human genome project [75], and other genome projects such as HapMap [74], contain enormous amount of information about evolutionary relationships among species [22], causes of varieties in populations [30], history of populations [28], and information about genes controlling common diseases in human populations [67]. Mathematical models in population genetics are essential for analysing these data [16].

1.1 Historical Background

The theoretical population genetics was initially developed in 1920's and 1930's by R.A Fisher (1890-1962)[20], S.Wright (1889-1988) [65], and Haldane (1892-1964) [31]. Their objective was to formulate evolutionary theory based on Mendelian hereditary mechanism [16]. Since then the theory have been developed in different directions, resulting to development of mathematical models such as genetic algebras in order to model populations [55, 64, 59] and employing mathematical tools such as Bayesian Networks and Markov Random Fields in order to model the inheritance process [41, 69].

There are many evolutionary forces involved in change of genetic composition of populations namely, mutation, natural selection, recombination and genetic drift. Similar to modeling physical systems, in which the model is made based on some simplifying assumptions, in most of the mathematical models in population genetics, a hypothetical model based on some simplifying assumptions is made. As an example, in the models that we will discuss in the subsequent chapters, the size of population under study is considered to be very large such that the effects of fluctuations in the population size are negligible. In a general categorization we can name two approaches for modeling the change of frequencies of genetic variants in a population:[2]

- The Discrete Time Model: In this model which is an extension of the Hardy-Weinberg model (see Section 2.5.5 and [30]), the generations are assumed to have no overlaps with each other. This model has been developed mainly by Geiringer [23], Bennett [4], Liyubich [38] and Dawson [12].

- The Continuous Time Model: In this model (which is not widely discussed in the literature) the change in genetic composition of populations is assumed to be a gradual process, therefore at each instance of time as the result of recombination the genetic composition of the population is changing. One of the most recent mathematical models developed in this framework is Baake and Baake model [3, 2] which will be discussed in Chapter 5 of this thesis.

1.2 Population Dynamics

One of the most important problems in all scientific fields is the prediction of the state of the system under study, knowing its initial condition [62]. For example in Newtonian mechanics, the system under study is a particle, and the state of the particle is its coordination in the Cartesian space. In that case we are interested to know the position of the particle at each instance of time. Knowing the particle's initial

position in space, Newton's laws of motion, enables us to predict its exact position at each instance of time. In practice, in some cases, it is impossible to know the exact initial condition of the system under study, (e.g. Quantum Mechanics) or it is not feasible to seek for the exact initial condition (complex systems like fluids or gases in thermodynamics). In such cases we need to analyse the system from a probabilistic point of view (stochastic approach). We can consider each possible state of the system as an *event*, and we assign a number $P \in [0, 1]$ to each event, which is representative of the likelihood of occurrence of that event.

In population genetics, the system under study is a collection of individual genotypes (depending on population under study, individuals have be haploid or diploid genotypes [30]) (see Section 2.2), and the state of the system is identified by alleles and their arrangement on the genome of each individual. Similar to the case of complex mechanical systems, it is almost impossible to know the exact genetic characteristics of each individual and also the exact manner in which each individual transfers its genetic information to its offspring. For this reason we model a population as a sample space of random events, in which each event corresponds to a certain configuration of genetic variants, and we are interested in finding the likelihood of each event. In the study of dynamics of populations we are interested in changes in the frequency of genetic variants over time. Therefore, the goal is to describe the probability of a variant at time t in terms of probability of that variant at the initial time $t = 0$.

As we will discuss in the coming chapters, the parameters involved in the equation describing the change of probability distribution, contain information about the physical position of the genetic variants under study, therefore we can use this information for physical mapping of the genes [46, 56].

1.3 The Thesis Work

Because of the interdisciplinary nature of this thesis, two chapters of the thesis (chapters 2 and 3) are devoted to provide background information both in biological and mathematical theories which are used in the subsequent chapters.

The discrete time model with the formalism proposed by Dawson [12] is discussed in Chapter 4. The similarity of the mathematical notations used in that model with the formalism used in the theory of random sets [45, 24], was a motivation for reformulating Dawson's model in terms of random sets. Random sets are generalizations of the familiar concept of random variables in probability theory, widely used in data fusion and stochastic geometry [24]. This Chapter of the thesis suggests identifying each haplotype configuration by a subset of the set of all loci under study, then the probability of having a certain haplotype in population is described in terms of probability mass function of the corresponding random set. This model is used to find the recursion law for the change of haplotype frequencies between successive generations. Also we have discussed special cases of the model for the case in which one and two loci are considered and we have compared these results with the HW principle.

In Chapter 5 we discuss the continuous time model based on the measure theoretic approach of Baake and Baake [2]. In this model the change of haplotype frequencies has been described by a differential equation in which the solution to this equation is a time-dependent probability measure, which describes the change of haplotype distributions. In this thesis, we show that the coefficients of this solution can be interpreted as the distribution of blocks identical by descent (BID) (see Section 5.3.1 and [?, 5]). In addition, we discuss the behaviour of distribution of BID's under two assumptions: the assumption of the existence of hotspots of recombination, and assumption of the absence of these hotspots on genome. Also briefly, we provide a graph theoretical interpretation for the solution of the differential equation in Baake and Baake 2003.

1.4 Organization of Thesis

This thesis is divided into six chapters including this introduction.

Chapter 2 is devoted to provide background knowledge in biology and population genetics.

In Chapter 3 we discuss the basic mathematical concepts which will be used in this thesis, including a brief review on main notions of measure theory and probability theory, graphical models including Bayesian Networks and Markov Random Fields and finally random sets.

Chapter 4 discusses the discrete time model for dynamics of population in a random mating model in the absence of mutation and natural selection. The difference equation describing the change of haplotype distributions is derived using the notion of random sets.

Chapter 5 derives the solution for the differential equation proposed by Baake and Baake [2, 3], using the notion of BID's and special cases for the solution regarding the existence or absence of hotspots of recombination is discussed also a graph theoretic interpretation for the solution has been investigated.

1.5 Contributions

This thesis proposes using random sets for modeling haplotypes in a population. As an example of the application of the random set theory in population genetics, Chapter 4, derives the recursion equation for the change of haplotype distributions using random sets. In this chapter of the thesis, the relationship between mathematical terms in discrete time model, (with the formalism proposed by K.Dawson [12]) and the random set theoretic notions, such as commonality measure and pmf of random sets, is established.

In Chapter 5 of the thesis, the relationship between the solution to the differential equation described by Baake and Baake [2, 3] and the concept of blocks identical by

descent is investigated.

This thesis derives the solution to the non-linear ODE based on the concept of blocks identical by descent in a population. The result is reformulated in terms of the recombination operators as defined in [3]. In addition, this method of solving the ODE, shows that the coefficients of the solution, describe the distribution of the block structures in a population.

Chapter 2

Basic Concepts of Genetics

This chapter introduces the main biological concepts and the actual processes which cause changes in genetic composition in populations and also causes of differences between different populations of species. In this chapter we describe the basic concepts of evolutionary theory and population genetics.

2.1 Cells, Chromosomes and DNA

The cell is the basic unit of all living organisms. Some organisms, such as bacteria, are made of a single cell. By contrast, most of non-microbial organisms are made up of many cells. In a general division, there are two types of cells: eukaryotic and prokaryotic cells; most multi cellular organisms are constituted of eukaryotic cells [14]. Eukaryotic cells are made of a variety of specialized structures like membrane-bound organelles and the nucleus which is the largest organelle within the cell and contains the genetic material. Bacteria are said to be prokaryotic. This type of cell is small and does not have a very complex internal organization. The genetic material in prokaryotic cells is free within the cell [36].

In organisms with sexual reproduction, there are two types of cells: *somatic* cells, which are all the body cells except sex cells, and *sex* cells which also are referred as *gametes*, which is a general term for sperms or eggs. In the nucleus of somatic cells, genetic

materials (chromosomes) are in pairs, in the sense that there are two copies of each chromosome (and consequently each gene) in the cell. For this reason, these types of cells are often called as *diploid* cells. In contrast, sex cells (or gametes) have only one copy of each gene (chromosomes are not in pairs) and are called haploid [36].

Cells are able to replicate themselves (for example somatic cells generate other somatic cells or they generate sex cells), through this process, a copy of the genetic material is produced in the nucleus of the cell and these copies will be transferred to the daughter cells.(see section 2.3) [36].

2.1.1 Chromosomes and DNA

A chromosome, contains a long continuous piece of DNA, located in the nucleus of the cell. In somatic cells of all diploid individuals, chromosomes appear in the form of pairs of sister chromosomes (one inherited from the father and the other from the mother of the individual). For example in all human somatic cells, there are 23 pairs of chromosomes, of which one pair, X and Y chromosomes, carry sex linked traits. Non-sex chromosomes, are called *autosomes*[30]. Through the process of mitosis and cell division (see Section 2.3) copies of chromosomes are made and these copies will be passed to the offspring.

Deoxyribonucleic acid molecules, known as DNA, are the basic structures carrying hereditary information in all living organisms (except some viruses in which RNA does this task) [22]. DNA usually consists of two complementary strands twisted around each other to form a right-handed double helix. Each chain consists of four kinds of nucleotides:*adenine (A),guanine (G),thymine (T),cytosine (C)* [22]. The two chains of DNA are joined throughout their lengths by hydrogen bonds between complimentary pairs of nucleotides. In order to know the nucleotide composition of a double helix (the arrangement of nucleotide along the DNA) we just need to know the nucleotide composition of one strand, because T nucleotide always pairs with an A, and a C with

a G. Therefore, for example if one strand is 5′ ATCG 3′ we know that the complementary strand is 5′CGAT3′, in which 5′ denotes the starting point of transcription and 3′ denotes the end point [22].

2.2 Genetic Information Processing

A *Gene*, traditionally is defined as the entity transmitted from parents to offspring that influences hereditary traits [36]. Traits like hair color, skin color, height and behaviour, depend on a combination of genetic structure of the individual and environmental factors. From a molecular point of view, a coding gene, is a segment of the DNA which can eventually be translated into amino acid sequences, which are building blocks of a protein chain. Decoding of genetic information to amino acid sequences is done in the cell through two separate steps: *Transcription* and *Translation* [36].

In the transcription step, only informative parts of the DNA (genes) will be used to make a sequence of a special type of RNA, called messenger RNA usually denoted by mRNA. An mRNA, similar to DNA, can be considered as as sequence of four nucleotides with the only difference that instead of T nucleotide we have a U (Uracil) nucleotide. In the next step, translation, mRNA sequence will be used as a template which contains the required information about the arrangement of amino acids in a protein sequence. In this step every three nucleotides (codons) in the coding region of the DNA, will be translated into one amino acid, this translation is done using a dictionary called the genetic code, which provides a universal rule for protein synthesis for all organisms (there are some rare exceptions in which a slightly different genetic code is used for translation) [?, 22]. As we see in Table 2.1 , genetic code is degenerated, in the sense that more than one codon (triplets of nucleotides) can be translated into the same amino acid. For example, both ATT and ATC codons stand for Isoleucine (Ile) amino acid.

TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys	ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
TTA	Leu	TCA	Ser	TAA	STOP	TGA	STOP	ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
TTG	Leu	TCG	Ser	TAG	STOP	TGG	Trp	ATG	Met	ACG	Thr	AAG	Lys	AGG	Arg
CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
CTC	Leu	CCC	Pro	CAC	His	CGC	Arg	GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly
CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly
CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Table 2.1: The Standard Genetic Code

The Position of a gene on a chromosome is called the *locus* (loci in plural form). A gene can appear in many forms (which one form is the dominant form), each of these forms of a gene are called *alleles* of that gene. Observable traits of an individual (also called the phenotype) depend on the pair of alleles that the individual has inherited from his parents. For example, assume that there are three possible states for the eye color of an individual: brown, green or blue. If the individual has the pair brown-brown or brown-blue or brown-green in the corresponding locus, the eye color will be brown. Note that although the individual may have blue or green alleles paired with the brown allele, as brown is dominant (or wild) type, the phenotype will be brown. The unordered pair of alleles (i.e. brown-blue, blue-green,... for eye color) is called the *genotype*. In general, each phenotype is influenced by a certain genotype. Different genotypes may result to the same phenotype and also environmental factors can affect the phenotype of an individual. A genotype may be composed of one gene or it can have several numbers of genes. Therefore the corresponding phenotype can be influenced by a single gene (single gene traits) or many genes (complex traits or multifactorial traits).

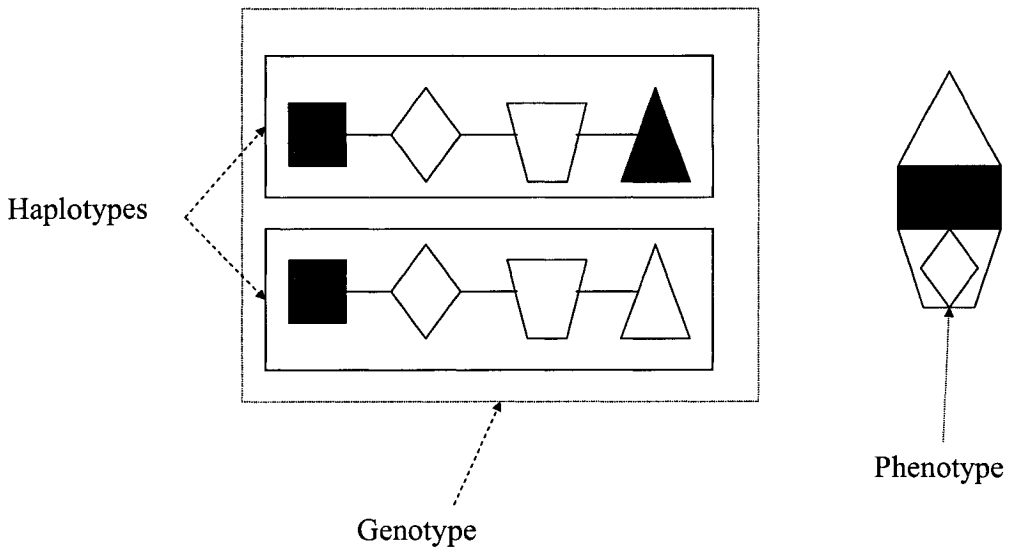


Figure 2.1: A Schematic Representation of Haplotype, Genotype and Phenotype

As we mentioned before, a diploid genotype is a pair of two sequences of alleles each called a haploid genotype or *haplotype*. A haplotype can be composed of just one allele (in case of simple traits), or it may be composed of hundreds of alleles. Figure 2.1 is a demonstration of relation between genotype, haplotype and phenotype. The complete set of genes in a haploid set of chromosomes of a particular organism, is called the *genome* of that organism. The human genome, for example, contains approximately three billion nucleotides [36]. Figure 2.2 summarizes the information processing in the cell.

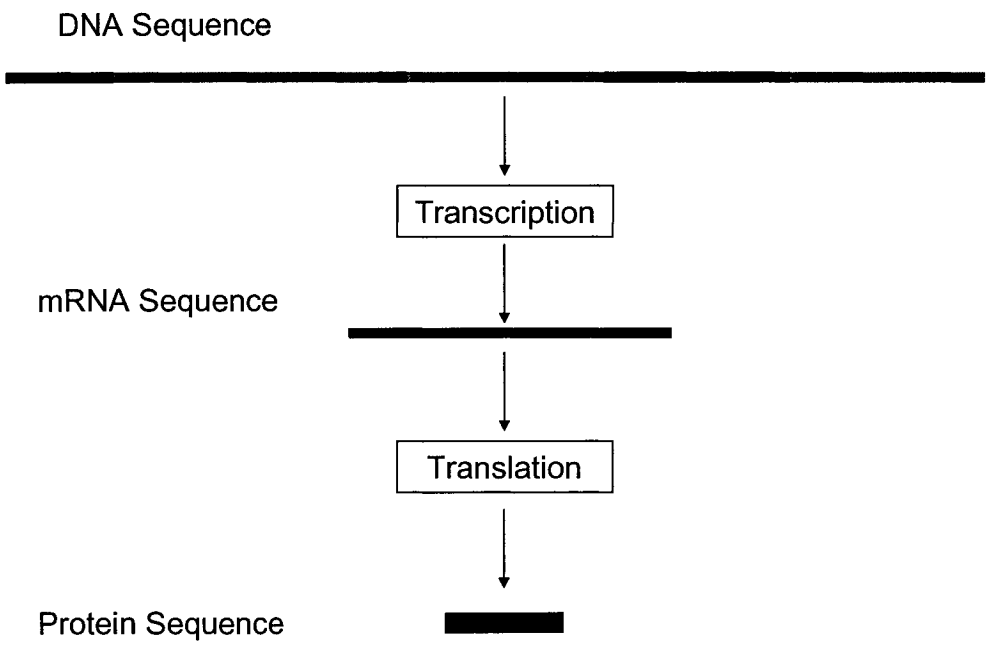


Figure 2.2: Decoding the DNA information into a protein sequence

In Figure 2.2, informative parts of the DNA are transcribed into an mRNA sequence through RNA processing, the mRNA sequence is used as a template for protein synthesis through the translation step.

2.3 Cell Division and Inheritance

Two types of cell division occur in human and other sexually reproducing species:*mitosis* and *meiosis* [36].

Mitosis is the process by which identical genes will be transferred to all the somatic cells from the original single-celled zygote. In mitosis, the nuclear material divides so that the resulting daughter cells are genetically identical. Meiosis only occurs in the process of formation of gametes (sex cells). Through this process, a combination of paternal and maternal chromosomes, formed as a new single chromosome, will be passed to the sex cells. In the process of mating, each individual randomly passes a chromosome to the zygote, (Figure 2.3), which, as a result, offspring individuals will have a copy of paternal and one copy of maternal chromosomes in each of their cells [36, 30].

2.3.1 Mendel's Inheritance Laws

Johann Mendel (1822-84), through experiments with garden peas, noticed the random segregation of heredity factors in the process of mating [48]. In modern terminology we can reformulate his laws of inheritance as follows:

- Mendel's first law of segregation: during meiosis the two members of any pair of alleles possessed by an individual separate into different gametes and subsequently into different offspring, without any interaction with each other

- Mendel's second law of Independent Assortment: during meiosis all combinations of alleles are distributed to daughter cells with equal probability, independent of each other.

For the case of one gene with two alleles, suppose that an individual has allele A on his paternal chromosome and allele a on his maternal chromosome , then it is equally likely that the gamete produced by the individual receives an A or a allele, in mathematical

terms:

$$\Pr(\text{gamete receives an A allele}|\text{individual is Aa type}) =$$

$$\Pr(\text{gamete receives an a allele}|\text{individual is Aa type}) = 1/2$$

In which $P(\cdot|\cdot)$ is the conditional probability (see Chapter 3). Now suppose two Aa individuals mate with each other, then the probability of having an Aa offspring is:

$$\Pr(\text{father passes an } a \text{ allele and mother passes an } A \text{ allele}) +$$

$$\Pr(\text{mother passes an } a \text{ allele and father passes an } A \text{ allele}) =$$

$$\Pr(\text{father passes an } a \text{ allele}) \Pr(\text{mother passes an } A \text{ allele}) +$$

$$\Pr(\text{mother passes an } a \text{ allele}) \Pr(\text{father passes an } A \text{ allele}) =$$

$$\frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Above, we have used the fact that if two events are independent then their joint probability is the product of the marginal probabilities [42]. As we will see later, Mendel's second law is not valid for linked loci; in fact, linkage causes departure from independent assortment due to the physical attachment of the linked loci. In Section 2.5.6, we will see that departure from independent assortment enables us to estimate the distance between genes under study.

2.3.2 Recombination

The process by which a combination of parental chromosomes is formed, is called *recombination* and plays an important role in the emergence of new types in a population [30].

As we discussed in Section 2.2 chromosomes appear in pairs as a system of homologous chromosomes. Homologous chromosomes have the same appearance and the same number of loci, but the corresponding loci may be occupied by the same or by different alleles of the same gene. Roughly speaking, during the process of meiosis, homologous

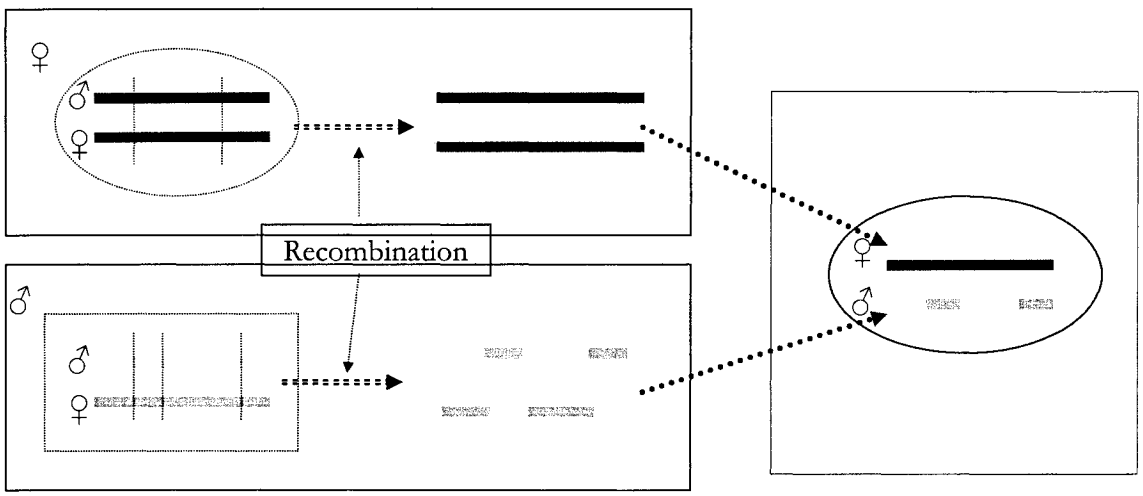


Figure 2.3: Representation of random mating

Each individual randomly passes a copy of its recombined genetic information to the offspring

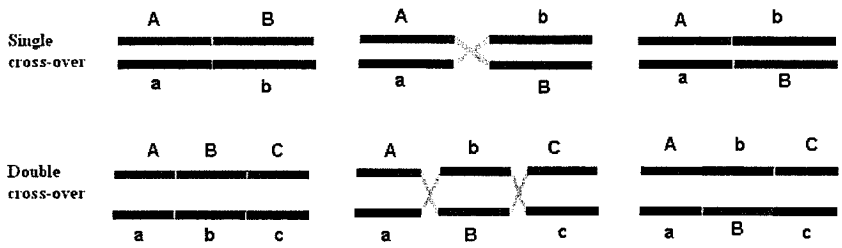


Figure 2.4: Recombination and Crossover

chromosomes break in the same segments and switch the corresponding segments. In other words, homologous chromosomes interchange some segments. This phenomena which is called crossing over is one of the main reasons for emergence of new haplotypes in a population. A simplified recombination model for one and two crossover events have been shown in Figure 2.4. As we see in Figure 2.4, the original diploid cell has haplotypes $(AB), (ab)$ but the gamete produced by this type, contains (Ab) or (aB) haploids, also in the second case, a $(ABC), (abc)$ diploid has produced two different haplotypes (AbC) and (aBc).

2.4 Molecular Evolution

The aim in the process of producing offspring, in all the species, is to pass genetic information without any change to the offspring. But occasional mistakes in copying DNA sequences may occur, giving rise to new sequences. These errors in copying are the main sources of emergence of new species in the nature. In this Section we will have a brief review on the main concepts of evolutionary genetics in a molecular level.

2.4.1 Mutations

The errors in replication of a DNA sequence are called *Mutations.* Mutations are the main source of variation and emergence of new types in nature [22]. Mutations can occur in either somatic or sex cells, but as somatic mutations are not passed on to the offspring, they will not have any effect on the process of evolution [30]. For this reason, usually the term mutation is referring to mutations in sex cells. Mutations may affect a single nucleotide, *point mutations* or a set of adjacent nucleotides *segmental mutations* [22]. Mutations can also be classified by the type of change caused by the mutational event. From this perspective we can distinguish four types of mutations [22]:

1. Substitution mutations: In this type of mutation, one nucleotide will be substituted by another nucleotide, for example, a G nucleotide will be replaced by a C, in other words, in the process of copying a mistake will happen in copying G nucleotide, and instead of G, a C will be written at the corresponding position in DNA of the offspring.

2. Deletions: in this type of mutation one or more nucleotides will be removed from the DNA

3. Insertions: One or more nucleotide will be added to the DNA sequence

4. Inversions: A 180^0 rotation of a segment of the DNA sequence

Mutations are assumed to occur at random through time. Also, in many cases it can be assumed that they occur at random along the genome, i.e. different nucleotides at different sites on genome have equal chance of experiencing a change. But it is not always true, some sites are more likely to experience a mutation event than others and some transitions are more frequent than others [22].

2.4.2 Changes in Gene Frequencies

The process of evolution, can be viewed as the process of changes in gene frequencies in a population over time. An important concern in population evolutionary genetics is to describe the process by which the frequency of a mutant allele will change in time under the effect of evolutionary forces.

Many molecular changes usually have small effect on the phenotype and therefore they will not affect the fitness of their carriers. But in some cases these changes will give rise to the emergence of a new allele in the population.

The relative proportion of an allele in population is referred as allele frequency [22] (or gene frequency), for example, suppose we consider a population of N haploid individuals, two alleles A_1 and A_2 are present at a certain locus, and the number of individuals carrying allele A_1 at that locus is n_1 and number of allele A_2 in the population is n_2. Then, the frequency of allele A_1represented by $f(A_1)$ is n_1/N and $f(A_2) = n_2/N$.

The process of evolution can be viewed as a process with two main drives: the emergence of new alleles in a population (meaningful mutation) and the process of change in allele frequencies within a population (natural selection and random genetic drift).

Most of the mutations which can affect the amino acid composition have negative effect on the fitness of the organism, therefore in time, as their carriers have less opportunity to survive (less productivity or shorter life time) these mutations will be eliminated from the population. These types of mutations are called deleterious mutations and the type of selection acting in this case is called purifying or negative selection. In some rare cases the mutated allele, will be as fit as the best allele in the population. These types of mutations are called neutral mutations and obviously, natural selection treats both alleles in the same way. In some extremely rare cases, a new mutation increases the fitness of its carrier. This type of mutation is called advantageous mutation and natural selection favours the mutant allele, in this case the selection is a positive selection, and the mutant allele will become fixed in the population after a number of

generations [22].

The other factor which may affect the fixation of certain type of mutation in the population is random fluctuations in the environment causing elimination of certain types (without being less fit than others) this phenomenon is called random genetic drift [30].

2.4.3 Evolutionary Trees

In a simplified model, we can consider that at the beginning, there has been only one type of living organism on earth (a putative organism which is called Cenancestor [22].) As a result of mutation and natural selection, the genome of the original sequence has been alternated in the way that other alleles of genes have emerged. At some points these changes are sufficient to result in a significant observable change in the organism, giving rise to the emergence of a new species. Therefore we can assume that the genomes of all existing organisms are copies of the ancestral genome, which the errors in copying in interaction with the process of natural selection are the source of divergence between their carriers. Therefore we can conclude that the species which are closer to each other (in a taxonomical sense) have a larger part of their genome in common. This can be verified by looking and comparing the available data. For example, Human and Gorilla have 98% of their genome in common, and only 2% of the genome differs between these two close relatives. In fact the difference between gene sequences can be used to construct phylogenetic trees and also estimate the divergence time between species. This field is one of the major research areas in bioinformatics and several algorithms and software have been developed in order to compare genetic composition of different types and estimate their divergence time and the relationship between them [22].

2.5 Basic Concepts of Population Genetics

2.5.1 Populations

In a broad sense, population genetics is the study of natural differences between organisms. From this perspective, evolutionary theory can also be viewed as part of this broad field. Genetic differences that are common among organisms of the same species are called *genetic polymorphisms* whereas genetic differences accumulated between different species - as a result of process of evolution, are called *genetic divergence* [30]. In a more specific sense when we refer to population genetics we consider the study of genetic polymorphism.

In population genetics, the word "population", usually does not refer to an entire species (for example the entire population of human beings), it rather refers to a group of individuals of the same species living within a sufficiently restricted geographical area that any member can potentially mate with any other member of the opposite sex in the population, hence the focus is on the local interbreeding units of possibly large, geographically structured populations. Such local populations are the fundamental units of population genetics. In population genetics usually we refer to these local populations simply as populations [30].

2.5.2 Polymorphism

A polymorphic gene is a gene for which the most common allele has frequency of less than 0.95 [30], and monomorphic gene is one that is not polymorphic. In any large population, almost all the genes have rare alleles associated with them. Rare alleles are with frequencies of less than 0.05.

Most of the genetic variations are hidden, in the sense that they are not apparent at the phenotypic level. However, genetic variations that can result in visible variations among the individuals are also common in most natural populations. In most cases, visible genetic variations (or phenotypic variations) are not traceable to the effect of individual

genes, and that is because of the fact that most of the traits are multifactorial, meaning that they are influenced by several genes, in interaction with the environmental factors [30].

In the case of human population, for example, any two copies of the human genome differ from one another by approximately 0.1% of nucleotide sites [74].

A *Single Nucleotide Polymorphism*, SNP, is a difference between chromosomes in the base present at a particular site in the DNA sequence. For example, some chromosomes in a population may have a C at that site (the 'C allele'), whereas others have a T (the 'T allele') [74].

2.5.3 Significance of Genetic Polymorphisms

Genetic variation provides a set of markers for the genetic studies of organisms in their natural environment. This method is especially important to apply on organisms for which genetic manipulation is not feasible or laboratory rearing or domestication is not possible. Studies of genetic variation are useful in inferring the ancestral history of a group of organisms. Also genetic polymorphisms are useful in investigating the genetic relationships among subpopulations in a species [30].

One of the most important applications of using genetic polymorphisms as markers in human populations, is to find out if they are linked with harmful genes that cause a certain disease. As we will discuss later, the recombination rate between linked loci reveals information about the physical position of an unknown locus with respect to a known marker; in fact, this method can be used for fine mapping of the human genome [56, 67].

2.5.4 The Effect of Random Mating and Recombination

The frequency of a specified genotype in a population is the genotype frequency. The mathematical relationship between allele frequencies and genotype frequencies, and the rate of change of these frequencies, can be used to infer the recombination rate between

loci involved in the genotype under study [30] (see Section 2.5.6). When there is more than one locus involved in a certain genotype, other than mutation and natural selection, recombination has an important effect in change of genotype frequencies. When a mutation event at a certain locus occurs, it results in the emergence of a new allele and consequently a new haplotype. Due to the effect of recombination and random mating of gametes, the frequency of haplotype will be changed in the population [30](see Chapter 4.

In order to describe the effect of random mating in changing haplotype distributions, we need to have a proper model for the system. In order to avoid certain complexities, we also need to make some assumptions about the system. One of the most common models in population genetics is the non-overlapping generation model (or discrete time model). In this model we assume a very simple life history. Specifically we assume that the members of the previous generation die before the members of the next generation are born. The only real life realizations of this model are populations of certain short-lived insects or annual plants that have short growing seasons. But this model is also a good approximation for other populations.

In order to be able to describe the change in frequencies, we also need to make some assumptions about the mating patterns in a population. One of the simplest assumptions is the random mating assumption, in which we assume that mating occurs at random with respect to the genotype under study. With random mating the chance that an individual mates with another having a prescribed genotype is equal to the frequency of that genotype in the population. This assumption is valid with respect to some traits (for example blood group in human population) but not realistic with respect to others (for example skin color in humans).

We can summarize these assumptions as follows:

1. The organism under study is diploid

2. Reproduction is sexual

3. Mating is Random

4. Population size is very large

5. Generations are non-overlapping

6. The effect of migration, mutation and natural selection can be ignored comparing to the effect of recombination

For the rest of this study we will be confined to these assumptions, in Chapter 5, we will consider a continuous time model in the sense that generations are not necessarily non-overlapping.

2.5.5 The Hardy-Weinberg Principle

The assumptions that we made in the previous section, constitute the Hardy-Weinberg model (HW model) [?, 68]. Now let us apply this model to predict the genotype frequencies in which the genotype under study has just one locus and the gene under study has two possible alleles denoted by A and a. With these two alleles the possible genotypes in the population are:AA, Aa, aa. In HW model, the mathematical relation between allele frequencies and genotype frequencies is described as:

$$f(AA) = p^2, \quad f(Aa) = 2pq, \quad f(aa) = q^2 \tag{2.1}$$

where $f(A) = p, f(a) = q$, are the allele frequencies, satisfying $p+q = 1$. The rationale behind the relations described in (2.1) lies behind the assumption of random mating. Let us denote the genotype frequencies by $f(AA) = P, f(Aa) = Q, f(aa) = R$ where $P + Q + R = 1$. As the individuals in the population are diploids, the total number of sites (loci) under study in a population of N diploid individuals is $2N$, as AA individuals have two A alleles on each locus, the total number of A alleles on carried by AA individuals is $2NP$. Also Aa individuals have one A allele, therefore the total

number of A alleles carried by Aa individuals in population is NQ, hence the total number of A alleles in the population is: $n(A) = 2NP + NQ$. as the total number of sites under study is $2N$ therefore:

$$f(A) = n(A)/2N = \frac{2NP + NQ}{2N} = P + Q/2 \tag{2.2}$$

With the same argument we can show that:

$$f(a) = n(a)/2N = R + Q/2 \tag{2.3}$$

Now let us consider the effect of random mating in allele frequencies of the next generation. As we have ignored the effect of mutation, possible alleles in the next generation are the same as before , therefore possible genotypes are AA, Aa, aa. And their frequencies are P', Q', R' respectively. Applying the total probability theorem [42], in mathematical terms we can write:

$$\begin{aligned} P' &= \Pr(AA \text{ in the next generation}) = \Pr(AA, AA \times AA) + \Pr(AA, AA \times Aa) \\ &+ \Pr(AA, Aa \times AA) + \Pr(AA, Aa \times Aa) \\ &= \Pr(AA|AA \times AA)\Pr(AA \times AA) + \Pr(AA|AA \times Aa)\Pr(AA \times Aa) \\ &+ \Pr(AA|Aa \times AA)\Pr(Aa \times AA) + \Pr(AA|Aa \times Aa)\Pr(Aa \times Aa) \end{aligned} \tag{2.4}$$

in which $AA \times Aa$ means mating event between type AA and Aa, also $\Pr(\cdot, \cdot)$ represents the joint probability of two events, and $\Pr(\cdot|\cdot)$ is the conditional probability (see Chapter 3). Note that here we have considered $AA \times Aa$ (an AA male mating with an Aa female) and $Aa \times AA$ (an Aa male mating with an AA female) as two different events. Based on Mendelian independent segregation model (Section 2.3.1) we have:

$$\Pr(AA|AA \times AA) = 1, \quad \Pr(AA|AA \times Aa) = 1/2, \quad , \Pr(AA|Aa \times Aa) = 1/4 \tag{2.5}$$

Also Mendelian model predicts that $\Pr(AA|Aa \times aa), \Pr(AA|AA \times aa), \Pr(AA|aa \times aa) = 0$ so we didn't include these terms in (2.4).

As mating is at random, mating probabilities are proportional to the product of frequency of each of the genotypes involved in mating therefore:

$$\Pr(AA \times AA) = \Pr(AA)\Pr(AA) = PP = P^2$$
$$\Pr(AA \times Aa) = \Pr(AA)\Pr(Aa) = PQ$$
$$\Pr(Aa \times AA) = \Pr(Aa)\Pr(AA) = QP$$
$$\Pr(Aa \times Aa) = \Pr(Aa)\Pr(Aa) = QQ = Q^2$$

Therefore we can write equation (2.4) as:

$$\begin{aligned} P' &= 1P^2 + 1/2(PQ) + 1/2(QP) + 1/4Q^2 \\ &= P^2 + PQ + 1/4Q^2 = (P + Q/2)^2 = p^2 = P \end{aligned} \tag{2.6}$$

With the same argument we have:

$$\begin{aligned} Q' &= 2PQ/2 + 2PR + Q^2/2 + 2QR/2 = 2(P + Q/2)(R + Q/2) = Q = 2pq \\ R' &= Q^2/4 + 2QR/2 + R^2 = (R + Q/2)^2 = R = q^2 \end{aligned} \tag{2.7}$$

Also if we calculate the allele frequencies p', q' from the genotype frequencies we will have: $p' = p$ and $q' = q$. Hence in this case allele frequencies and genotype frequencies will remain the same at each generation. Also we notice that the assumption of random mating is equivalent to the assumption of random union of gametes, therefore knowing the frequency of the gametes carrying the haplotype under study at each generation, we can find the corresponding genotype frequencies at the next generation by multiplying these frequencies. When the genotype frequencies doesn't change over time (i.e. it doesn't change as the result of random mating) we say that the population has reached to equilibrium with respect to the genotype under study, therefore based on HW principle, a population reaches its equilibrium state after one generation with respect to a single locus genotype.

2.5.6 Linkage and Linkage Disequilibrium (LD)

Let us consider two biallelic genes A and B with alleles represented by A_1, A_2 and B_1, B_2, with frequencies p_1, p_2, q_1, q_2 respectively, which $p_1 + p_2 = 1, q_1 + q_2 = 1$. HW principle states that genotype frequencies A_1A_1, A_1A_2, A_2A_2 are $p_1^2, 2p_1p_2, p_2^2$ and B_1B_1, B_1B_2, B_2B_2 frequencies are $q_1^2, 2q_1q_2, q_2^2$, and they will stay constant over time. As we saw in the previous section, A_1 and A_2 alleles are in random association with each other , i.e. $\Pr(A_1, A_2) = \Pr(A_1)\Pr(A_2)$ (in mathematical terms, as we will see in Chapter 3 it means that the two events A_1, A_2 are two independent events), this independence or random association, is because of the fact that the two alleles A_1 and A_2 are located on two different chromosomes and their segregation is independent of each other. When the alleles of the genes are in random association, the frequency of a gamete carrying any particular combination of alleles equals the product of the frequencies of those alleles. When genes are in random association we say that they are in a state of *linkage equilibrium* and if they are not in random association we say that the genes are in a state of *linkage disequilibrium* or LD for short. Therefore with linkage equilibrium the haplotype frequencies of the particular case of two biallelic genes are:

$$\begin{aligned} A_1B_1 &: \quad p_1q_1 \\ A_1B_2 &: \quad p_1q_2 \\ A_2B_1 &: \quad p_2q_1 \\ A_2B_2 &: \quad p_2q_2 \end{aligned}$$

At a particular generation, the alleles of different genes may fail to be in random association, for example $\Pr(A_1B_2) \neq p_1q_1$, but given the HW assumptions, the state of linkage equilibrium will be attained after several generations. The rate of approaching to the state of equilibrium may be very slow, and as we will show later it depends on the recombination rate between the loci under study.

Let us consider the genotype A_1B_1/A_2B_2, this genotype can produce four types of gametes, carrying haplotypes: $1)A_1B_1$; $2)A_2B_2$; $3)A_1B_2$ $4)A_2B_1$.

Haplotypes $1, 2$ are non-recombinant haploids, because the alleles are in the same arrangement as they were in the parental chromosome, and haplotypes $3, 4$ are recombinant types, and they are the result of a crossover and recombination between two genes A, B. In other words, during the process of meiosis a break has happened at a point between genes A and B and as the result of recombination the segment of the chromosome containing A_1 has recombined with the segment of the sister chromosomes containing B_2, therefore the recombinant type contains A_1 and B_2.

The recombination rate between genes depends on whether they are present on the same chromosome and, if so, on the physical distance between them. If two genes are located on two different chromosomes the recombination fraction is $r = 1/2$; Indicating that the probability that the two genes preserve their arrangement, is 0.5, implying no correlation between segregation of two genes (and hence no actual physical attachment between them). There is a direct relationship between recombination rate between two linked genes and their physical distance on the chromosome [30, 31], if they are far apart then $r = 1/2$, meaning that they easily will be separated during the formation of gametes in meiosis and in fact they behave as they are located on two different chromosome (note that when two genes are far apart from each other there is a high probability of a crossover event between them), but when they are close to each other, the probability that a crossover event happens at a point between the two is low, and when they are very close to each other, the probability of such an event is almost zero. Hence in this case $r = 0$, which is the smallest possible recombination rate. This property enables us to use the recombination rate as a measure of distance between two genes and hence knowing the value for r and location of one of the genes on the chromosome we can estimate the location of the other gene. Therefore an individual with

the genotype A_1B_1/A_2B_2 will produce the following haplotypes with the proportions indicated below:

The two non-recombinant haplotypes

A_1B_1 with frequency $(1-r)/2$

A_2B_2 with frequency $(1-r)/2$

The two recombinant haplotypes

A_1B_2 with frequency $r/2$

A_2B_1 with frequency $r/2$

Now, let us see how the frequency of a haplotype will change over time. For two genes with two alleles we have four possible haplotypes namely $A_1B_1; A_1B_2; A_2B_1; A_2B_2$. For simplicity let us denote their frequencies as $P_{11}; P_{12}; P_{21}; P_{22}$ respectively, where $P_{11} + P_{12} + P_{21} + P_{22} = 1$, and let us denote the corresponding frequencies in the next generation by $P'_{11}; P'_{12}; P'_{21}; P'_{22}$. For now let us focus on the A_1B_1 haplotype. In any generation a chromosome carrying A_1B_1 either could have undergone recombination between the genes (an event with probability r), or could have not experienced a recombination event (with probability $1-r$). Among the A_1B_1 chromosomes that did not undergo recombination, the frequency of A_1B_1 is the same as it was in the previous generation; among the chromosomes that did undergo recombination the frequency of A_1B_1 chromosomes is simply the frequency of $A_iB_1/A_1B_j, \quad i,j = 1,2$ genotypes in the previous generation; Because mating is random, the overall frequency of A_iB_1/A_1B_j genotypes is p_1q_1, therefore the frequency of type A_1B_1 in the next generation, denoted by P', is:

$$P'_{11} = (1-r)P_{11} + rp_1q_1 \tag{2.8}$$

Now if we subtract p_1q_1 from both sides of (2.8), we will have:

$$P'_{11} - p_1q_1 = (1-r)(P_{11} - p_1q_1) \tag{2.9}$$

Let us introduce a new term D defined as:

$$D \triangleq P_{11} - p_1q_1$$

Hence equation (2.9) will be written as:

$$D' = (1-r)D \tag{2.10}$$

In which D' is the value for D at the next generation. In general if we define $D_t \triangleq P_{11}^{(t)} - p_1q_1$ we have

$$D_t = (1-r)D_{t-1} = (1-r)^2D_{t-2} = \cdots = (1-r)^tD_0 \tag{2.11}$$

Where D_0 is the value of D in the founding population. As $0 \leq r \leq 1/2$, $(1-r)^t$ goes to zero as t increases, therefore as the number of generations increases $D_t = P_{11} - p_1q_1$,decreases hence:

$$\lim_{t\to\infty} D_t = 0; \tag{2.12}$$

Hence when t is large, $P_{11} = p_1q_1$ meaning that the linkage equilibrium has been attained. We can consider D as a measure of linkage disequilibrium, indicating deviation of P_{11} from the product of p_1q_1. And as we see from equation (2.11), linkage disequilibrium will decay as the number of generation increases. D measure is a good measure of linkage disequilibrium when we have two loci. The rate by which D measure decays depends on the value of r, i.e. if r is closer to $1/2$ (when two genes are far apart, D goes to zero very fast (few generations) but as r become smaller (genes are closer to each other on the chromosome), D will decay very slowly, i.e. it takes many generations

until population reaches the state of equilibrium with respect to these two genes.
The value of D that holds for $P_{11} - p_1 q_1$ also holds for the other possible haplotypes:

$$\begin{aligned} P_{11} &= p_1 q_1 + D \\ P_{12} &= p_1 q_2 - D \\ P_{21} &= p_2 q_1 - D \\ P_{22} &= p_2 q_2 + D \end{aligned} \tag{2.13}$$

If we calculate D from (2.13) in terms of haplotype frequencies we have:

$$D = P_{11} P_{22} - P_{12} P_{21} \tag{2.14}$$

As we will see in Chapter 4), we can define a general measure of LD for an arbitrary number of loci, and D measure will be a special case of that measure.
If we consider loci on a chromosome as random variables (see Chapter 3), taking values from the space of possible alleles, linkage equilibrium, can be viewed as the state in which all these random variables are independent of each other, and linkage disequilibrium can be viewed as departure from this state. This concept, is the basis of our future discussions, in Chapter 4 we will discuss different measures of linkage and linkage disequilibrium in populations with respect to an arbitrary number of loci.

Chapter 3

Mathematical Background

In this chapter, we provide an introduction to the mathematical concepts that will be used in subsequent chapters. Our approach to probability theory will be a measure theoretic approach, in the sense that probability is considered as an additive measure defined on the probability space. In Sections 3.1 and 3.2 we will introduce the basic concepts in measure theory and probability theory, in Section 3.3 we will introduce graphical models for representing independence of random variables, and Section 3.4 is a brief introduction to the concept of random sets which in Chapter 4 will be used as an alternative mathematical model to describe haplotype probabilities.

3.1 Measure, Measurable Space and Measurable Function

Let $\aleph$ be an arbitrary set. A collection $\mathcal{A}$ of subsets of $\aleph$ is called a *σ-algebra* on $\aleph$ if:

1. $\aleph \in \mathcal{A}$,

2. For each set A that belongs to $\mathcal{A}$, the set $A' \triangleq \aleph - A$ also belongs to $\mathcal{A}$,

3. For each countable sequence $\{A_i\}$ of sets that belong to $\mathcal{A}$ the set $\cup_{i=1}^{\infty} A_i$ belongs to $\mathcal{A}$

Therefore a σ-algebra on $\aleph$ is a family of subsets of $\aleph$ that contains $\aleph$ itself and is closed under complementation, formation of countable unions and countable intersections [10]. When $\mathcal{A}$ is a σ-algebra on the set $\aleph$, any subset of $\aleph$, which belongs to $\mathcal{A}$, is called *$\mathcal{A}$-measurable.*

Definition 3.1.1. *Borel σ-algebra*

The Borel σ-algebra on $\mathbb{R}^n$, is defined as the σ-algebra on $\mathbb{R}^n$, which is the smallest σ-algebra that includes the collection of open subsets of $\mathbb{R}^n$. We denote a Borel σ-algebra on $\mathbb{R}^n$ by $\mathcal{B}(\mathbb{R}^n)$[11].

Definition 3.1.2. *Countable Additivity*

Let $\aleph$ be a set, and let $\mathcal{A}$ be a σ-algebra on $\aleph$. And let $\{A_i\}$ be a sequence of disjoint members of $\mathcal{A}$, i.e. $A_i, A_j \in \mathcal{A}$, $A_i \cap A_j = \emptyset, \forall i \neq j$. A function μ which takes its values from the σ-algebra $\mathcal{A}$ and returns numbers belonging to the extended half line $[0, +\infty]$ is said to be countably additive if it satisfies [10]:

$$\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i) \tag{3.1}$$

Definition 3.1.3. *Measure*

A measure, (or a countably additive measure) on $\mathcal{A}$ is a function $\mu : \mathcal{A} \rightarrow [0, +\infty]$ that satisfies [10]:

1. $\mu(\emptyset) = 0$

2. μ is countably additive (satisfies (3.1))

Definition 3.1.4. *Measurable Space*

If $\aleph$ is a set, and $\mathcal{A}$ is a σ-algebra on $\aleph$, and if μ is a measure on $\mathcal{A}$, then the triple $(\aleph, \mathcal{A}, \mu)$ is often called a measure space. The pair $(\aleph, \mathcal{A})$ is often called a measurable space [10].

Definition 3.1.5. *Measurable Function*

Let $(\aleph_1, \mathcal{A}_1), (\aleph_2, \mathcal{A}_2)$ be two measurable spaces, and let $f : \aleph_1 \rightarrow \aleph_2$, then the function

f is called $\mathcal{A}_1/\mathcal{A}_2$-measurable, if [11]:

$$\forall A \in \mathcal{A}_2, f^{-1}(A) \in \mathcal{A}_1, \tag{3.2}$$

in which the inverse function $f^{-1}(\cdot)$ is defined as

$$f^{-1}(A) \triangleq \{\alpha \in \aleph_1 : f(\alpha) \in A\} \tag{3.3}$$

3.2 Probability Space and Probability Measure

The degree of likelihood of an event, can be represented mathematically as an additive measure defined on the probability space [19, 11]. A probability space can be defined as follows,

Definition 3.2.1. *Probability Space*

Consider the measurable space $(\Omega, \mathcal{F})$, in which Ω is the sample space (the set of all possible outcomes of a random experiment) and $\mathcal{F}$ is the σ-algebra defined on Ω.

A mapping $P : \mathcal{F} \to [0, 1]$ satisfying the following conditions:

1. P is a countably additive measure defined on $(\Omega, \mathcal{F})$
2. $P(\Omega) = 1$

is called a probability measure. The triple $(\Omega, \mathcal{F}, P)$ is called the *Probability Space* [11].

3.2.1 Random Variable

The outcome of a random experiment usually is not a real number, for example if we consider the experiment of n trials of tossing a coin, the outcome takes values from the set $\Omega = \{H, T\} \times \cdots \times \{H, T\} = \{H, T\}^n$. But usually we are interested in some numerical attributes to the outcome, for example the number of heads for each $\omega \in \Omega$. A random variable is a measurable function X, which assigns a number or a vector to each outcome of a random experiment, more precisely we can define a random variable as [11]:

Definition 3.2.2. *Random Variable*

Consider the probability space $(\Omega, \mathcal{F}, P)$, a measurable function $X : \Omega \to \mathbb{R}^n$, is called a n-dimensional random variable:

$$X : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$$

3.2.2 The Distribution Function and The Density Function

By defining a random variable on a probability space $(\Omega, \mathcal{F}, P)$ we will have an induced probability space by random variable X, which is: $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_X)$. Where the induced probably measure by the random variable X is defined as:

$$P_X(B) = P\{\omega \in \Omega : X(\omega) \in B\}, \quad \forall B \in \mathcal{B}(\mathbb{R}^n) \tag{3.4}$$

For now let us consider the case of $n = 1$, hence we restrict our attention to measure space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

We can describe all the events of interest in terms of open subsets of $\mathbb{R}$, which in general have the form: $A = \{x : a < x < b\}$, where a, b are elements of $\mathbb{R} \cup \{-\infty, +\infty\}$. The induced probability measure P_X by random variable X, can be described in terms of a distribution function which is defined as follows [?, 11]:

Definition 3.2.3. *The Cumulative Distribution Function (cdf)*

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a random variable defined on $(\Omega, \mathcal{F})$, the function $F_X : \mathbb{R} \to [0, 1]$ defined as:

$$F_X(x) \triangleq P(\{\omega \in \Omega : X(\omega) \leq x\}) \tag{3.5}$$

Is called the cumulative distribution function (cdf) of the random variable X [42]. From the definition of a probability measure and a random variable defined on the measurable space $(\Omega, \mathcal{F})$, we can show that the cdf of a random variable satisfies the following conditions:

1. $0 \leq F_X(x) \leq 1$

2. $\lim_{x\to\infty} F_X(x) = 1$

3. $\lim_{x\to-\infty} F_X(x) = 0$

4. if $a < b$, then $F_X(a) \leq F_X(b)$

5. for $h > 0$, $F_X(b) = \lim_{h\to 0} F_X(b+h) = F_X(b^+)$

Definition 3.2.4. *Probability Density Function*

For some common distribution functions (such as Gaussian distribution function) we can define the derivative with respect to variable x. The probability density function (pdf), denoted by $f_X(x)$, of a *continuous* random variable, is defined as the derivative of $F_X(x)$ (if a derivative exists) with respect to x [42]:

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{3.6}$$

And also by definition we can define a cdf in terms of pdf as follows:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \tag{3.7}$$

The pdf, represents the probability measure of a small open ball with center x, i.e. the probability that X is in a small interval $(x, x+\Delta x)$ is:

$$P(\{x < X \leq x+\Delta x\}) = F_X(x+\Delta x) - F_X(x) = \frac{F_X(x+\Delta x) - F_X(x)}{\Delta x}\Delta x \tag{3.8}$$

When Δx is very small then:

$$P(\{x < X \leq x+\Delta x\}) \approx f_X(x)\Delta x \tag{3.9}$$

We notice that always we can calculate the probability measure of a Borel set $B \in \mathcal{B}(\mathbb{R})$ induced by random variable X, in terms of the cdf of X [11]. For example in the case

of continuous random variable X, consider the open interval (a, b), $b > a$, then:

$$\begin{aligned} P_X(a < x < b) &= P(\{\omega : a < X(\omega) < b\}) = 1 - P(\{\omega : a < X(\omega) < b\}^c) \\ &= 1 - P(\{\omega : a \geq X\} \cup \{\omega : X \geq b\}) \\ &= 1 - [P(\{\omega : X < a\}) + (1 - P\{\omega : X < b\})] \\ &= F_X(b) - F_X(a) \end{aligned} \tag{3.10}$$

In which we have used the additivity property of probability measure over disjoint unions. If a pdf exists, we can describe the probability measure of any Borel set, i.e. $P_X(a < x < b)$, in terms of the pdf as:

$$P_X(a < x < b) = \int_a^b f_X(t)dt \tag{3.11}$$

Definition 3.2.5. *Discrete Random Variable*

A random variable X is said to be discrete, if there exist a countable set $S = \{x_i\}$ where:

$$\sum_{x_i \in S} P(\{\omega; X(\omega) = x_i\}) = 1 \tag{3.12}$$

For discrete random variables, we can assign a probability mass function (pmf) to each point $x_i \in S$, represented by $p_X(x_i)$ satisfying:

$$\sum_i p_X(x_i) = 1 \tag{3.13}$$

We can define the pdf of a discrete random variable as follows:

$$f_X(x) = \sum_i p_X(x_i)\delta(x - x_i) \tag{3.14}$$

In which $\delta(x - x_i)$ is the Dirac delta function, defined as follows:

Definition 3.2.6. *Dirac Delta Function*

Let $u(x)$ be the unit step function defined as:

$$u(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

Then the delta function $\delta(t)$ is defined in terms of the unit step function by the following equation:

$$u(x) = \int_{-\infty}^{x} \delta(t)dt \tag{3.15}$$

3.2.3 Some Common Discrete Random Variables

Discrete random variables are important in situations where counting is involved (for example the number of cross-over events along a chromosome). In this section, we introduce two specific discrete random variables, which will be used in following chapters.

Definition 3.2.7. *The Bernoulli RV*

Consider the measure space $(\Omega, \mathcal{F})$. Let $A \in \mathcal{F}$. The indicator function $I_A(\omega)$ is a random variable defined by the following equation [11]:

$$I_A(\omega) \begin{cases} 1 & \text{if} \quad \omega \in A \\ 0 & \text{if} \quad \omega \notin A \end{cases} \tag{3.16}$$

$I_A(\omega)$ is a discrete random variable with range $S_X = \{0, 1\}$ and its pmf is:

$$p_I(1) = p; \quad \text{and} \quad p_I(0) = 1 - p \tag{3.17}$$

where $p \triangleq P(\{\omega : \omega \in A\})$

Definition 3.2.8. *: The Poisson Random Variable*

In many applications we are interested to count the number of occurrences of a random event in a certain period of time or in a certain region in space. The Poisson random variable is a random variable representing the number of occurrences of a completely random event, in a certain interval. The pmf of the Poisson RV is defined as:

$$p_N(k) \triangleq P(N = k) = \frac{\alpha^k}{k!} e^{-\alpha}, \quad k = 0, 1, 2, \ldots \tag{3.18}$$

where α is the average number of event occurrences in a specified interval [42].

3.2.4 Joint Probability

In many applications, we are interested to know the probability of the joint occurrence of several events. For example let X, Y be two random variables defined on $(\Omega, \mathcal{F}, P)$. We represent the joint probability of (X, Y) by $P_{XY}(x, y)$ or for short $P(X, Y)$. Which is the probability of simultaneous occurrence of $X = x, Y = y$. In terms of the underlying probability space, $P(X, Y)$ can be defined as [11]:

$$P(X = x, Y = y) \triangleq P(\{\omega \in \Omega : \omega \in A \cap B\}), \quad A, B \in \mathcal{F}, A = X^{-1}(x), B = Y^{-1}(y) \tag{3.19}$$

We can extend this definition for m random variables defined on $(\Omega, \mathcal{F}, P)$:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m) = P(\{\omega \in A_1 \cap A_2 \cap \cdots \cap A_m\}), \quad A_i = X_i^{-1}(x_i) \tag{3.20}$$

3.2.5 Conditional Probability and Independence

Another important probability measure which describes the relation between two or more random variables is the conditional probability measure. Suppose X, Y are two random variables defined on the probability space $(\Omega, \mathcal{F}, P)$. The conditional probability measure $P_{X|Y}(X = x|Y = y)$ or for short $P(X|Y)$, is defined as the probability that $X = x$ knowing that $Y = y$. In terms of the underlying probability space we can define the conditional probability as:

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}, \quad A, B \in \mathcal{F} \tag{3.21}$$

In terms of random variables X, Y we have:

$$P(X|Y) = \frac{P(X, Y)}{P(Y)} \tag{3.22}$$

We can use equation (3.22) to define joint probability of X, Y in terms of the conditional probability as:

$$P(X, Y) = P(X|Y)P(Y) \tag{3.23}$$

Equation (3.23) can be generalized for m random variables $X_1, \ldots, X_m$ as:

$$P(X_1, \ldots, X_m) = P(X_m|X_1, \ldots, X_{m-1})P(X_{m-1}|X_1, \ldots, X_{m-2}) \ldots P(X_2|X_1)P(X_1) \tag{3.24}$$

Definition 3.2.9. *Independence of Random Variables*

We say two random variables X, Y are independent of each other, represented by $X \perp Y$, if and only if [42]:

$$P(X|Y) = P(X) \tag{3.25}$$

From (3.25) and (3.23) we can define independence of two random variables as:

$$P(X, Y) = P(X)P(Y) \tag{3.26}$$

This can be generalized for m random variables. Hence we say random variables $X_1, \ldots, X_m$ are independent of each other if and only if:

$$P(X_1, \ldots, X_m) = P(X_1)P(X_2) \ldots P(X_m) \tag{3.27}$$

Definition 3.2.10. *Conditional Independence*

Consider random variables X, Y, Z, we say that X is independent of Y conditioned on Z, $X \perp Y|Z$, if and only if [42, 35]:

$$P(X|Y, Z) = P(X|Z) \tag{3.28}$$

or:

$$P(X, Y|Z) = P(X|Z)P(Y|Z) \tag{3.29}$$

Let I be any subset of the index set $S = \{1, 2, \ldots, m\}$. We represent the joint probability $P(X_{I_1}, \ldots, X_{I_k})$ by $P(X_I)$, where $k = |I|$ is the cardinality of I. Knowing the joint

probability $P(X_S) = P(X_1, \ldots, X_m)$ we can calculate $P(X_I)$ for $\forall I \subset S$ by summing over all possible values of random variables X_j where $j \in S - I$. Hence

$$\begin{aligned} P(X_I) &= \sum_{X_{J_1},\ldots,X_{J_l}} P(X_1, \ldots, X_m), \quad J = S - I, l = |J| \\ &\stackrel{\nabla}{=} \sum_{X_{S-I}} P(X_S) \end{aligned} \tag{3.30}$$

For example, in the special case of $m = 2$, $P(X_1) = \sum_{X_2} P(X_1, X_2)$. We can also calculate the marginal probability with respect to the conditional probability, using (3.23) we have:

$$P(X_1) = \sum_{X_2} P(X_1|X_2)P(X_2) \tag{3.31}$$

3.2.6 The Expected Value of Random Variables

The expected value or mean of a random variable X is defined by :

$$E(X) = \int_{-\infty}^{+\infty} t f_X(t) dt \tag{3.32}$$

If X is a discrete random variable, from (3.14) and (3.32) we have:

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} t \sum_k p_X(t)\delta(t - x_k)dt \\ &= \sum_k x_k P_X(x_k) \end{aligned} \tag{3.33}$$

As an example, consider the Poisson random variable, the expected value of the Poisson random variable is:

$$\begin{aligned} E(N) &= \sum_{k=0}^{\infty} k P_N(k) = \sum_{k=0}^{\infty} k \frac{\alpha^k}{k!} e^{-\alpha} \\ &= \sum_{k=0}^{\infty} \frac{\alpha^k}{(k-1)!} e^{-\alpha} = \sum_{k=0}^{\infty} \frac{\alpha^{k-1}\alpha}{(k-1)!} e^{-\alpha} \\ &= e^{-\alpha}\alpha \sum_{k=1}^{\infty} \frac{\alpha^{k-1}}{(k-1)!} = e^{-\alpha}e^{\alpha}\alpha = \alpha \end{aligned} \tag{3.34}$$

Which is compatible with the definition of constant α as the average occurrence of a Poisson random variable.

We can generalize the concept of expected value of a random variable to any function g of a random variable, hence in general we have:

$$E(g(X)) = \int_{-\infty}^{+\infty} g(t) f_X(t) dt \tag{3.35}$$

and

$$E(g(X)) = \sum_k g(x_k) P_X(x_k) \tag{3.36}$$

for discrete random variables.

3.3 Graphical Representation of Random Variables

Consider the set of random variables $\{X_1, \ldots, X_m\}$ defined on the probability space $(\Omega, \mathcal{F}, P)$. Suppose we want to find the joint probability $P(X_I)$ where $I \subset S = \{1, \ldots, m\}$. In general, if there is no independence relation between some of these random variables from (3.24) we have:

$$P(X_I) = P(X_{I_1}, \ldots, X_{I_l}) = P(X_{I_1}) \prod_{i=2}^{l} P(X_{I_i} | X_{I_{i-1}}, \ldots, X_{I_1}) \tag{3.37}$$

But if somehow we know that there is some independence relationship among these random variables, we can simplify equation (3.37), by removing some of the extra conditionings. For example consider the case $m = 3$, and suppose that we have the following independence relationship among random variables X_1, X_2, X_3:

$$\begin{aligned} X_2 &\perp X_1 \\ X_3 &\perp X_1 | X_2 \end{aligned} \tag{3.38}$$

Then the joint probability $P(X_1, X_2, X_3)$ can be written as:

$$P(X_1, X_2, X_3) = P(X_3 | X_1, X_2) P(X_2 | X_1) P(X_1) = P(X_3 | X_2) P(X_2) P(X_1) \tag{3.39}$$

We can consider this product as factorizing the joint probability in terms of some *local* functions [35]:

$$h_1(X_3, X_2) = P(X_3|X_2); \quad h_2(X_2) = P(X_2); \quad h_3(X_1) = P(X_1)$$

Usually for representing the independence relationship among a set of random variables we use graphical models, in which the nodes of the graphs are representatives of the random variables and the edges connecting the nodes are representatives of dependency among the nodes [35, 41].

There are two types of graphical representations: *directed graphs* and *undirected graphs* [35]. In the following sections we introduce these models and their properties.

3.3.1 Bayesian Networks

First let us consider the directed graphical representations. A directed graph is a pair $(\mathcal{V}, \mathcal{E})$ (also represented by $\mathcal{G}(\mathcal{V}, \mathcal{E})$), where $\mathcal{V}$ is a set of nodes and $\mathcal{E}$ is a set of (oriented) edges. In order to use the graph for representing random variables, we consider a one-to-one mapping from nodes to random variables, and we say that the random variables are indexed by the nodes in the graph. Therefore for each $i \in \mathcal{V}$, there exist a random variable X_i. Hence the associate set of nodes for random variables $\{X_1, \ldots, X_m\}$, will be $\mathcal{V} = \{1, \ldots, m\}$

If there is a directed edge from node i to node j we say that node i is parent of node j (and j is the child of node j) [35]. Each node has a set of parent nodes, which can be empty set. We denote the set of parents of nodes i by π_i. To each pair (i, π_i) we assign a function $f(i, \pi_i) = P(x_i|x_{\pi_i})$.

A path between nodes $i, j \in \mathcal{V}$ is a sequence of directed edges starting at node i and ending at node j. If there exist a directed path starting at a node and ending at the same node, we say that the graph has a cycle, a directed graph without any cycle is called a *directed acyclic graph* or *DAG* [35].

Definition 3.3.1. *Bayesian Networks*

A Bayesian network is a *DAG*, $\mathcal{G}(\mathcal{V}, \mathcal{E})$, in which nodes $i \in \mathcal{V}$ represents random variables X_i, and the joint probability $P(X_\mathcal{V})$ can be factorized to local conditional probabilities as:

$$P(X_1, \ldots, X_v) \triangleq P(X_\mathcal{V}) = \prod_{i \in \mathcal{V}} P(X_i | X_{\pi_i}) \tag{3.40}$$

where π_i is the set of parents of node X_i, and $v = |\mathcal{V}|$ is the number of nodes.

In a Bayesian network, each node is conditionally independent of nodes which are not descendants given the value of its parents. Figure (3.1), represents a Bayesian network, with $\mathcal{V} = \{1, 2, 3, 4, 5\}$ representing random variables $X_1, \ldots, X_5$. For this network we have:

$$\pi_1 = \emptyset, \quad \pi_2 = 1, \quad \pi_3 = \{1\}, \quad \pi_4 = \{2\}, \quad \pi_5 = \{3, 4\} \tag{3.41}$$

and joint probability $P(X_1, \ldots, X_5)$ can be written as:

$$P(X_1, \ldots, X_5) = \prod_{i}^{5} P(X_1 | X_{\pi_i}) = P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2)P(X_5|X_3, X_4) \tag{3.42}$$

As another example of a Bayesian network we can consider Markov chains, which are special cases of Bayesian networks, Figure 3.2 shows a Markov chain for three random variables X_1, X_2, X_3, $\pi_1 = \emptyset, \pi_2 = \{1\}, \pi_3 = \{2\}$, therefore:

$$P(X_3|X_2, X_1) = P(X_3|X_2) \tag{3.43}$$

If we consider the indices to represent time, it can be viewed as the property of a Markov process in which *future* is independent of the *past* given the *present* state [42].

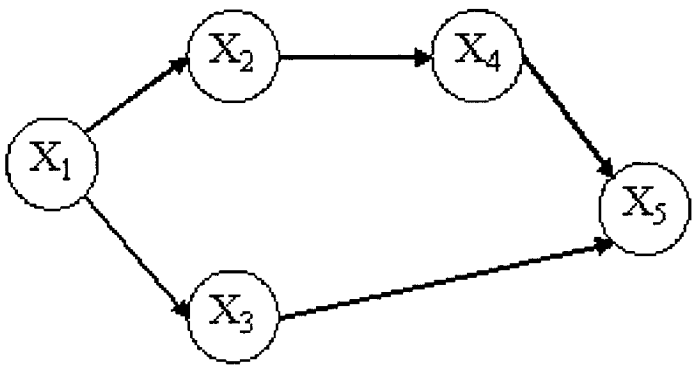


Figure 3.1: A Bayesian network representation for random variables $X_1, \ldots, X_5$

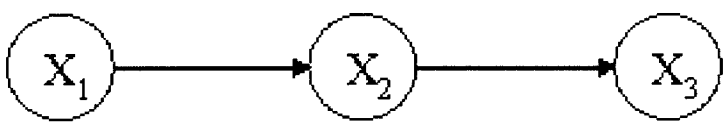


Figure 3.2: A Bayesian network representation for the Markov chain X_1, X_2, X_3

3.3.2 Markov Random Fields

In this section we introduce the undirected graphical models for representing random variables, known as *Markov random fields* or MRF for short.

An undirected graphical model, is a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of nodes that are in one-to-one correspondence with a set of random variables, and $\mathcal{E}$ is a set of undirected edges. We are interested to finding the joint probability of $X_1, \ldots, X_m$ in terms of multiplication of local functions defined on subsets of $\mathcal{V}$. In the case of undirected graphs, unlike the Bayesian networks it is not very easy to consider these local functions as conditional probabilities (due to the problem of consistency of conditional probabilities [35], rather we define these functions as arbitrary functions with the constraint that their product to be equal the joint probability of $P(X_\mathcal{V})$. In order to be able to factor the joint probability in terms of local functions first we need to define a topology on the graph, by which we define the connections between the nodes [35].

Definition 3.3.2. *Neighborhoods*

A neighbourhood system on $\mathcal{V}$ is a family $N = \{\mathcal{N}_v\}, v \in \mathcal{V}$ of subsets of $\mathcal{V}$ such that for all $v \in \mathcal{V}$:

1. $v \notin \mathcal{N}_v$

2. $u \in \mathcal{N}_v \Rightarrow v \in \mathcal{N}_u$

For a Markov random field corresponding to a given set of random variables, the neighbourhood system is defined in a way to satisfy the following condition:

$$P(X_v | X_{\mathcal{V}-\{v\}}) = P(X_v | X_{\mathcal{N}_v}) \tag{3.44}$$

In the other words, the Markov property implies that the probability of any realization of a node, given its neighbours, is independent of the rest of the nodes in the graph [34, 35].

Definition 3.3.3. *Clique*

A *clique* in an undirected graph, is a set of nodes which are mutually connected to each other, in other words, a clique is a fully-connected subset of the nodes.

Definition 3.3.4. *Maximum Clique*

The maximum cliques of a graph are the cliques that cannot be extended to include additional nodes without losing the property of being fully connected [35]. Therefore all the cliques are subsets of maximal cliques.

Definition 3.3.5. *Potential Function*

Let C be a set of indices of a maximal clique in an undirected graph G, and let $\mathcal{C}$ be the set of all such C. A *potential function*, $\psi_{X_C}(x_C)$, is a function on the possible values of $x_C = (x_{C_1}, \ldots, x_{C_{|C|}})$, satisfying $\psi_{X_C}(x_C) \in \mathbb{R}^+$.

The joint probability $p(X_\mathcal{V})$ can be factored as:

$$P(X_\mathcal{V}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_{X_C}(x_C) \tag{3.45}$$

where Z is a normalization factor defined as:

$$Z \triangleq \sum_{x} \prod_{C \in \mathcal{C}} \psi_{X_C}(x_C) \tag{3.46}$$

Figure 3.3, represents a MRF with six nodes, in which nodes X_2, X_5, X_6 constitute a clique. In this section we introduced two main categories of graphical models to represent random variables, namely Bayesian networks and Markov random fields. And we see that in both cases knowing the structure of the graph we can factorize the joint probability distribution $P(X_1, \ldots, X_m) \triangleq P(X_\mathcal{V})$ in terms of a product of local functions, which in the case of Bayesian networks the local functions have the form $P(X_i|\pi_i)$, and for Markov random fields these factors are potential function $\psi_C(X_C)$, defined on the cliques of the graph.

Graphical models have been widely used in genetic analysis especially for representing

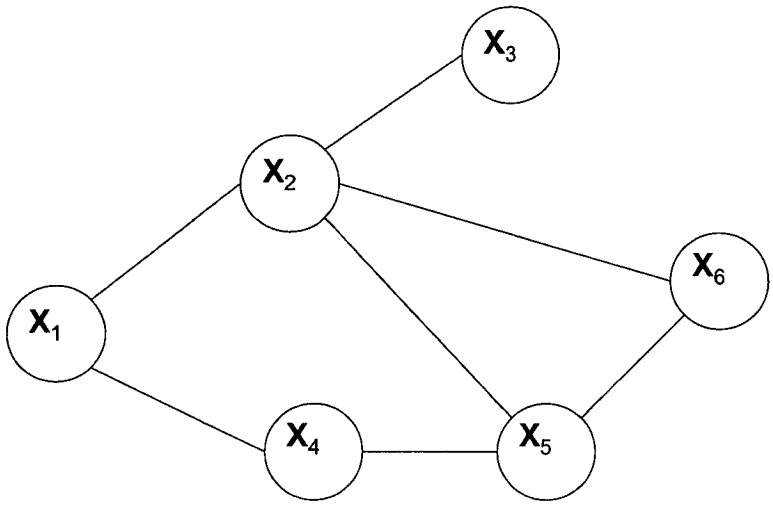


Figure 3.3: A MRF representation for six random variables $X_1, \ldots, X_6$

pedigrees and modeling the segregation of genes [41, 69]. A detailed review on graphical models in genetics can be found in Lauritzen [41]. In Section 5.5 we will use a Markov random field model for modeling blocks of genes which have remained intact from an ancestral population.

3.4 Random Sets

We can generalize the concept of random variables introduced in Section 3.2.1, to a more general concept of *random sets*. In this case instead of mapping an event $\omega \in \Omega$ into a variable $X \in \mathbb{R}^n$, we map the event ω, into a subset of some set S [45]. As a simple example let us consider S to be a finite countable set. And let us denote the power set of S (the collection of all subsets of S) by $\mathcal{P}(S)$. If we assign an element $A \in \mathcal{P}(S)$, to each $\omega \in \Omega$, then A is random set corresponding to the event ω.

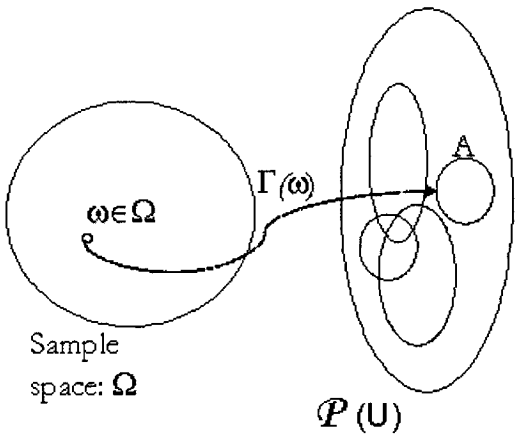


Figure 3.4: A representation of assigning a random set

Definition 3.4.1. *Random Element*

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $(U, \mathcal{U})$ be a measurable space. A random element Z is a map:

$$Z : \Omega \to U$$

which is $\mathcal{F}/\mathcal{U}$-measurable, i.e.

$$X^{-1}(B) \in \mathcal{F}, \quad \forall B \in \mathcal{U}$$

When $U = \mathbb{R}^n, \quad n = 1, 2, \ldots$ then $Z \stackrel{\nabla}{=} X$ is a random variable and when $U = \mathcal{P}(\Theta)$, for some set Θ, then $Z \stackrel{\nabla}{=} \Gamma$ is a random set. Therefore random sets are set valued random elements [45, 24].

Definition 3.4.2. *Probability Distribution Function for Random Sets*

Probability distribution function or for random set $\Gamma(\omega)$ is defined as:

$$\begin{aligned} &F : \mathcal{P}(U) \to [0, 1] \\ &F(A) = P(\{\omega \in \Omega : \Gamma(\omega) \subseteq A\}), \quad \forall A \in \mathcal{U} \end{aligned} \tag{3.47}$$

Let us consider the probability that $\Gamma(\omega) = A, \quad A \in \mathcal{U}$. We denote this probability by $f(A)$. The function $f : \mathcal{U} \to [0, 1]$ can be considered as the probability mass function

of the random set $A \in \mathcal{U}$, and satisfies the following condition:

$$\sum_{A \in \mathcal{U}} f(A) = 1 \tag{3.48}$$

The density function $f(A)$ is related to the probability distribution function as follows [24]:

$$F(A) = \sum_{B \subseteq A} f(B) \tag{3.49}$$

If the function $F(A)$, satisfies $F(\emptyset) = 0$, then it is often referred to as *belief* function for random set A [61].

3.4.1 Möbius Transform

Möbius transformation is a powerful tool when we deal with set functions. It can be shown (see [35]) that if g and f are mappings from $\mathcal{P}(U)$ to $\mathbb{R}$ then the following two statements are equivalent:

1. $g(A) = \sum_{B \subseteq A} f(B)$
2. $f(A) = \sum_{B \subseteq A} (-1)^{|A-B|} g(B)$

in which $|A - B|$ is the number of elements of $A - B$.

By using Möbius transform we can write equation (3.49) as:

$$f(A) = \sum_{B \subseteq A} (-1)^{|A-B|} F(B) \tag{3.50}$$

which describes the relation between distribution function and mass function of a random set [24].

3.4.2 Commonality Measure

Another set function of interest in random-set theory is the commonality number or commonality measure [61]. Commonality number of a random set A is defined as:

$$Q(A) = \sum_{B \supseteq A} f(B), \quad \forall A \in \mathcal{U} \tag{3.51}$$

Commonality measure is related to the probability distribution function of the random set as follows [61]:

$$F(A) = \sum_{B \subseteq A^c} (-1)^{|B|} Q(B) \tag{3.52}$$

In which A^c is the complement of the set A, i.e. $U - A$

In Chapter 4 we will use random sets to model the set of haplotypes in a population in terms of subsets of the set of loci under study, and we will show that the recursion equation describing change of haplotype distributions as described by Dawson [12], can be reformulated in terms of commonality measure of random subsets of the set of loci under study [58].

Chapter 4

Discrete Time Model

As we discussed in Chapter 2, evolutionary forces such as mutation, recombination and natural selection are the main sources of change in genetic composition of populations [22]. In Section 2.5.5, we discussed the equation describing change of haplotype frequencies for a haplotype of length 2, i.e., two loci (equations (2.9),(2.11)). In this chapter, we will discuss models for describing change of haplotype distributions, extended to an arbitrary number of loci. In this model, the only source of change of haplotype frequencies is considered to be recombination and the other evolutionary forces such as mutations and natural selection are not taken into account [12]. The assumptions for mating patterns and population size are the same as the Hardy-Weinberg model discussed in Section 2.5.5 [33]. One of the main problems in generalizing the HW model to an arbitrarily large number of loci, is solving the equation describing the change of probabilities, in terms of haplotype frequencies in the founding population [23, 4].

For the case of two loci, we showed in Section 2.5.6 [30] that if we describe the state of the system in terms of D measure for LD, we can relate the state of population at any generation to its initial state, which is the state of the founding population [30]. The main idea in this chapter is generalizing this method for an arbitrary number of loci.

One of the mathematical models in theoretical population genetics is the algebraic rep-

resentation of populations [32, 55, 59].[1] There are different algebras arising in genetics, such as special train algebras, genetic algebras, baric algebras and the most general model which is algebras with genetic realization [55]. In these models individuals (haplotypes or genotypes) are represented as bases for the algebra and the frequencies are coefficients of these bases, hence a population is uniquely identified by a linear combination of these bases. Random mating is represented by a product defined on the algebra. In this thesis, as the main focus is on stochastic modeling of populations , we will not focus on algebraic models. For more information about the relation between the models described in this chapter and genetic algebras see Geiringer [23], also for a review on the properties of genetic algebras see [55, 64].

In the following sections, we discuss the discrete time model (or non-overlapping generations model) which is based on the models developed by Geiringer [23], Bennett [4], and Lyubich [38] with a random-set theoretic interpretation of Dawson's model [12, 58].

In this model, generations are non-overlapping and population size is very large and is assumed to be constant. Also the effect of mutation and natural selection is assumed to be negligible.

A continuous time model with a measure theoretic approach developed by Baake 2001 and 2003 [2, 3], will be discussed in Chapter 5.

4.1 The Probabilistic Model

Here we focus on haploid individuals, and we represent the set of all individuals in the population by $\mathcal{X}$. Therefore, an element $x \in \mathcal{X}$ represents a haploid individual (i.e. a gamete). Let us denote the set of all loci in a gamete, by $G = \{0, 1, \ldots, g\}$. The

[1] Claude Shannon, who is the father of modern communication theory, wrote his PHD dissertation at MIT on developing an algebra for theoretical genetics; he developed his model independent of other mathematicians who at the time were working on genetic algebras [60].

configuration of alleles on the loci corresponding to any subset S of G is a haplotype. Let us represent possible alleles of a gene by integers $\Lambda = \{0, 1, \ldots m-1\}$ where m is the total number of possible alleles. We can uniquely represent any haplotype by the array $X_S = (X_1, \ldots, X_n)$ where $X_i \in \Lambda$ represents the allele on locus i, and the subscript S denotes the set of loci in the haplotype under study, and $n = |S|$ is the length of the haplotype [23, 12]

A population of gametes can be partitioned according to the haplotype under consideration, where individuals belonging to the same partition share the same arrangements of genes with respect to the specified region. The number of individuals in each partition divided by the population size, is the frequency of the corresponding haplotype. As the size of the population increases the frequency of haplotypes converges to its probability measure defined on the sample space [42], therefore as we consider a very large population, the frequency of a haplotype is almost equal to its probability measure. Let us denote the space of all haplotypes corresponding to the set of loci S, by $\mathcal{X}_S$, therefore the frequency of haplotype X_S is equal to $P(X_S = (X_1, \ldots, X_n))$ which $X_S \in \mathcal{X}_S$. From now on for simplicity we drop the subscript S, and simply refer to the sample space $\mathcal{X}_S$ as $\mathcal{X}$.

In this section, we focus on biallelic genes [12], hence in this case, $\Lambda = \{0, 1\}$, therefore we can represent a haplotype as a sequence of 0 and 1's, for example $X = (X_1, X_2, X_3) = (1, 0, 1)$, represents a haplotype of length 3 in which the gene on the first locus appears with its second allele, the second gene appears with its first allele and the third one with its second allele [12]. For a haplotype of length n , there are 2^n possible configurations. For instance, when there are 3 loci we have $2^3 = 8$ possible haplotypes denoted by :

$$(0,0,0), (0,0,1), (0,1,1), (0,1,0), (1,0,1), (1,1,1), (1,1,0)$$

As there are only two possible alleles, we can identify each configuration of alleles by the loci carrying allele '1'. For example a haplotype of length 3, with configuration $(0,0,0)$ can be uniquely identified by $\emptyset$, and haplotype $(1,0,0)$ by $\{1\}$ and $(1,0,1)$by $\{1,3\}$ and so on. This way we can set a one-to-one correspondence between all possible haplotypes and 2^n subsets of $S = \{1,\ldots,n\}$ [12]. This mapping can be considered as a random set $\Gamma : \mathcal{X} \to \mathcal{P}(S)$, which is a probabilistic mapping assigning a subset $U = \Gamma x$ of S to each haplotype $x \in \mathcal{X}$, such that Γx represents the set of loci carrying 1 allele, and the loci belonging to $S - \Gamma x$ carrying 0 alleles [58]. Therefore the probability of having a haplotype which all the loci are occupied by 1 allele is the probability that the haplotype $x \in \mathcal{X}$ is mapped to the set S. Let the probability that x is mapped to $\Gamma x = U \subseteq S$, to be denoted by $g_S(U)$, in which subscript S denotes the set of loci under consideration. $g_S(U)$ is the probability of the event that all the loci belonging to U are carrying 1 allele, and the rest, $S-U$, are carrying 0 alleles [12]. $g_S(U)$ represents the probability of haplotype X_S which is identified by the random set U. As we will see later, $g_S(U)$ is the probability mass function of the random set $U \subseteq S$ [58].

As we will discuss later, it is more useful to describe the change of haplotype probabilities in terms of the commonality measure (see Section 3.4.2 and [61]) of the random set corresponding to the haplotype. The commonality measure for the random set U, denoted by $Q(U)$ is:

$$Q(U) = \sum_{A:U\subseteq A\subseteq S} g_S(A) \tag{4.1}$$

$Q(U)$ is the probability of the event that in an arbitrary haplotype, a subset U of the loci is occupied by 1 allele. Knowing the value $Q(U)$, we can find the probability of haplotype corresponding to U, using the Möbius inverse transformation of (4.1), (see Section 3.4.1 and [35, 24]):

$$g_S(U) = \sum_{A:U\subseteq A\subseteq S} (-1)^{|A-U|} Q(A) \tag{4.2}$$

If we let $U = \emptyset$ in (4.1) then:

$$Q(\emptyset) = \Pr\{x \in \mathcal{X} : x \to \Gamma x, \emptyset \subseteq \Gamma x\} = \Pr\{x \in \mathcal{X}\} = 1 \tag{4.3}$$

Hence

$$Q(\emptyset) = \sum_{A:A\subseteq S} g_S(A) = 1 \tag{4.4}$$

As we see from (4.4), summing $g_S(A)$ over all subsets A of S we get one, hence $g_S(A)$ satisfies the condition for being a probability mass function for the random set U [58, 61].

4.1.1 Recombination Rates

The configuration of alleles in each generation depends on their configuration in the previous generation and the probability of transmitting certain set of loci from parents to children. In order to be able to describe the change of haplotype distributions in a population from one generation to the next we need to represent the transmission probabilities for each set of loci. These probabilities only depend on the loci and not on the alleles.

As a convention, we label loci inherited from the father by 1 and those inherited from the mother of the individual by 0 [12]. Now let us define a probabilistic mapping $\Gamma' : \mathcal{X} \to \mathcal{P}(S)$, where $\Gamma' x$ assigns a random set Γ' to each individual x such that $\Gamma' x$ denotes the set of loci on the haplotype x that have been inherited from the father of the individual labelled by 1 and the rest (maternal loci) from the mother labelled by 0. Now let us denote the probability that a haplotype x being mapped to $\Gamma' x = R \subseteq S$ by $r_S(R)$ then [58]:

$$r_S(R) \triangleq \Pr\{x \in \mathcal{X} : \Gamma' x = R \subseteq S\} \tag{4.5}$$

If the father of the individual transmits a subset R of loci, the rest of the genes on the child's haplotype, (genes on loci $S - R$), have to come from the mother. Therefore

probability that father transmits R is equal to the probability that the mother transmits $S-R$, but as we assume symmetry between paternal and maternal genes (we don't consider sex-linked genes), the probability that the mother transmits $S-R$ is equal to the probability that father transmits $S-R$ [12], therefore:

$$r_S(R) = r_S(S-R) \tag{4.6}$$

Now let us define the parameter $\rho_S : \mathcal{P}(S) \to \mathbb{R}$ which is the probability that all the loci which belong to $U \subseteq S$ have come from the same parental gamete, in other words $\rho_S(U)$ is the probability of non-recombination among the set of loci U [12]. Therefore:

$$\rho_S(U) = \Pr[\{x \in \mathcal{X} : U \subseteq \Gamma' x \subseteq S\} \cup \{x \in \mathcal{X} : S-U \subseteq \Gamma' x \subseteq S\}] \tag{4.7}$$

If we exclude cases in which $U = \emptyset$ and $U = S$, we have:

$$\{x \in \mathcal{X} : U \subset \Gamma' x \subset S\} \cap \{x \in \mathcal{X} : S-U \subset \Gamma' x \subset S\} = \emptyset \tag{4.8}$$

Hence, by the additivity of probability measure, from (4.7) we get:

$$\rho_S(U) = \Pr\{x \in \mathcal{X} : U \subseteq \Gamma' x \subseteq S\} + \Pr\{x \in \mathcal{X} : S-U \subseteq \Gamma' x \subseteq S\} \tag{4.9}$$

On the other hand, we have:

$$\Pr\{x \in \mathcal{X} : U \subseteq \Gamma' x \subseteq S\} = \sum_{A:U\subseteq A\subseteq S} r_S(A) \tag{4.10}$$

and

$$\begin{aligned} \Pr\{x \in \mathcal{X} : S-U \subseteq \Gamma' x \subseteq S\} &= \sum_{A:S-U\subseteq A\subseteq S} r_S(A) \\ &= \sum_{A:U\subseteq A\subseteq S} r_S(S-A) = \sum_{A:U\subseteq A\subseteq S} r_S(A) \end{aligned} \tag{4.11}$$

Therefore by (4.6) we have:

$$\rho_S(U) = \sum_{A:U\subseteq A\subseteq S} 2r_S(A) \tag{4.12}$$

By Möbius inversion of (4.12), we have:

$$2r_S(U) = \sum_{A:U\subseteq A\subseteq S} (-1)^{|A-U|}\rho_S(A) \tag{4.13}$$

If we put $U = \emptyset$ in (4.7), we have:

$$\{x \in \mathcal{X} : \emptyset \subseteq \Gamma' x \subseteq S\} \cup \{x \in \mathcal{X} : S \subseteq \Gamma' x \subseteq S\} = \{x \in \mathcal{X} : \Gamma' x \subseteq S\} \tag{4.14}$$

Hence:

$$\rho_S(\emptyset) = \Pr\{x \in \mathcal{X} : \Gamma' x \subseteq S\} = \sum_{A:A\subseteq S} r_S(A) = 1 \tag{4.15}$$

Therefore $r_S(A)$ is the probability mass function for the random set $\Gamma' x = A$. With a change of variables in (4.7) we can rewrite this equation as follows:

$$\rho_S(U) = \Pr[\{x \in \mathcal{X} : U \subseteq \Gamma' x \subseteq S\} \cup \{x \in \mathcal{X} : U \subseteq S - \Gamma' x \subseteq S\}] \tag{4.16}$$

Now if we let $S = U$ in (4.16) we get [12]:

$$\begin{aligned}
\rho_S(U) \quad &= \Pr[\{x \in \mathcal{X} : U \subseteq \Gamma' x \subseteq U\} \cup \{x \in \mathcal{X} : U \subseteq U - \Gamma' x \subseteq U\}] \\
&= \Pr[\{x \in \mathcal{X} : U = \Gamma' x\} \cup \{x \in \mathcal{X} : U - \Gamma' x = U\}] \\
&= \Pr\{x \in \mathcal{X} : U = \Gamma' x\} + \Pr\{x \in \mathcal{X} : \Gamma' x = \emptyset\} \\
&= r_U(U) + r_U(\emptyset) = 2r_U(U) = 2r_U(\emptyset)
\end{aligned} \tag{4.17}$$

If we put $U = \emptyset$ in (4.13) we get:

$$2r_S(\emptyset) = \sum_{A:A\subseteq S} (-1)^{|A|}\rho_S(A) = \rho_S(S) \tag{4.18}$$

4.2 Change of Haplotype Probabilities

As a result of meiosis and random union of gametes, the frequency of haplotypes in a population will be different in the next generation, unless the population has reached

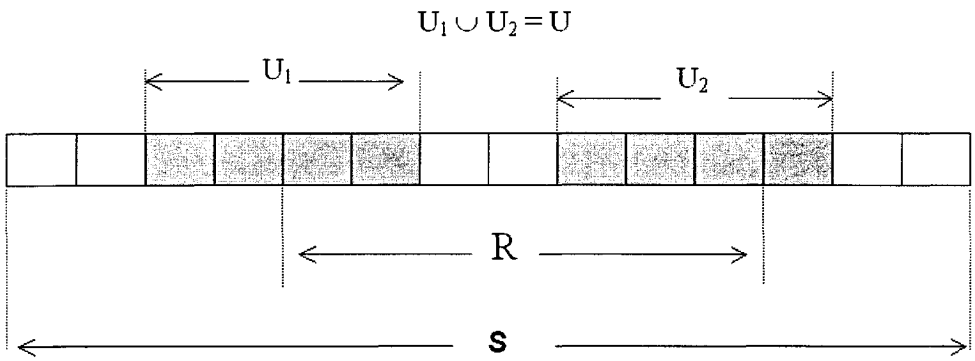


Figure 4.1: Representation of an arbitrary haplotype.
The structure of the haplotype represented in Figure 4.1 is such that the region R has come from the father of the individual and the region represented by $U = U_1 \cup U_2$ is the set of loci which carry alleles '1'

the equilibrium [30].

The subject of this section is to find a relation between haplotype frequencies in $(t+1)$'th generation knowing the state in the t'th generation. Therefore we want to determine $g_S^{t+1}(U)$ as a function of $g_S^t(U)$

The structure of a haplotype in $(t+1)$'th generation, in general, is such that a part of it $(R \subseteq S)$ has come from parent 1 and the rest $(S - R \subseteq S)$ from parent 2. This is because of recombination effect in meiosis.

Suppose we want to find the probability of a haplotype with a structure showed in Figure 4.1. The probability of this event in the space $\mathcal{X}$ is:

$$\Pr(R \text{ has come form parent 1}) \times \Pr(\text{parent 1 has allele 1 on segment } R \cap U \subseteq U)$$
$$\times \Pr(\text{Parent 1 has allele 1 on segment } (S - R) \cap U \subseteq U)$$

Therefore:

$$
\begin{aligned}
g_S^{t+1}(U) &= \sum_{R:R\subseteq S} r_S(R)\Pr\{x : U\cap R\subseteq \Gamma x\subseteq U\}\Pr\{x : (S-R)\cap U\subseteq \Gamma x\subseteq U\} \\
&= \sum_{R:R\subseteq S} r_S(R) \sum_{A:U\cap R\subseteq A\subseteq U} g_S^t(A) \sum_{\substack{B:(S-R)\cap U\\ \subseteq B\subseteq U}} g_S^t(B) \\
&= \sum_{R:R\subseteq S} r_S(R) \sum_{\substack{A:A\subseteq S\\ U\cap R=A\cap R}} g_S^t(A) \sum_{\substack{B:B\subseteq S\\ (S-R)\cap B=(S-R)\cap U}} g_S^t(B)
\end{aligned}
\tag{4.19}
$$

Equation (4.19) describes the change of haplotype frequencies over time [12], we can simplify this equation, by deriving the recursion equation for the commonality measure of random set U, i.e. $Q(U)$, defined by (4.1):

$$
Q_t(U) = \sum_{A:U\subseteq A\subseteq S} g_S^t(A) \Rightarrow Q_t(S) = g_S^t(S) \tag{4.20}
$$

Now if we put $U = S$ in (4.19) we get [12]:

$$
g_S^{t+1}(S) = \sum_{R:R\subseteq S} r_S(R) \sum_{A:R\subseteq A\subseteq S} g_S^t(A) \sum_{(S-R)\subseteq B\subseteq S} g_S^t(B) \tag{4.21}
$$

On the other hand from basic set theoretic rules, we have:

$$
S\cap R = A\cap R \Rightarrow R = A\cap R \Rightarrow R\subseteq A\subseteq S \tag{4.22}
$$

Hence:

$$
g_S^{t+1}(S) = \sum_{R\subseteq S} r_S(R)Q_t(R)Q_t(S-R) \tag{4.23}
$$

From (4.20) and (4.23) we have:

$$
Q_{t+1}(S) = \sum_{R\subseteq S} r_S(R)Q_t(R)Q_t(S-R) \tag{4.24}
$$

Equation (4.24) is valid for all set of loci, hence we have:

$$Q_{t+1}(U) = \sum_{A \subseteq U} r_U(A) Q_t(A) Q_t(U - A); \quad \forall U \subseteq S \tag{4.25}$$

Equation (4.25), describes the change in frequency of haplotypes over time, i.e. it provides a recursive relation for calculation of haplotype probabilities at each generation [23, 12]. In the following examples, we will derive the Hardy-Weinberg principle and equation (2.8), from Chapter 2, as special cases of (4.25)

Example 4.2.1. *The case of one locus (HW principle)*

In the case of only one locus, we have $S = \{1\}, |S| = 1$. Hence we can partition the set of all haplotypes into two subsets $\mathcal{X}_0, \mathcal{X}_1$ such that $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1$, $\mathcal{X}_0 \cap \mathcal{X}_1 = \emptyset$, and:

$$\mathcal{X}_1 = \{x \in \mathcal{X} : \Gamma x = \{1\}\} \quad \text{and} \quad \mathcal{X}_0 = \{x \in \mathcal{X} : \Gamma x = \emptyset\} \tag{4.26}$$

Hence all the haplotypes belonging to the set $\mathcal{X}_1$ contain a gene which is appeared with allele 1 at the specified locus, and $\mathcal{X}_0$ represent the set of all haplotypes with 0 allele on that locus. In this case equation (4.25) will become:

$$Q_{t+1}(\{1\}) = r_{\{1\}}(\emptyset) Q_t(\{1\}) Q_t(\emptyset) + r_{\{1\}}(\{1\}) Q_t(\{1\}) Q_t(\emptyset) \tag{4.27}$$

But by (4.6), $r_{\{1\}}(\emptyset) = r_{\{1\}}(\{1\})$, Hence:

$$Q_{t+1}(\{1\}) = 2 r_{\{1\}}(\emptyset) Q_t(\{1\}) Q_t(\emptyset) \tag{4.28}$$

By (4.18), $2 r_{\{1\}}(\emptyset) = \rho(\{1\}) = 1$, Hence:

$$Q_{t+1}(\{1\}) = \rho(\{1\}) Q_t(\{1\}) = Q_t(\{1\}) \tag{4.29}$$

Equation (4.29) implies that when $|S| = 1$, case of one factor, the population reaches it's equilibrium state after one generation and the haplotype frequencies remain constant.

Example 4.2.2. *The case of two loci.* Now let $S = \{1, 2\}$, in this case, $\mathcal{X} = \mathcal{X}_{11} \cup \mathcal{X}_{10} \cup \mathcal{X}_{01} \cup \mathcal{X}_{00}$, where:

$$\mathcal{X}_{11} \triangleq \{x \in \mathcal{X} : \Gamma x = \{1, 2\}\}; \quad \mathcal{X}_{10} \triangleq \{x \in \mathcal{X} : \Gamma x = \{1\}\};$$
$$\mathcal{X}_{01} \triangleq \{x \in \mathcal{X} : \Gamma x = \{2\}\}; \quad \mathcal{X}_{00} \triangleq \{x \in \mathcal{X} : \Gamma x = \emptyset\}$$

In this case, from (4.25) we have:

$$\begin{aligned} Q_{t+1}(\{1,2\}) &= r_{\{1,2\}}(\emptyset)Q_t(\emptyset)Q_t(\{1,2\}) + r_{\{1,2\}}(\{1\})Q_t(\{1\})Q_t(\{2\}) \\ &+ r_{\{1,2\}}(\{2\})Q_t(\{1\})Q_t(\{2\}) + r_{\{1,2\}}(\{1,2\})Q_t(\{1,2\})Q_t(\emptyset) \\ &= 2r_{\{1,2\}}(\emptyset)Q_t(\{1,2\}) + 2r_{\{1,2\}}(\{1\})Q_t(\{1\}Q\{2\} \end{aligned} \tag{4.30}$$

To simplify the notations, let:

$$\begin{aligned} & Q_t(\{1,2\}) \stackrel{\nabla}{=} Q_{12}^{(t)} \quad Q_i \stackrel{\nabla}{=} Q(\{i\}); \\ & r_{\{1,2\}}(\{i\}) \stackrel{\nabla}{=} r_i; \quad r_{\{1,2\}}(\{1,2\}) \stackrel{\nabla}{=} r_{12}; \\ & \rho(\{1,2\}) \stackrel{\nabla}{=} \rho_{12}; \rho(\{i\}) \stackrel{\nabla}{=} \rho_i, \\ & i = 1,2 \end{aligned}$$

Now we can rewrite (4.30) as:

$$Q_{12}^{(t+1)} = 2r_{12}Q_{12}^{(t)} + 2r_1Q_1^{(t)}Q_2^{(t)} \tag{4.31}$$

On the other hand by (4.13) we have:

$$2r_{12} = \rho_{12} \quad \text{and} \quad 2r_1 = \rho_1 - \rho_{12} \tag{4.32}$$

But from (4.15) we know that

$$\sum_{U:U\subseteq S} r_S(U) = 1$$

Hence:

$$2r_1 + 2r_{12} = 1 \tag{4.33}$$

Therefore by (4.32) we have:

$$\rho_1 - \rho_{12} + \rho_{12} = \rho_1 = 1 \tag{4.34}$$

Hence we can write equation (4.31) as:

$$\begin{aligned} Q_{12}^{(t+1)} &= \rho_{12}Q_{12}^{(t)} + (\rho_1 - \rho_{12})Q_1^{(t)}Q_2^{(t)} \\ &= \rho_{12}Q_{12}^{(t)} + (1 - \rho_{12})Q_1^{(t)}Q_2^{(t)} \end{aligned} \tag{4.35}$$

Subtracting $Q_1^{(t)}Q_2^{(t)} = Q_1Q_2$ from both sides of (4.35) we get:

$$\begin{aligned} Q_{12}^{(t+1)} - Q_1Q_2 &= \rho_{12}Q_{12}^{(t)} + (1-\rho_{12})Q_1^{(t)}Q_2^{(t)} - Q_1Q_2 \\ &= \rho_{12}(Q_{12}^{(t)} - Q_1Q_2) \end{aligned} \tag{4.36}$$

Note that here we defined the recombination rate in terms of the probability of non-recombination between two loci, therefore $\rho_{12} = 1 - r$ which r is the probability of recombination between two loci, defined in Section 2.5.6. Hence equation (4.36) is the same as equation (2.9).

4.3 Solving the Difference Equation

The parameter $Q_t(U)$ describes the probability of finding a haplotype, in which $U \subseteq S$ is occupied by allele '1'. Equation (4.25) describes the change of this parameter over time. Because of the quadratic nature of (4.25), it is difficult to solve this equation in terms of the haplotype distributions in the founding population, i.e. $Q_0(U)$ [23, 12]. Bennett [4], found that we can define the state of a population, in terms of a new set of variables, which are related to the haplotype frequencies, and decay geometrically over time, in other words, if we denote the new variable by $b_t(U)$, then:

$$b_{t+1}(U) = \lambda(U)b_t(U) \tag{4.37}$$

in which $\lambda(U)$, is a parameter independent of time. Furthermore, it can be shown that $\lambda(U)$ is equal to the recombination rate among the set of loci denoted by U, namely $\rho(U)$ [12]. Hence the rate of decay for the parameter $b_t(U)$ is equal to the recombination rate. For example, if the loci represented by U are close to each other, the rate of decay is very low, and if they are loosely linked (i.e. far apart of each other), then $b_t(U)$ decays rapidly.

In this section, we discuss the properties of "Bennet Variables" [12] with the general formalism developed by Dawson [12]. For the detailed proofs and calculations, see [12, 4].

Definition 4.3.1. *Partition of a Set*

By a partition Θ of the set U, we mean a collection of disjoint subsets of U, such that, if $A_1, A_2, \ldots, A_m \in \Theta$ then $A_i \cap A_j = \emptyset, \quad \forall i \neq j$ and $U = A_1 \cup A_2 \cup \ldots A_m$.

Let $\pi(U)$ denote the collection of the all partitions of U, i.e.:

$$\pi(U) = \{\Theta : \Theta = \{\Theta_1, \ldots, \Theta_{|\Theta|}\}, \Theta_i \cap \Theta_j = \emptyset, \bigcup_{i=1}^{|\Theta|} \Theta_i = U\} \tag{4.38}$$

Now let us define the new variable $b_t(U)$, where $U \subseteq S$, as:

$$b_t(U) \triangleq \sum_{\Theta \in \pi(U)} C(\Theta) \prod_{A \in \Theta} Q_t(A) \tag{4.39}$$

in which $C(\Theta)$ are constants and are only dependent on the partition Θ. Also, we can find the inverse transformation describing the parameter Q in terms of the new set of variables [12]:

$$Q_t(U) = \sum_{\Theta \in \pi(U)} \bar{C}(\Theta) \prod_{A \in \Theta} b_t(A) \tag{4.40}$$

The main step to complete the calculation of $b_t(U)$, is to find constants $C, \bar{C}$. The main idea in Dawson [12] is to substitute (4.40) in (4.25), and using equation (4.37), to find parameter $\bar{C}$ (and the same method for finding C). Here without going through the calculations for finding $C, \bar{C}$, we give the results derived by Dawson for cases in which $|\Theta| = 1, 2, 3$ for $\bar{C}$ and C.

The value of $\bar{C}(\{U\})$, in which the partition $\Theta = \{U\}$, is an arbitrary constant and for convenience we choose

$$\bar{C}(\{U\}) = 1, \quad C(\{U\}) = 1 \tag{4.41}$$

For partitions of U, into two partitions $\Theta = \{U_1, U_2\}$,(i.e. $|\Theta| = 2$), we have:

$$\bar{C}(\{U_1, U_2\}) = \frac{2r_U(U_1)}{\rho(U_1)\rho(U_2) - \rho(U)} \tag{4.42}$$

and

$$C(\{U_1, U_2\}) = -\frac{2r_U(U_1)}{\rho(U_1)\rho(U_2) - \rho(U)} \tag{4.43}$$

And for the case of $|\Theta| = 3$, we have:

$$\begin{aligned}\bar{C}(\{U_1, U_2, U_3\}) \quad &= \frac{1}{\rho(U_1)\rho(U_2)\rho(U_3) - \rho(U)} \times \left(\frac{2r_U(U_1)2r_{U_2\cup U_3)}(U_2}{\rho(U_2)\rho(U_3) - \rho(U_2 \cup U_3)}\right. \\ &\left. + \frac{2r_U(U_2)2r_{U_1\cup U_3}(U_1)}{\rho(U_1)\rho(U_3) - \rho(U_1 \cup U_3)} + \frac{2r_U(U_3)2r_{U_1\cup U_2}(U_1)}{\rho(U_1)\rho(U_2) - \rho(U_1 \cup U_2)}\right)\end{aligned} \tag{4.44}$$

and

$$\begin{aligned}C(\{U_1, U_2, U_3\}) \quad &= \frac{1}{\rho(U_1)\rho(U_2)\rho(U_3) - \rho(U)} \times \left(\frac{\rho(U_1)2r_U(U_1)2r_{U_2\cup U_3)}(U_2)}{\rho(U_1)\rho(U_2 \cup U_3) - \rho(U)}\right. \\ &\left. + \frac{\rho(U_2)2r_{U_1\cup U_3}(U_1)2r_U(U_2)}{\rho(U_2)\rho(U_1 \cup U_3) - \rho(U)} + \frac{\rho(U_3)2r_{U_1\cup U_2}(U_1)2r_U(U_3)}{\rho(U_3)\rho(U_1 \cup U_2) - \rho(U)}\right)\end{aligned} \tag{4.45}$$

By finding the constants $C(\Theta), \bar{C}(\Theta)$, we can determine the parameter Q (and hence the haplotype frequencies) in terms of parameter $b_t(U)$ which decays geometrically over time (and hence can relate the haplotype frequencies at time t, to the haplotype frequencies in the founding population).

Bennett variable defined by (4.39) can be considered as a measure for linkage disequilibrium for an arbitrary number of loci, and it decays geometrically over time. The rate of decay is equal to the probability of non-recombination among the set of loci under study. In other words, the rate of decay for Bennett's measure of linkage disequilibrium is equal to the probability that the set of loci U, have been transmitted from

the parents of the individual, without any change during the process of meiosis. Let us consider the case in which $S = \{1, 2\}$ (the case of two loci), and suppose we consider the haplotype $(1, 1)$ which means that loci $1, 2$ are occupied by 1 allele. Therefore $U = S$, (U denotes the set of loci which are occupied by 1 allele). For the case of two loci:

$$\pi(U) = \{\{\{1\}, \{2\}\}, \{\{1, 2\}\}\}$$

From (4.41) and (4.42), we have

$$C(\{\{1, 2\}\}) = 1$$

and

$$C(\{\{1\}, \{2\}\}) = -\frac{2r_{\{1,2\}}(\{1\})}{\rho(\{1\})\rho(\{2\}) - \rho(\{1, 2\})}$$

Therefore by (4.39), we have:

$$b_t(\{1, 2\}) = Q_t(\{1, 2\}) - \frac{2r_{\{1,2\}}(\{1\})}{\rho(\{1\})\rho(\{2\}) - \rho(\{1, 2\})} Q_t(\{1\})Q_t(\{2\}) \tag{4.46}$$

On the other hand from (4.32),(4.34) we have:

$$\begin{aligned} &2r_{\{1,2\}}(\{1\}) = 2r_{\{1,2\}}(\{2\}) = \rho(\{1\}) - \rho(\{1, 2\}) \\ &\rho(\{1\}) = \rho(\{2\}) = 1 \end{aligned} \tag{4.47}$$

Hence:

$$b_t(\{1, 2\}) = Q_t(\{1, 2\}) - Q_t(\{1\})Q_t(\{2\}) \tag{4.48}$$

Therefore, for the case of two loci, Bennett's measure of linkage disequilibrium is equal to the D measure introduced in Chapter 2, and its rate of decay is $\rho = 1 - r$, which r is the recombination fraction defined in Section 2.5.6 [30]

4.4 Uses of LD in Mapping Human Genome

Studying linkage disequilibrium among a set of loci is one the most important problems in genetic analysis. The main application of linkage analysis is in mapping complex traits such as common diseases in human populations [27, 46, 56]. There are two complimentary approaches in the identification of the underlying genes for complex traits: (i) pedigree analysis and (ii) association studies [6]. The pedigree analysis leads us to find the loci which co-segregate together in a family pedigree, while association studies seek to identify particular variants that are associated with the phenotype at the population level [6].

We say that disease-marker association exists if alleles at the marker locus occur with different relative frequencies in affected and unaffected individuals, in the other words, association exists if the probability of being affected at the population level is not independent of probability of finding a certain allele M at the marker locus [49]:

$$\Pr(\text{disease}, M) \neq P(\text{disease})P(M)$$

When a disease is in association with a marker locus there are two possibilities. The first case is when the marker locus is causative, meaning that it directly influences the disease risk. The other case is when alleles at the marker locus are correlated with alleles at the disease locus, but do not directly influence disease risk, in this case the disease locus is in linkage disequilibrium with the marker locus [49].

The rate of decay of LD (which is the recombination fraction between the two loci) is used to define map distance between two loci [67, 27, 46] . The genetic map distance between two loci [31] is defined as the expected value of crossover events to occur between them in a gamete [41]. The unit for this distance is called Morgan (which in practice, centi Morgan denoted by cM, is commonly used). There are various mapping functions in order to relate genetic map distance to recombination fractions [51]. One

of the commonly used functions due to Haldane [31] assumes that crossovers occur as a Poisson process with rate 1 per Morgan, also, it assumes that the crossover events occur independent of each other along the chromosome (no-interference assumption). Under this model, the relationship between the genetic distance between any two loci, denoted by λ, and the corresponding recombination fraction, r, is given by:

$$r = \frac{1}{2}(1 - e^{-2\lambda}) \tag{4.49}$$

with the inverse function [41, 31]

$$\lambda = -\frac{1}{2}\ln(1 - 2r) \tag{4.50}$$

4.4.1 Measures of LD

Other than the measure D for pair wise LD there are other measures used in linkage analysis. One of the commonly used measures for LD between two loci with alleles A, a and B, b is defined as [50, 15]:

$$D' = \frac{D}{D_{Max}} \tag{4.51}$$

in which

$$D = P(AB) - P(A)P(B)$$

and

$$D' = \begin{cases} \min(P(A)P(b), P(a)P(B)), & \text{if} \quad D > 0 \\ \min(P(A)P(B), P(a)P(b)), & \text{if} \quad D < 0 \end{cases} \tag{4.52}$$

D' is a normalized measure for LD and is independent of the allele frequencies, hence, it enables us to compare linkage disequilibrium between two different pairs of loci.

Defining a measure of LD among an arbitrary number of loci, is still an open problem in mathematical population genetics. In Section 4.3, we discussed the b measure

defined by (4.39) which is a generalization of the pairwise LD measure D. Other measures of linkage disequilibrium are defined in order to reflect the degree of statistical association among a set of loci, from which we can name entropy based measure of LD [50], and recursive definition for LD in which LD for n loci is defined in terms of LD for $n-1$ loci [26]. For more information about the measures for LD and a comparison of these measures see [29, 15].

Chapter 5

Continuous Time Model

In a discrete time model, the change of haplotype probabilities is described in terms of a difference equation, i.e., an equation which relates the probability distribution at time t to the probability distribution at time $t-1$ (equations (4.19),(4.25)). If we consider a continuous time model for change of the genetic composition of the population, the dynamic of the population can be described in terms of a differential equation. Here we discuss the method proposed by Baake [2], in which the process of change of haplotype frequencies is considered as a differential operator acting on the set of all probability measures defined on the measure space corresponding to the population under study.[1]

5.1 The Model

Let $(\mathcal{X}, \mathbb{X})$ be the measurable space corresponding to the population $\mathcal{X}$ and let $\mathcal{M}(\mathcal{X})$ be the set of all probability measures defined on $(\mathcal{X}, \mathbb{X})$. Therefore, for all $\mu \in \mathcal{M}(\mathbf{X})$:

$$\mu(\mathcal{X}) = 1$$

[1]The model described here is a simplified version of the model described in [2]. In the original model, the set of all additive measures is considered. In this thesis, we have considered the special case of probability measures i.e. the set of all positive measures satisfying $\mu(\mathcal{X}) = 1$ in which $\mathcal{X}$ is the sample space. For more information see [3, 2].

Let $S = \{1, \ldots, n\}$ denote the set of all loci under consideration. In general, a gene can have more than two loci, let $\mathcal{X}_i = \{0, \ldots, m_i\}$ be the set of all possible alleles that can appear on loci i. Therefore each possible realization of $X \in \mathcal{X}$ can be considered as a member of the set $\mathcal{X}_1 \times \cdots \times \mathcal{X}_N$, in which $'\times'$ represents the Cartesian product of sets. For example, consider the case in which there are three loci, and each can take three possible alleles, therefore

$$\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}_3 = \{0, 1, 2\}$$

and

$$\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 = \{(i, j, k) : i, j, k = 0, 1, 2\}$$

Therefore, as $\mathcal{X}$ represents the set of all possible haplotypes, we can write:

$$\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n \tag{5.1}$$

Let $\mathcal{M}_i$ to be the set of all possible probability measures defined on the space $\mathcal{X}_i$, (i.e. all possible allele frequencies corresponding to locus i). It can be shown that $\prod_{i=0}^{n} \mathcal{M}_i \subset \mathcal{M}(\mathcal{X})$ [2].

In this model, in addition to the set of loci denoted by S, we consider the set of *links* between loci as a set $L = \{0, 1, 2, \ldots, n-2\}$, in which $l \in L$ represents the link between locus $l+1$ and $l+2$. To each link $l \in L$ we assign a binary random variable $\alpha_l \in \{0, 1\}$, with the interpretation that $\alpha_l = 1$ means that a break between locus $l+1$, and $l+2$ has happened. Each configurations of breaks $\tilde{\alpha} = (\alpha_0, \ldots, \alpha_{n-2})$ can be identified by the set of links which have value 1, for example in the case of three loci, (i.e. $S = \{1, 2, 3\}$, $L = \{0, 1\}$), $\tilde{\alpha} = (1, 0)$ corresponds to the subset $A = \{0\}$ of L and it is representative of a haplotype that a break has happened between its first and second loci. We notice that each break configuration corresponds to a partition of S. Therefore, we can set a one-to-one correspondence between subsets of L and $\pi(S)$, the set of possible groupings

of the loci, and we call each group of loci as a *block*. For example, in the previous case of three loci,

$$\mathcal{P}(L) = \{\{0\}, \{1\}, \{0,1\}, \emptyset\}$$

and we have the following correspondence between $\mathcal{P}(L)$ and the set of all possible ways that we can separate a set of loci from the others:

$$\{0\} \Leftrightarrow \{[1], [2,3]\}$$
$$\{1\} \Leftrightarrow \{[1,2], [3]\}$$
$$\{0,1\} \Leftrightarrow \{[1], [2], [3]\}$$
$$\emptyset \Leftrightarrow \{[1,2,3]\}$$

In general, we have the following relation between any subset of L and different collection of blocks:

$$\{a_1, \ldots, a_q\} \Leftrightarrow \{[1, a_1+1], [a_1+2, \ldots, a_2+1], \ldots, [a_{q-1}+2, \ldots, a_{q+1}], [a_{q+2}, \ldots, n]\} \tag{5.2}$$

in which

$$\{a_1, \ldots, a_q\} \subseteq L$$

and $n = |S|$, is the number of loci under study.

5.1.1 Recombination Operators on Probability Measures

In the model described in [2], recombination is represented by an operator acting on the set of all probability measures $\mathcal{M}(\mathcal{X})$. To construct recombination operators, first we define a marginalization operator $\psi_i : \mathcal{X} \to \mathcal{X}_i$ defined as follows:

$$\psi_i(X) = X_i, \quad X \in \mathcal{X}, \quad X_i \in \mathcal{X}_i \tag{5.3}$$

In general, let $I = \{I_1, \ldots, I_{|I|}\} \subseteq S$ then:

$$\psi_I(X) = X_I = (X_{I_1}, \ldots, X_{I_{|I|}}), \quad X_{I_1} \in \mathcal{X}_{I_1}, \ldots, X_{I_{|I|}} \in \mathcal{X}_{I_{|I|}} \tag{5.4}$$

This mapping, induces a mapping on $\mathcal{M}(\mathcal{X})$, which can be interpreted as marginalization [2] (see Section 3.2). We denote this induced mapping, by Ψ and we have: $\Psi_i : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X}_i)$ and is defined as:

$$\Psi_i \cdot \omega(E) = \omega(\psi_i^{-1}(E)), \quad \forall E \subseteq \mathcal{X}_i \tag{5.5}$$

in which $\omega \in \mathcal{M}(\mathcal{X})$ is a probability measure. In other words, $\Psi_i \cdot \omega(E)$ is the probability measure of all events in $\mathcal{X}$ such that their i'th component is in the set $E \subset \mathcal{X}_i$. In the special case that E is a singleton of $\mathbb{X}$, $E = W = (W_1, \ldots, W_n) \in \mathcal{X}$ we have

$$\Psi_i \cdot \omega(W) = \omega(\psi_i^{-1}(W)) = \omega(W_i)$$

which can be interpreted as the frequency of the allele W_i.

Similar to ψ_i operators, we can define $\Psi_I : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X}_I)$ as:

$$\Psi_I \cdot \omega(E) = \omega(\psi_I^{-1}(E)), \quad \forall E \subseteq \mathcal{X}_I \tag{5.6}$$

in which $\mathcal{X}_I = \mathcal{X}_{I_1} \times \cdots \times \mathcal{X}_{I_{|I|}}$.

In particular we are interested in the index sets of the form $I = (1, 2, \ldots, a+1)$ or $J = (a+2, \ldots, n)$, in which $a \in L$. For simplicity we introduce the following notation

$$\Psi_{<a} \stackrel{\nabla}{=} \Psi_I, \quad I = (1, 2, \ldots, a+1)$$

and

$$\Psi_{>a} \stackrel{\nabla}{=} \Psi_I, \quad I = (a+2, \ldots, n)$$

Therefore $\Psi_{<a} \cdot \omega$ where $a \in L$ and $\omega \in \mathcal{M}(\mathcal{X})$, is the probability measure of all the elements in $\mathcal{X}$ such that they have the same value on sites 1 up to $a+1$.

Definition 5.1.1. *The Elementary Recombination Operator*

The recombination operator $R_a : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$, is defined as:

$$R_a(\omega) \triangleq (\Psi_{<a} \cdot \omega)(\Psi_{>a} \cdot \omega), \quad \forall a \in L \tag{5.7}$$

In particular

$$R_a(\omega)(X) = [(\Psi_{<a} \cdot \omega)(\Psi_{>a} \cdot \omega)](X) = \omega(\psi_{<a}^{-1}(X_{<a}))\omega(\psi_{>a}^{-1}(X_{>a}))$$

We can extend the definition of R_a to any subset of L as follows. Let $G \subseteq L$ then: [2]

$$R_G \triangleq \prod_{a \in G} R_a \tag{5.8}$$

Hence R_a can be considered as a special case of R_G for which $G = \{a\}$. Following without going through the proofs, we state some properties of the recombination operator (see [2]).

1. For any scalar $s \in \mathbb{R}$ and $\omega \in \mathcal{M}(\mathcal{X})$, we have $R_a(s\omega) = |s|R_a(\omega)$, in which $|a|$ is the absolute value of a

2. For all values of a we have: $R_a R_a = R_a^2 = R_a$, in other words, R_a is an idempotent [37] operator

3. For all values of $a, b \in L$, $R_a R_b = R_b R_a$, in other words, recombination operators are commutative

5.2 Change of Haplotype Frequencies

In this model, we assume that each individual (for which its genetic composition is represented by $X = (X_1, \ldots, X_n)$ such that $X \in \mathcal{X}$ and $X_i \in \mathcal{X}_i$) carries a Poisson clock at each link $a \in L$, with Poisson parameters $\rho_a > 0$, which do not depend on the individual. If the clock at link a of the individual of type X rings, a partner of type

Y is picked randomly (random mating assumption) from the population for recombination at that link [2]. Therefore the recombined pair is $(X_1, \ldots, X_{a+1}, Y_{a+2}, \ldots, Y_n)$ and $(Y_1, \ldots, Y_{a+1}, X_{a+2}, \ldots, X_n), \forall a \in L$.

Let $Z_X(t)$ be the random variable that represents the number of individuals of type X at time t, and let $Z(t)$ be the random vector with components $Z_X(t)$. For example in the case of $S = \{1, 2\}$, $\mathcal{X}_1 = \mathcal{X}_2 = \{0, 1\}$, we have four types $X^{(1)} = (0, 0), X^{(2)} = (0, 1), X^{(3)} = (1, 0), X^{(4)} = (1, 1)$, then $Z(t) = (Z(t)_{X^{(1)}}, Z(t)_{X^{(2)}}, Z(t)_{X^{(3)}}, Z(t)_{X^{(4)}})$. Also let us denote the unit vector corresponding to X by U_X, consider the example of two loci, $U_{X^{(2)}} = (0, 1, 0, 0)$. As a result of the process of recombination the number of different types in the population will be changed over time. Suppose that the Poisson clock at link a of an individual of type $X = (X_1, \ldots, X_n)$ rings and the individual of type X mates with an individual of type $Y = (Y_1, \ldots, Y_n)$ $(Y \neq X)$. Therefore we have a decrease in the number of individuals of types X and Y by one, and at the same time we have an increase in the number of individuals of types $(X_1, \ldots, X_{\alpha+1}, Y_{\alpha+2}, \ldots, Y_n) \stackrel{\nabla}{=} [X_{<\alpha}, Y_{>\alpha}]$ and $(Y_1, \ldots, Y_{\alpha+1}, X_{\alpha+2}, \ldots, X_n) \stackrel{\nabla}{=} [Y_{<\alpha}, X_{>\alpha}]$.

Therefore, for the change in the value of $Z(t)$ we can write:

$$Z \rightarrow Z - U_X - U_Y + U_{[X_{<\alpha}, Y_{>\alpha}]} + U_{[Y_{<\alpha}, X_{>\alpha}]} \stackrel{\nabla}{=} Z' \tag{5.9}$$

The rate of such transitions is equal to the probability of the event that individuals of types X and Y mate and recombination happens at link a, as we assume that mating is random and parameter ρ_a is independent of the individual's type we have:

$$\Pr(Z \rightarrow Z') = \rho_a P(Z_X) P(Z_Y) \tag{5.10}$$

5.2.1 The Differential Equation

Now suppose that we want to find the rate of change in the frequency of type W in a time interval $[t, t+\Delta t]$, therefore we want to find:

$$\dot{P}_t(W) = \lim_{\Delta t \to 0} \frac{P_{t+\Delta t}(W) - P_t(W)}{\Delta t} \tag{5.11}$$

Let us start from a discrete time model, i.e. assume that generations have a life span denoted by Δt. We can reformulate the discrete time difference equation for the change of haplotype probabilities [3] as:

$$\begin{aligned} P_{t+\Delta t}(W) = &\sum_{a \in L} P(W|\text{one cross over event at link } a) P_{\Delta t}(\text{one cross over event at link } a) \\ &+ \sum_{a<b \in L} P(W|\text{two cross over events at links } a,b) P_{\Delta t}(\text{two cross over events at links } a,b) \\ &+\text{terms involving more than two crossovers} \end{aligned} \tag{5.12}$$

in which by $P(W|\text{one cross over event})$ we mean the probability of having the type W in the current generation given one cross over event has happened during the formation of this type at link a, and

$P_{\Delta t}(\text{one cross over event at link } a)$ is the probability of this event during the time period Δt, and finally we need to sum over all the links in L to marginalize over all the links in L. As we will eventually let $\Delta t \to 0$, we can ignore terms involving more than two crossovers.[2] Based on the Poisson clock model for cross over events, the probability of having a cross over event at link a during the time period Δt, denoted by γ_a is:

$$\gamma_a = 1 - e^{-\rho_a \Delta t} \tag{5.13}$$

Hence, the probability of having just one cross over event at link a is:

$$P_{\Delta t}(\text{one cross over event at link } a) = \gamma_a \prod_{b \neq a} (1 - \gamma_b), \quad \forall a, b \in L \tag{5.14}$$

[2]It is because of the fact that if the time interval is short, probability of occurrence of more than two crossover events will be very small

and the probability of having two crossover events will be:

$$P_{\Delta t}(\text{two cross over events at links } a,b) = \gamma_a\gamma_b \prod_{c\neq a,b}(1-\gamma_c), \quad \forall a<b; c\in L \quad (5.15)$$

On the other hand:

$$P(W|\text{one cross over event at link } a) = P(W_{<a})P(W_{>a}) \quad (5.16)$$

The reason for factoring the conditional probability of W, is that if a cross over has happened at link a then two segments of W, ($W_{<a}$ and $W_{>a}$) have come from two different individuals, hence given that there is no correlation between the two individuals, we can assume that probability of $W_{<a}$ is independent of probability of $W_{>}a$, and therefore we can write the joint probability $P(W_{<a}, W_{>a})$ as the product of two events. In Section 5.5, we will discuss more about this type of factorization. In the same way, for two crossovers, we can write:

$$P(W|\text{two cross over event at links } a,b) = P(W_{[1,a+1]\cup[b+2,n]})P(W_{[a+2,b+1]}) \quad (5.17)$$

in which by $[i,j]$ we mean the closed interval that includes loci k such that $i\leq k\leq j$ Hence we can rewrite equation (5.12) as:

$$\begin{aligned} P_{t+\Delta t}(W) \quad &= \sum_{a\in L}\gamma_a \prod_{b\neq a}(1-\gamma_b)P_t(W_{<a})P_t(W_{>a}) \\ &+ \sum_{a<b\in L}\gamma_a\gamma_b \prod_{\delta\neq a,b}(1-\gamma_\delta)P_t(W_{[1,a+1]\cup[b+2,n]})P_t(W_{[a+2,b+1]}) \end{aligned} \quad (5.18)$$

If we expand $\gamma_a = 1-e^{-\rho_a\Delta t}$, using the Taylor expansion, we have:

$$\gamma_a = \rho_a\Delta t + O((\Delta t)^2) \quad (5.19)$$

substituting (5.19) in (5.18) and expanding products and reordering the terms we have:

$$P_{t+\Delta t}(W) = P_t(W) + \Delta t\sum_{a\in L}\rho_a(P_t(W_{<a})P_t(W_{>a}) - P_t(W)) + O((\Delta t)^2) \quad (5.20)$$

Therefore,

$$\frac{P_{t+\Delta t}(W) - P_t(W)}{\Delta t} = \sum_{a \in L} \rho_a (P_t(W_{<a}) P_t(W_{>a}) - P_t(W)) + O(\Delta t) \tag{5.21}$$

Hence from (5.11) and (5.21) we will get:

$$\dot{P}_t(W) = \frac{dP_t(W)}{dt} = \sum_{a \in L} \rho_a (P_t(W_{<a}) P_t(W_{>a}) - P_t(W)) \tag{5.22}$$

using the operator Ψ introduced in (5.6) and denoting P by $\omega \in \mathcal{M}(\mathcal{X})$, we can rewrite equation (5.22) as:

$$\begin{aligned} \dot{\omega}_t(W) &= \sum_{a \in L} \rho_a [(\Psi_{<a} \cdot \omega)(\Psi_{>a} \cdot \omega) - \omega](W) \\ &= \sum_{a \in L} \rho_a [R_a(\omega) - \omega](W) \end{aligned} \tag{5.23}$$

in which we have used the definition of recombination operator introduced in (5.7). Now (suppressing the time dependence for simplicity) we can write the differential equation describing change of probability measure as:

$$\dot{\omega} = \sum_{a \in L} \rho_a (R_a - \mathbf{1})(\omega) \tag{5.24}$$

In which $\mathbf{1}$ is the identity operator. This equation is the corresponding continuous time equation for dynamics of populations. The next step in this analysis, is solving the differential equation (5.24). The solution to the equation (5.24) enables us to find the frequency of haplotypes knowing the initial condition of the population, namely the haplotype frequencies in the founding population. One of the main difficulties in solving equation (5.24) is the fact that this equation is a non-linear ODE, which stems from the non-linearity of the recombination operators R_a. The solution to this equation is one of the topics that will be discussed in the next section.

5.3 Blocks of Genes

In this section, we will show that the solution of the differential equation (5.24), can be directly derived using the notion of "Blocks Identical by Descent" in a population, which are blocks of genes which have remained intact from the ancestral population [5, 25, 52].

5.3.1 Blocks Identical by Descent (BID)

As we discussed in Chapter 2, DNA sequences of a population of individuals at any given generation, are copies of DNA sequences of an ancestral population , subject to some changes in copying. These changes in copying DNA sequences have different origins, for example mutation (an error in copying a nucleotide) or crossover and recombination, in which the copied version a of DNA instead of being a copy of one DNA string in population is a combination of two DNA sequences , existed in past generation. As an example of the effect of cross over and recombination consider there are two original sequences: ARISTOTELLOGIC and PLATOSREPUBLIC, and there is a copier with two readers labelled by R_1 and R_2, the copier takes sequences $S_1^{(0)}$ = ARISTOTELLOGIC and $S_2^{(0)}$ =PLATOSREPUBLIC, and R_1 starts reading sequence $S_1^{(0)}$ and writes the characters of sequence one into the output sequence $S_1^{(1)}$ and the same for R_2 which starts reading and copying sequence $S_2^{(0)}$. Now consider two readers R_1 and R_2 switch their task with probability ρ_i after reading character i in each sequence, i.e., after reading the third character in $S_1^{(0)}$ with probability ρ_3, R_1 continues reading (and copying) sequence $S_2^{(0)}$ and R_2 continues its task using sequence $S_1^{(0)}$; After finishing copying $S_1^{(0)}$ and $S_2^{(0)}$, in the next round the copier takes $S_1^{(1)}$ and $S_2^{(1)}$ as the inputs and this process continues up to $t'th$ round of copying, and the final outputs will be $S_1^{(t)}$ and $S_2^{(t)}$. So starting from sequences $S_1^{(0)}$ and $S_2^{(0)}$ one possible final output sequence will look like: $S_1^{(t)}$ =ARITOOTEPLOLIC and $S_2^{(t)}$ =PLASTSRELUBGIC. As we see in the output sequences $S_1^{(t)}, S_2^{(t)}$ there are *Blocks* of characters preserved

from the original sequences, for example "ARI", "PLO", "LIC" in the $S_1^{(t)}$ and their corresponding blocks in $S_2^{(t)}$. From this point of view the solution to the differential equation (5.24) will enable us to calculate the probability of any possible sequence at a given time. For example in the above example, we might be interested in the probability of having sequence ARISTOREPUBLIC in the second generation which has the same probability of finding the sequence PLATOSTELLOGIC in that generation.

5.4 Link Configurations

In Section 5.1 we introduced the random variable $\tilde{\alpha} = (\alpha_0, ..., \alpha_{n-2})$ where $\alpha_j \in \{0, 1\}$ and n is the number of loci under study. $\tilde{\alpha}$ represent the state of the links between the loci, for example $\alpha_i = 1$ indicates that there has been a crossover event between locus $i+1, i+2$ and $\alpha_i = 0$ indicates that there has been no cross over event, hence $\alpha_i = 0$ means that X_{i+1}, X_{i+2} have been located on the same chromosome in the past generations.

The probability of the haplotype $X = (X_1, \ldots, X_n) \in \mathcal{X}$ can be written as the marginalized joint probability distribution of $(X, \tilde{\alpha})$ over all the possible values of $\tilde{\alpha}$, hence:

$$P_t(X) = \sum_{\tilde{\alpha}} P_t(X; \tilde{\alpha}) = \sum_{\tilde{\alpha}} P_t(X|\tilde{\alpha}) P_t(\tilde{\alpha}) \tag{5.25}$$

Before continuing the calculation of $P(X)$, let us take a look at the interpretation of $\tilde{\alpha}$; Suppose that we have $n = 10$, and let one possible value of $\tilde{\alpha}$ at time t to be $\tilde{\alpha} = (0, 0, 0, 1, 0, 0, 0, 1, 0, 0)$, this means that up to time t only links 3 and 7 have experienced a crossover event, therefore given this configuration, we know that in the corresponding haplotype, loci $\{1, 2, 3, 4\}$ have not experienced any recombination, in the words, alleles $X_1, \ldots, X_4$ have been together on the ancestral haplotype sequence, so they belong to the same individual. In the same way, loci $X_5, \ldots, X_8$ belong to another ancestral individual and X_8, X_9, X_{10} also belong to an-

other individual that existed in the founding population. In this case, we say that $[X_1, \ldots, X_4], [X_5, \ldots, X_8], [X_8, X_9, X_{10}]$ are three blocks of genes identical by descent. If we assume that mating occurs at random and there is no correlation between ancestral individuals, we can conclude that the probability distribution of a haplotype X at time t given $\tilde{\alpha} = (0001000100)$ can be factored to a product of marginals:

$$P_t(X = x_1, \ldots, x_{10} | \tilde{\alpha} = (0001000100)) = P_0(X_1, X_2, X_3, X_4) P_0(X_5, X_6, X_7) P_0(X_8, X_9, X_{10}) \tag{5.26}$$

Let G denote the index set of all links g for which $\alpha_g = 1$ and let L to be the set of all links:

$$G = \{g \in L : \alpha_g = 1\}, \bar{G} = \{h \in L : \alpha_h = 0\}, G \subset L \tag{5.27}$$

For example, $\tilde{\alpha} = (0001000100)$, then we have $G = \{4, 8\}$ and $G = L$ represents the configuration in which all the links have experienced at least one crossover event. This way, we can set a one-to-one correspondence between the subsets of L and each realization of $\tilde{\alpha}$, therefore each value of $\tilde{\alpha}$ can be identified by a subset of L, which represents the set of all links which have experienced at least one cross over event. Now we can generalize our results in 5.26 as:

$$P_t(X = (X_1, \ldots, X_n) | \tilde{\alpha}) = P_0(X_{<g_1}) P_0(X_{>g_{|G|}}) \prod_{\kappa \in G} P_0(X_{g_\kappa + 1}, \ldots, X_{g_{\kappa+1}}) \tag{5.28}$$

In which G is defined in (5.27). Using the marginalization operators defined in Section, 5.1.1 we can write the equation (5.28) as:

$$P_t(X = (X_1, \ldots, X_n) | \tilde{\alpha}) = (\Psi_{<g_1} \cdot P_0)(\Psi_{>g_{|G|}} \cdot P_0) \prod_{\kappa \in G} \Psi_{[g_\kappa + 1, g_{\kappa+1}]} \cdot P_0 \tag{5.29}$$

Using the definition of recombination operators equation (5.29) can be written as:

$$P_t(X = (X_1, \ldots, X_n) | \tilde{\alpha}) = \prod_{g \in G} R_g(P_0) = R_G(P_0) \tag{5.30}$$

Now, using (5.25),(5.30) we can write the probability of a haplotype X at time t as:

$$P(X) = \sum_{\tilde{\alpha}} P_t(\tilde{\alpha}) R_G(P_0), \quad G = \{g : \alpha_g = 1\} \tag{5.31}$$

In the next section we will compute the coefficients $P_t(\tilde{\alpha})$, and we will discuss the biological interpretation for these coefficients.

5.4.1 Distribution of Link Configurations

As we discussed in Section 5.2, the crossover events happen based on a Poisson process. When the Poisson clock attached to the link a rings, readers in our imaginary copier (see Section 5.3.1) switch their task. Therefore, the probability that the link a has not experienced any crossover up to time t is:

$$\begin{aligned} &\Pr\{\text{Link } a \text{ has not experienced a crossover up to time } t\} = e^{-\rho_a t} \\ &\Pr\{\text{Link } a \text{ has experienced at least one crossover up to time } t\} = 1 - e^{-\rho_a t} \end{aligned} \tag{5.32}$$

Let us denote $P_t(\tilde{\alpha})$ by $\gamma_G(t)$, in which $\tilde{\alpha}$ has been identified by $G \subset L$, defined by (5.27). Probability of a certain configuration for $\tilde{\alpha}$ is

$$\begin{aligned} P_t(\tilde{\alpha}) \quad &\stackrel{\nabla}{=} \gamma_G(t) \\ &= \Pr(\forall a \in G,\ a \text{ have at least experienced one crossover event up to time } t) \\ &\times \Pr(\forall a \in \bar{G},\ a \text{ have not experienced any crossover event up to time } t) \end{aligned} \tag{5.33}$$

If we assume independent crossover events (no interference assumption), we have:

$$\gamma_G(t) = \prod_{a \in \bar{G}} e^{-\rho_a t} \prod_{b \in G} (1 - e^{-\rho_b t}) \tag{5.34}$$

If we expand the expression in (5.34), it can be shown that [2]:

$$\gamma_G(t) = \sum_{K \subset L} (-1)^{|G-K|} e^{-\sum_{a \in \bar{K}} \rho_a t} \tag{5.35}$$

Let us denote $e^{-\sum_{\alpha\in\bar{K}}\rho_\alpha t}$ by $\Gamma_K(t)$:

$$\Gamma_K(t) \triangleq e^{-\sum_{\alpha\in\bar{K}}\rho_\alpha t} \tag{5.36}$$

Then we have:

$$\gamma_G(t) = \sum_{K\subset L}(-1)^{|G-K|}\Gamma_K(t) \tag{5.37}$$

Now using Möbius inversion formula, we have:

$$\Gamma_K(t) = \sum_{G\subset K}\gamma_G(t) \tag{5.38}$$

We notice that if we put $K = L$ in (5.36) we have, $\Gamma_L(t) = 1$ (in which by convention we assume that the empty sum is zero), therefore:

$$\Gamma_L(t) = \sum_{G\subset L}\gamma_G(t) = 1 \tag{5.39}$$

Equation (5.39), proves that:

$$\sum_{\tilde{\alpha}} P_t(\tilde{\alpha}) = 1 \tag{5.40}$$

$P_t(\tilde{\alpha})$ is the probability that loci $(x_{g_\kappa+1}, \ldots, x_{g_{\kappa+1}})$ have been together in the ancestral population, or in the other words it is the probability of having a *block structure* defined by $\tilde{\alpha}$ at time t. Notice that each configuration $\tilde{\alpha}$ represents a partition of the set of loci Θ to disjoint subsets of S (set of all loci under consideration) such that:

$$\Theta = \{B_i : B_i \subset S, B_i \cap B_j = \emptyset, \cup_i B_i = S\} \tag{5.41}$$

For example $\tilde{\alpha} = (010)$, when we have four loci, is representative of :$\{[1,2],[3,4]\}$ and $\tilde{\alpha} = (001)$ is representative of $\{[1,2,3],[4]\}$. Combining equations (5.31) and (5.34), and having in mind that there is a one to one correspondence between $\tilde{\alpha}$ and $G \subset L$ we have:

$$\begin{aligned} P_t &= \sum_{G\subset L}\prod_{\alpha\in\tilde{G}} e^{-\rho_\alpha t}\prod_{b\in G}(1-e^{-\rho_b t})R_G(P_0) \\ &= \sum_{G\subset L}\gamma_G(t)R_G(P_0) \end{aligned} \tag{5.42}$$

5.4.2 Solution to the Differential Equation

Equation (5.42) describes the probability of a haplotype at each instance of time, and it is derived based on the concept of blocks identical by descent. In this section we show that equation (5.42) in fact solves the differential equation (5.24). Therefore, the goal is to show that:

$$\begin{aligned}\dot{P}_t &= \frac{d}{dt}\sum_{G\subset L}\gamma_G(t)R_G(P_0)\\ &= \sum_{G\subset L}\dot{\gamma}_G(t)R_G(P_0) = \sum_{a\in L}\rho_a(R_a-\mathbf{1})(P_t)\end{aligned} \tag{5.43}$$

Denoting $\sum_{a\in L}\rho_a(R_a-\mathbf{1})(P_t)$ by $\Phi_{rec}(P_t)$ we have:

$$\begin{aligned}\Phi_{rec}(P_t) &= \Phi_{rec}(\sum_{G\subset L}\gamma_G(t)R_G(P_0))\\ &= \sum_{a\in L}\rho_a(R_a-\mathbf{1})(\sum_{G\subset L}\gamma_G(t)R_G(P_0))\\ &= \sum_{a\in L}\rho_a\sum_{G\subset L}\gamma_G(t)(R_aR_G(P_0)-\mathbf{1}R_G(P_0))\\ &= \sum_{a\in L}\rho_a\sum_{G\subset L}\gamma_G(t)(R_{G\cup\{a\}}(P_0)-R_G(P_0))\end{aligned} \tag{5.44}$$

absorbing $\{a\}$ in the subset G of L we will have:

$$\begin{aligned}\Phi_{rec}(P_t) &= \sum_{a\in L}\rho_a[\sum_{\substack{G\subset L\\ a\in G}}\gamma_{G\backslash\{a\}}(t)R_G(P_0)-\sum_{\substack{G\subset L\\ a\notin G}}\gamma_G(t)R_G(P_0)]\\ &= \sum_{G\subset L}[\sum_{a\in G}\rho_a\gamma_{G\backslash\{a\}}(t)-\sum_{b\in\bar{G}}\rho_b\gamma_G(t)]R_G(P_0)\end{aligned} \tag{5.45}$$

in which $G\backslash\{a\}$ is the set of all elements of G except $\{a\}$, $(G-\{a\})$, and $\bar{G}$ is the complementary set of G, i.e. $\bar{G}=L\backslash G$. On the other hand if we directly calculate $\dot{\gamma}_G(t)=\frac{d}{dt}\gamma_G(t)$ we will have:

$$\dot{\gamma}_G(t)=\sum_{a\in G}\rho_a\gamma_{G\backslash\{a\}}(t)-\sum_{b\in\bar{G}}\rho_b\gamma_G(t) \tag{5.46}$$

Therefore, from (5.45) and (5.46) we have:

$$\dot{P}_t=\sum_{a\in L}\rho_a(R_a-\mathbf{1})(P_t)=\sum_{G\subset L}\dot{\gamma}_G(t)R_G(P_0) \tag{5.47}$$

Therefore equation (5.42) is the solution to the differential equation (5.24) and it describes the change of haplotype probabilities at each instance of time in terms of recombination rates (parameter γ) and the frequencies in the founding population.

5.4.3 Block Distributions in The Absence of Hotspots

The value ρ_a depends on the physical distance of the corresponding loci to the link a, namely loci $a+1, a+2$. If they are far apart then ρ_a has a high value (high rate of crossover events between two loci) and if they are very close, ρ_a will have a small value. For now let us assume that the loci S are evenly distributed along the chromosome, i.e. the distance between consecutive loci are all equal. If we assume that there is an equal chance of experiencing a crossover event for all links, i.e.

$$\forall a \in L, \rho_a = \rho \tag{5.48}$$

equation (5.34) will be read as:

$$P_t(\tilde{\alpha}) = e^{-(|L|-|G|)\rho t}(1-e^{-\rho t})^{|G|} \tag{5.49}$$

In which $|L|$ is the total number of links. The assumption in (5.48) can be interpreted as the absence of hotspots of recombination on the chromosome under study [71]. Hence, assuming that there are no recombination hotspots, $P_t(\tilde{\alpha})$ depends only on the number of crossovers and not to the position of crossover events (which is reflected by configuration of $\tilde{\alpha}$). Figures (5.1,5.2) show the behaviour of the function $P_t(\tilde{\alpha})$ introduced in (5.49) for $|L| = 10$ and $\rho = 0.01, 0.001$ respectively.

5.4.4 Hotspots of Recombination

Now let us consider that the loci under study are evenly distributed but there are some links which are more susceptible to experience a crossover event (i.e. some more fragile links), for this case let us imagine that the set of links denoted by η are the set which are more fragile and $\rho_a = \rho' = \epsilon\rho, 0 \leq \epsilon, \forall a \in \eta$, and for $a \notin \eta, \rho_a = \rho$ in this case

Figure 5.1: A plot of equation (5.49), for $\rho = 0.01$

As it is shown in this figure , the probability that more links experience a cross over, increases with the number of links, and for recent generations (higher values of t), almost all the links have experienced at least one cross over event. For small values of t, as we see in the plot, the proba-

bility that a small number of links have experienced a crossover is higher than for others

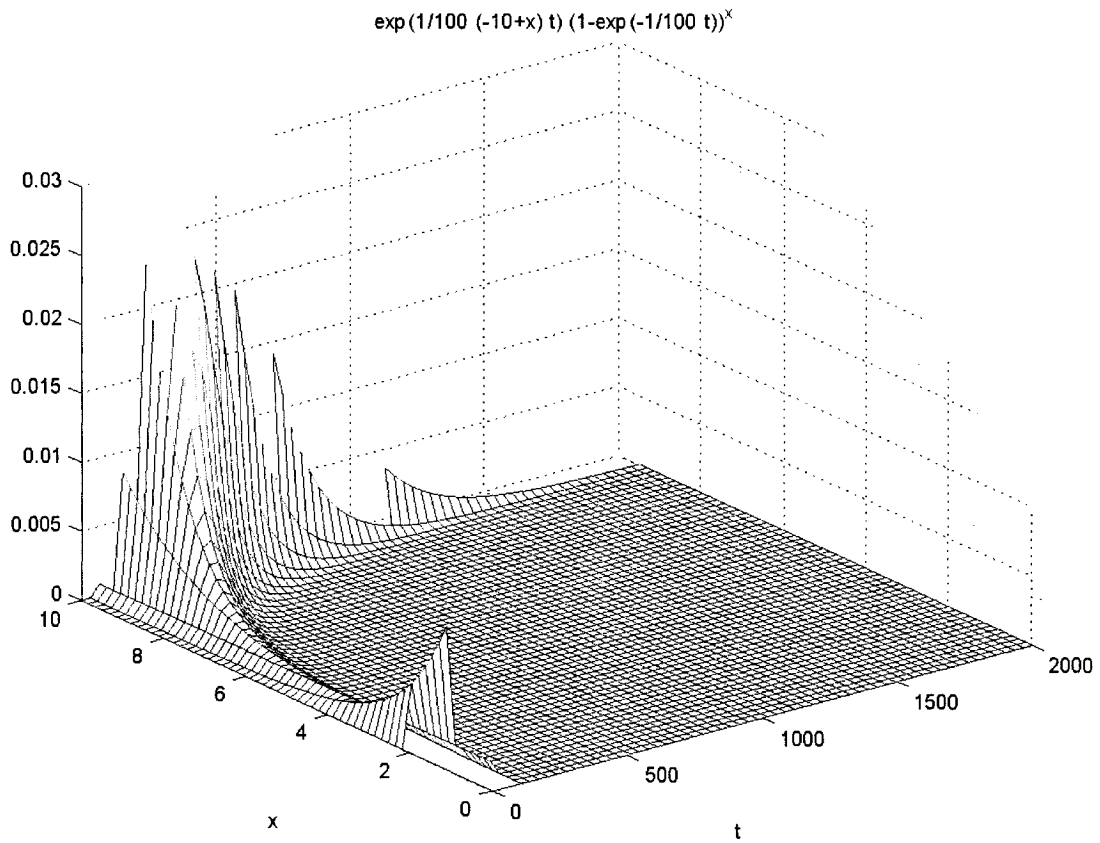
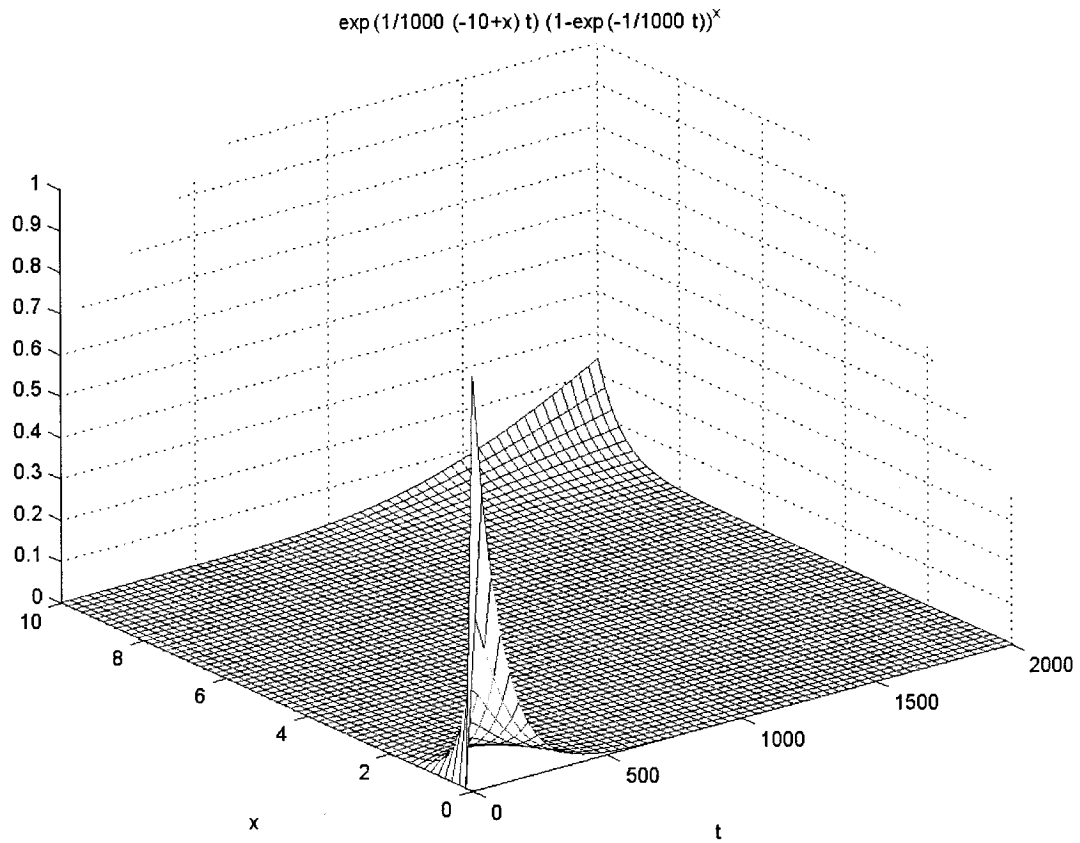


Figure 5.2: A plot of equation (5.49) for $\rho = 0.001$

In this plot, because of setting a small value for ρ, the probability that more links being hit by a crossover grows slowly. For smaller values of t at most 2 out of 10 links have experienced a cross over event



equation (5.34) will become:

$$P_t(\tilde{\alpha}) = \prod_{a\in(\bar{G}\cap\eta)\cup(\bar{G}\cap\bar{\eta})} e^{-\rho_a t} \prod_{a\in(G\cap\eta)\cup(G\cap\bar{\eta})} (1-e^{-\rho_a t})$$
$$= e^{-|\bar{G}\cap\eta|\epsilon\rho t} e^{-|\bar{G}\cap\bar{\eta}|\rho t} (1-e^{-\epsilon\rho t})^{|G\cap\eta|} (1-e^{-\rho t})^{|G\cap\bar{\eta}|} \quad (5.50)$$

We can rewrite equation (5.50) as:

$$P_t(\tilde{\alpha}) = e^{-\rho t[(|L|-|G|)+(1-\epsilon)(|G\cap\eta|-|\eta|)]} \left(1-e^{-\epsilon\rho t}\right)^{|G\cap\eta|} \left(1-e^{-\rho t}\right)^{|G|-|G\cap\eta|} \quad (5.51)$$

Note that if we put $\epsilon = 1, (\rho = \rho')$ then equation (5.51) will become same as (5.49).

$$e^{-\rho t(|L|-|G|)} \left(1-e^{-\rho t}\right)^{|G|} \quad (5.52)$$

As we see, in this equation the block distribution, is not only a function of number of crossover events, but also a function of number of hotspots and also the number of crossovers that have occurred exactly at the hotspots which is $(|\eta \cap G|)$.

5.5 A Graph Theoretical Interpretation

As we discussed in Section 5.1, each link configuration corresponds to a block structure on the haplotype under study, and we showed that there is a one-to-one correspondence between each value of $\tilde{\alpha}$ and the partitions Θ of loci. Equation (5.28) suggests that knowing the block structure, we can factor the joint probability $P(X_1, \ldots, X_n)$ into the product of some local functions defined on each block.

The assumption behind factorization in (5.28) is that the probability of finding a certain allele on each block, is only a function of the allele configurations on that block and not the rest of loci. In the other words:

$$P(X_i|X) = P(X_i|X_{B^{(i)}}) \quad (5.53)$$

in which $B^{(i)} \in \Theta$ is the block which the ith locus is located on that block and $B_k \cap B_l = \emptyset, \quad \forall k \neq l$ and $\bigcup_k B_k = S$. Hence equation (5.53) suggests that corresponding to each

block configuration (i.e. to each value of $\tilde{\alpha}$) we can define a neighbourhood system [34] $\mathcal{N}_\Theta$ such that:

$$\mathcal{N}_\Theta(i) = B^{(i)} \backslash \{i\}, \quad \forall i \in S \tag{5.54}$$

Therefore, corresponding to each value of Θ (hence each block structure) we have a Makov random field in which the corresponding neighbourhood system is defined by (5.54). Hence:

$$\Theta \Leftrightarrow \mathcal{G}(X, \mathcal{N}_\Theta)$$

We can divide the population under study $\mathcal{X}$, into sub-populations $\mathcal{X}_\Theta$, in which haplotypes belonging to the same sub-population share the same block structure. Therefore:

$$P(X) = \sum_\Theta P(X \in \mathcal{X}_\Theta) P(X | X \in \mathcal{X}_\Theta) \tag{5.55}$$

on the other hand if we know that $X \in \mathcal{X}_\Theta$ then we know that the joint probability of $X = (X_1, \ldots, X_n)$ can be factored to the product of local functions defined on the maximum cliques. On the other hand based on the definition of the neighborhood system in (5.54), each block corresponds to a maximum clique in the corresponding Markov random field. Ignoring the normalization factor we can write the joint probability in terms of the product of the local potential functions ϕ_B^Θ, hence:

$$P(X | X \in \mathcal{X}_\Theta) = \prod_{B \in \Theta} \phi_B^\Theta(X_B) \tag{5.56}$$

From (5.55) and (5.56) we have:

$$P(X) = \sum_\Theta P(X \in \mathcal{X}_\Theta) \prod_{B \in \Theta} \phi_B^\Theta(X_B) \tag{5.57}$$

On the other hand, as we discussed in Section 5.4, each partition Θ corresponds to a realization of $\tilde{\alpha}$ hence

$$P(X \in \mathcal{X}_\Theta) = P(\tilde{\alpha})$$

We can rewrite equation (5.28) as:

$$P(X|\tilde{\alpha}) = P(X|X \in \mathcal{X}_\Theta) = \prod_{B \in \Theta} P_0(X_B) \tag{5.58}$$

Hence regarding (5.57) and (5.58), if we interpret the local potential functions as the probability of X_B in the founding population (which B is a block corresponding to the partition Θ) we can consider a population $\mathcal{X}$ as a *mixture* of different Markov random fields, such that each Markov field corresponds to a certain block structure.

Chapter 6

Conclusion

Studying linkage disequilibrium among a set of loci is one the most important problems in genetic analysis. The main application of linkage analysis is in mapping complex traits such as common diseases in human populations [27, 56]. Modeling change of haplotype distribution is important for analysing available data from genome projects such as human genome project [75] and HapMap [74].

This thesis discusses the mathematical models describing change of haplotype distributions in a large random mating population in the absence of mutation, natural selection and migration. Two types of models, discrete time and continuous time model, are studied. A new mathematical framework for discrete time model is suggested and the special cases for one and two loci are studied. A simplified version of the continuous time model proposed by Baake and Baake [2], is presented and the solution to the differential equation which describes the change of haplotype distribution, is derived using the notion of BID's in genome [8, 25]. Finally, briefly the behaviour of the coefficients of the solution for the differential equation are studied under two hypothesises of existence of recombination hotspots and the absence of such a hotspots of recombination.

This chapter provides a summary of the thesis. In Section 6.1 a summary of the research is provided and key results are presented. Section 6.3 discusses the future

research directions and possible extensions of the models presented in this thesis.

6.1 Summary Of The Work

Chapter 4, discusses the discrete time model with the generalization of Bennett's [4] model proposed by Dawson [12]. In Section 4.1, the probabilistic model for representing populations of haploid individuals in the bi-allelic case, is made based on the theory of random sets. In this model, the genetic composition of each haplotype is described by a random subset of the set of all loci under study, ($\Gamma x \subseteq S$). In order to describe the recombination rates, another random set which represents the paternal segments of the haplotype is introduced. The haplotype frequencies as well as recombination rates are described in terms of the probability mass function of the corresponding random sets [45]. In Section 4.2, the recursion equation as is formulated in Dawson [12] and Geiringer [23], is derived in terms of probability mass function and commonality measure of random sets. Section 4.3, discusses the methods of solving the difference equation describing the change of haplotype frequencies. Dawson's [12] generalization of Bennett [4], is discussed and Bennett's measure of LD is evaluated for the case of one and two loci, and the results are compared with 'D' measure of pair wise LD. Section 4.4, discusses the applications of LD for mapping human genome and different measures of LD are briefly discussed.

Chapter 5, discussed the continuous time model for change of haplotype probabilities in a large population. we showed that considering the fact that there are blocks of genes identical by descent on genome, we can directly calculate the probability distribution of any haplotype X at any time t, and we showed that the calculated probability measure satisfies the differential equation describing the change of haplotype probabilities in population [2]. In addition, we showed that the coefficients γ_G of the solution for the differential equation , are in fact the probability of finding a specific block structure. In addition we showed that the probability of a block structure can be viewed as the

probability that a specific haplotype (considered as a sequence of random variables X_i) belongs to a Makov random filed corresponding to the block structure. The behaviour of the function $\gamma_G(t)$, is studied under the assumption of hotspots of recombination and uniformity of cross over events along the chromosome (no-hotspot assumption)

6.2 Applications of the Models

As we discussed in Section 4.4, one of the most important applications of LD analyses, is in mapping complex diseases in human populations. Most of the current models focus on pairwise analyses of the linkage [53, 44, 56]. However, for identification of complex trait diseases, simultaneous analysis of a set of loci is computationally more advantageous compared to the pairwise analysis [44].

The models for the change of haplotype distributions can be applied for both estimating the recombination rate among a set of loci, (knowing the haplotype frequencies in consecutive generations) and predicting the haplotype frequencies knowing the physical position of the loci in a haplotype.

Another problem of interest in population genetics, is finding the patterns of BID's in populations. Knowledge about the structure of BID's in a population helps us to define equivalence classes of loci on genome, i.e. if we know the block structure in a population of haplotypes, we can classify the polymorphic sites on genome based on the BID's. Every equivalence class can be identified by a single locus, which is located on the corresponding block. Hence, in association studies we can select these representative loci as the genetic markers [73].

One of the major disadvantages of the models described in this thesis, steams from the assumption of the infinite population size. Recent studies [70], and also simulated data, shows that in small populations, when we consider large number of loci, population size plays an important role in the structure of haplotype blocks and hence the change of haplotype frequencies over time. Therefore, in such cases, the infinite model cannot

be applied for data analysis.

6.3 Future Research Direction

A possible approach for finding the genes underlying complex diseases is considering the parameter $\rho(U)$ (Chapter 4) as the "degree of association" between the set of loci represented by random set U. Starting from this point, we might be able to use Dempster-Schafer theory for data mining [13, 61] to find the genes underlying complex traits such as common diseases in human populations.

Another problem of interest in population genetics is studying effects of various demographic and evolutionary factors on the haplotype block structure of a genome. These factors include genome size, recombination rate, population size, population subdivision and natural selection [8, 52]. One of the assumptions about the source of blocks in human genome, is hotspots of recombination [25], i.e. there are areas on genome which are more fragile and therefore recombination rate in those areas is higher than the other parts of the genome. Some other studies show that in a finite population even under the assumption of random distribution of crossovers, blocks can appear on genome [70]. The model described in Section 5.4, extended to the case of finite population size, can be a starting point to verify the validity of either of those assumptions.

Bibliography

[1] E. C. Anderson and J. Novembre, *Finding Haplotype Block Boundaries by Using the Minimum-Description-Length Principle*, Am. J. Hum. Genet. Vol.73,pp.336-354, 2003

[2] M. Baake and E. Baake, *An Exactly Solved Model for Mutation, Recombination and Selection*, Canad. J. Math. Vol. 55(1),pp.3-41, 2003

[3] E. Baake, *Mutation and recombination with tight linkage*, J Math Biol. Vol.42(5),pp.455-88, 2001

[4] J. H. Bennett, *On the theory of random mating*, Ann. Hum. Genet, Vol. 18,pp. 311-317, 1954

[5] S.J.E. Baird, N.H. Barton and A.M. Etheridge, *The distribution of surviving blocks of an ancestral genome*, Theoretical Population Biology, Vol.64, pp.451471, 2003

[6] I. B. Borecki and B. K. Suarez, *Linkage and Association: Basic Concepts*, Advances in Genetics, Vol. 42, pp.45-65, 2001

[7] N. H. Barton, *Estimating multilocus linkage disequilibrium*, Heredity Vol.84, pp.373-389, 2000

[8] N. H. Chapman and E. A. Thompson, *A model for the length of tracts of identity by descent in finite random mating populations*, Theor. Pop. Biol. Vol. 64, pp.141-150, 2003

[9] L. R. Cardon and G. R. Abecasis, *Using haplotype blocks to map human complex trait loci*, TRENDS in Genetics, Vol.19 no.3, pp.135-140, 2003

[10] D. L. Cohn, *Measure Theory*, Birkhäuser Boston, 1980

[11] C. D. Charalambous, *Stochastic Systems*, Lecture Notes, SITE, University of Ottawa, 2002

[12] K. Dawson, *The Decay of Linkage Disequilibrium under Random Union of Gametes: How to Calculate Bennett's Principal Components*, Theor. Pop. Biol., Vol.58,pp.1-20,2000

[13] A. P. Dempster, *Upper and lower probabilities induced by a multi-valued mapping*, Ann. Math. Stat. , Vol. 38, pp. 325339, 1967

[14] C.Damman, L.Yoon, *Discover Biology*,Sinauer Associates Inc. 2002

[15] B. Devlin , N. Risch, *A comparison of linkage disequilibrium measures for fine-scale mapping* Genomics, Vol.29, pp.311-322, 1995

[16] W. J. Ewens, *Mathematical Population Genetics: Theoretical Introduction*, Interdisciplinary Applied Mathematics , Vol.27, Speringer-Verlag, 2004

[17] I.M.H. Etherington, *Genetic algebras*, Proc. Roy. Soc. Edinburgh,Vol.59,pp.242-258, 1939

[18] P. Holgate, *Sequences of powers in genetic algebras*, J. London Math. Soc., Vol. 42,pp.489-496, 1967

[19] B. Fristedt and L. Gray, *A Modern Approach to Probability Theory* Birkh auser. 1997

[20] R. A. Fisher, *The Correlation between relatives on the supposition of Mendelian inheritance*, Trans. of the Roy. Soc. of Edinburgh, Vol.52, pp.399-433, 1918

[21] P. Fearnhead, *Haplotypes: the Joint Distribution of Alleles at Linked Loci* J. Appl. Prob. Vol.40,pp.505-512, 2003

[22] D. Graur and W. Li, *Fundamentals of Molecular Evolution*, Sinauer Associates Inc. 2000

[23] H.Geiringer, *On the Probability theory of Linkage in Mendelian Heredity*, Ann. Math. Stat., Vol.15,pp.25-57, 2000

[24] I. R. Goodman, R. Mahler and H. t. Nguyen, *Mathematics of Data Fusion* Kluwer Academic Publishers, 1997

[25] S.B. Gabriel, S.F. Schaffne, *et al*, *The Structure of Haplotype Blocks in the Human Genome*, Science, Vol. 296, 21 JUNE 2002

[26] R. Gorelick and M. D. Laubichler, *Decomposing Multilocus Linkage Disequilibrium*, Genetics Vol. 166, pp.1581-1583, 2004

[27] A. M. Glazier *et al*, *Finding Genes That Underlie Complex Traits*, Science, Vol. 298, pp.2345-2349, 2002

[28] R. C. Griffiths and S. Tavare, *Ancestral Inference in Population Genetics*, Stat. Science,Vol.9, Iss. 3, pp.307-319, 1994

[29] S. W. Guo, *Linkage disequilibrium measures for fine-scale mapping: a comparison*, Hum Hered, Vol.47(6), pp.301314,1997

[30] D. L. Hartl, *A Primer of Population Genetics*, Sinauer Associates Inc. 1997

[31] J. B. S. Haldane, *The combination of linkage values, and the calculation of distances between the loci of linked factors*, J. Genet, Vol.8, pp.299-309, 1919

[32] P. Holgate, *Characterisations of genetic algebras*, J. London Math. Soc. (2), Vol.6,pp.169-174, 1972

[33] G. H. Hardy, *Mendelian proportions in a mixed population*, Science, Vol.28, pp.49-50, 1908

[34] J. M. Hammersley and P. Clifford, *Markov fields of finite graphs and lattices*, Preprint, Univ. of California, Berkeley, 1968

[35] M. I. Jordan, *An Introduction to Probabilistic Graphical Models*, Lecture Notes, University of California, Berkely, 2003

[36] G. Karp, *Cell and Molecular Biology*,John Wiley & Sons, 2003

[37] E. Kreyszig, *Introductory Functional Analysis with Applications*, John Wiley & Sons, 1989

[38] Y.U.Lyubich , *Basic Concepts and Theorems of The Evolutionary Genetics of Free Populations*, Russ. Math. Survey,Vol.26,pp.51-123,2000

[39] Y.U.Lyubich, *Mathematical Structures in Population Genetics*, Biomathematics,Vol.22,Springer-Verlag,1992

[40] Y.U.Lyubich , *A new advance in the Bernstein Problem in mathematical genetics* Stony Brook IMS Preprint #1996/9, 1996

[41] S. L.Lauritzen and N. A. Sheehan *Graphical Models for Genetic Analyses*, TR R-02-2020, Department of Mathematical Sciences, Aalborg University, 2002

[42] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*, Addison-Wesley Pub. Co., 1994

[43] S.Lessard and J. Wakeley, *The two-locus ancestral graph in a subdivided population: convergence as the number of demes grows in the island model*, J. Math. Biol. , Vol.48, pp.275-292, 2004

[44] P.Luceka *et al.*, *Multi-Locus Nonparametric Linkage Analysis of Complex Trait Loci with Neural Networks*, Hum. Hered., Vol 48, pp.275284, 1998

[45] G.Matheron, *Random sets and integral geometry*, John Wiley and Sons,New York 1975

[46] N. Maniatis, A. Colling *et al*, *the first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis*, PNAS, Vol. 99, no. 4, pp.2228-2233, 2002

[47] N.E. Morton *et al*, *The optimal measure of allelic association*, PNAS, Vol. 98, no.9, pp.5217-5221, 2001

[48] G. Mendel, *Versuche über Pflanzenhybriden*, Verh. des Naturforschd. Vereines in Brünn, IV. Bd. Abhandlungen Brünn, pp. 3-47, 1866

[49] A. Morris and R. Mott, *Design and analysis of disease-marker association studies*, Research Report, University of Oxford, Bioinformatics Core, Summer 2004

[50] M. Nothnagel, R. FÜrst, and K.Rohde, *Entropy as a Measure for Linkage Disequilibrium over Multilocus Haplotype Blocks*, Hum Hered, Vol. 54, pp.186-198, 2002

[51] J. Ott, *Analysis Of Human Genetic Linkage*, 3rd edition,Johns Hopkins University Press, Baltimore MD, 1999

[52] S. Pääbo, *The mosaic that is our genome* Nature, Vol. 421, 2003

[53] J. K. Pritchard and M. Przeworski *Linkage Disequilibrium in Human: Models and Data*, Am. J. Hum. Genet. Vol. 69, pp.1-14, 2001

[54] I. Pe'er and J. S. Beckmann, *On the applicability of a haplotype map to un-assayed populations*, Hum. Genet. Vol.114, pp.214-217, 2004

[55] M. L. Reed, *Algebraic structure of genetic inheritance*, Bulletin of the American mathematical society, Vol. 34,pp.107-129,1997

[56] B. Rannala and M. Slatkin, *Methods for Multipoint Disease Mapping Using Linkage Disequilibrium*, Genetic Epidemiology Vol.19, pp.S71-S77, 2000

[57] B. D. Ripeley and F. P. Kelly, *Markov Point Processes*, J. Lon.Math. Soc. (2), Vol.15, pp.188-192, 1977

[58] A. R. Raji-Kermany, C. D. Charalambous, and D. Hickey, *Modeling Change of Haplotype Distributions in Random Mating Populations Using Random Sets*, CCECE 2004-CCGEI 2004, Niagara Falls, May 2004

[59] C. Shannon, *An Algebra for Theoretical Genetics* , PHD dissertation, MIT, 1940

[60] N. J. A. Sloane and A. D. Wyner, *Claude Shannon: Collected Papers*, IEEE Press, 1993

[61] G. Shafer, *A Mathematical Theory of Evidences*, Princeton University Press, 1976.

[62] A. Sommerfeld, *Mechanics*, Lectures on theoretical physics, Vol.1, Academic Press Inc., 1964

[63] H. T. T. Toivonen *et al*, *Data Mining Applied to Linkage Disequilibrium Mapping*, Am. J. Hum. Genet, Vol. 67, pp.133-145, 2000

[64] A. Wörz-Busecros, *Algebras in Genetics*, Lecture Notes in Biomathematics, Vol. 36, Spriner-Verlag, 1980

[65] S. Wright, *Evolution in Mendelian populations* Genetics,Vol.16, pp.97-159, 1931

[66] L.A. Wasserman, *Belief Functions and Statistical Inference* Can. J. Stat. Vol. 18 (3),pp.183-196,1990

[67] K. M. Weiss and A. G. Clark, *Linkage disequilibrium and the mapping of complex human traits*, TRENDS in Genetics Vol.18 No.1, January 2002

[68] W. Weinberg, *On the detection of heredity in man* (originally in German), Jh. Ver. Vaterl. Naturk. Wurttemb. Vol. 64, pp.368-382 , 1908

[69] E. P. Xing, R. Sharan, and M. I. Jordan, *Bayesian Haplotype Inference via the Dirichlet Process*, Lecture Notes in Bioinformatics, 2nd RECOMB sattelite workshop on computational methods for SNP's and haplotypes, 2004

[70] K. Zhang, J. M. Akey *et al. Randomly distributed crossovers may generate block-like patterns of linkage disequilibrium: an act of genetic drift* Hum Genet Vol.113, pp. 5159, 2003

[71] J. Zhang *et al, Genome wide Distribution of High-Frequency, Completely Mismatching SNP Haplotype Pairs Observed To Be Common across Human Populations*, Am. J. Hum. Genet., Vol:73, pp.1073-1081, 2003

[72] H. Zhao *et al, Assessing linkage disequilibrium in a complex genetic system. I. Overall deviation from random association*, Ann. Hum. Genet. Vol.63, pp.167-179, 1999

[73] Haplotype Map Project (Haplotype Mapping for human genome), http://www.hapmap.org

[74] *The International HapMap Project*, Nature, Vol.426, December 2003

[75] Human Genome Project, http://www.ncbi.nlm.nih.gov