Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Preparing SNP data for genetic linkage analysis

Anna Tzemach

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Preparing SNP data for genetic linkage analysis

Research Thesis

In Partial Fulfillment of The Requirements for the
Degree of Master of Science in Computer Science

Anna Tzemach

Submitted to the Senate of the Technion - Israel
Institute of Technology

Shvat, 5769 Haifa January, 2009

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

The Research Thesis or Project Thesis or Final Paper Was Done Under the Supervision of Prof. Dan Geiger in the Faculty of Computer science
The Generous Financial Help Of the Technion Is Gratefully Acknowledged.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Contents

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

List of Figures

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

List of Tables

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Abstract

Single nucleotide polymorphisms (SNPs) are stably inherited, highly abundant, and distributed throughout the genome. Current estimates are that SNPs occur as frequently as every 100-300 bases. This implies that in an entire human genome there are approximately 10 to 30 million potential SNPs. More than 4 million SNPs have been identified and the information made public. Their large number makes SNPs good candidates for linkage analysis, but introduces new problems that were not significant for highly polymorphic markers. Part of the problems originate in the genotyping process, others are the result of restrictions of current linkage software. In the present thesis we propose an algorithm for preprocessing SNP data and implement a tool, the SNPdistiller, that handles the complete process of preparing SNPs for linkage analysis, from the data after genotyping to the creation of an input file suitable for currently available linkage analysis tools. The tool begins by removing erroneous and unlikely SNPs from the data and continues by organizing SNPs into clusters that simulate behavior of high polymorphic and informative markers. The algorithm takes into consideration both the genetic data and the capabilities of the linkage analysis software. Experimental results demonstrate the performance of SNPdistiller on simulated and real datasets. The thesis ends by proposing further enhancements to the algorithms implemented in SNPdistiller

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

List of Symbols and Abbreviations

θ	Recombination value
g	Genotype
$L(x)$	Likelihood of x
$P(x)$	Probability of x
r^2 test	Test for measuring statistical association
cM	Centimorgan
INAD	Infantile NeuroAxonal Dystrophy
LD	Linkage Disequilibrium
LOD	Logarithm Of Odds
SNP	Single Nucleotide Polymorphisms
STR	Short Tandem Repeat

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Chapter 1

Introduction

Bioinformatics is the use of informatics tools and techniques in the study of molecular biology, genetic, or clinical data. The field of bioinformatics has grown greatly to cope with the large expansion of information generated by the mouse and human genome projects, as newer and more powerful generations of computers have emerged. It is now possible to employ the computing hardware and software at hand to generate novel methodologies for linking data across the different databanks generated by international projects and to derive clinical and biological relevance from all the information collected. The ultimate goal is to develop a computer program that can quickly and easily provide information correlating genes, their single nucleotide polymorphisms (SNPs), and the possible structural and functional effects on the encoded proteins, in relation to known information about complex diseases (Mah and Chia, 2007) In the present thesis we focus on processing and preparing SNP data for genetic linkage analysis tools.

Research suggests that 99% of an individual's DNA sequences are identical with those of another person. Of the remaining 1% of differences, over 80% are Single Nucleotide Polymorphisms (SNPs). An SNP is a single base substitution of one nucleotide with another, and both versions are observed in the general population at a frequency greater than 1%. Human DNA is comprised of only four chemical entities, A, G, C, and T, whose specific chemical order represents the alphabet of the genome. An SNP in individual "A" may have a

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

sequence of GAACCT, whereas in individual "B" the sequence is GAGCCT, the polymorphism being A/G. The most prominent public effort was spearheaded by The SNP Consortium (TSC) whose mission was to determine and map some 300,000 evenly spaced single nucleotide polymorphisms within the human genome.

Current estimates are that SNPs occur as frequently as every 100-300 bases. This implies in a complete human genome there are approximately 10-30 million potential SNPs. More than 4 million SNPs have been identfied, and the information has been made publicly available through the efforts of TSC and others. Many of these SNPs have unknown associations. Compilation of the public SNPs by NCBI has produced a subset of SNPs defined as a non-redundant set of markers used for annotation of reference genome sequences, and are thus referred to as reference SNPs (rsSNPs). Over 2.6 million SNPs have currently been assigned as rsSNPs, making SNPs excellent candidates for linkage analysis.

Genetic linkage analysis is useful for mapping disease genes. It enables the use of statistical tools to associate the functionality of genes with their location on the chromosome. This analysis uses a probabilistic model of inheritance of genetic materials and applies it to data in the form of pedigrees, where some of the individuals are annotated with information about the trait of interest and genetic makeup. As highly informative genetic marker maps have been developed, multipoint linkage analysis has become a crucial part in linkage analysis studies because of its superiority in pairwise linkage analysis for locating genes and detecting linkage. But the computational complexity required to perform such calculations increases exponentially because of the large number of markers that participate in the analysis, the high polymorphism of the markers under study, the size of the pedigree, and the number of untyped people in the pedigree. These factors greatly constrain the space and time requirements of existing programs. Some programs fail to run as the number of markers or the degree of polymorphism of the markers increase. Other programs can handle a large number of markers but can analyze only small to medium pedigrees.

Because of the great number of SNPs it is impossible to use them all for linkage analysis at once. At the same time, a small number of SNPs contains a small amount of information because of little variation between them. In

addition, errors occurring during genotyping make SNPs even less beneficial for linkage analysis.

In the present thesis we propose an algorithm for processing SNP data and implement a tool, SNPdistiller, that handles the complete process of preparing SNPs for linkage analysis, from preparing the data after genotyping to producing the input file for currently available linkage analysis tools. SNPdistiller begins by removing erroneous and unlikely SNPs from the data and continues by organizing SNPs into meaningful subgroups (clusters) that simulate high polymorphic and informative markers. We tested the tool on simulated and real data to measure its accuracy and performance. The thesis ends by proposing additional enhancements to the algorithms implemented in SNPdistiller.

The cleaning of SNP data is based on existing algorithms, reviewed in Chapter 4. The main contribution of the proposed solution is atomization of the process, which reduces the possible errors introduced by human handling. In Chapter 5 we propose new algorithms for clustering SNPs, using not only the genetic data and pedigree structure but also taking into consideration the complexity of the given data and adjusting to it. In Chapter 6 we demonstrate the performance of the algorithm on different sets of the simulated data, which are similar to the data structure of most real cases. In Section 6.2 we use real data to perform linkage analysis from beginning to end with SNPdistiller and compare the results with linkage analysis performed without SNPdistiller. After receiving promising results in the experiment, we run SNPdistiller on the SNP data set currently under research and propose genetic areas for future research. In Section 7.2 we go over some possible extensions to the proposed solution. Appendix A provides a detailed user manual for SNPdistiller.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Chapter 2

Genetic Background

Genetics as a science and the laws of inheritance originate in the works of Mendel which were published in the Journal of the Brunn Society of Natural Science in 1866 (Mc Farland, 1993) but were not adequately developed until the 20th century.

2.1 Basic Concepts in Genetics

DNA, or deoxyribonucleic acid, is the hereditary material in nearly all organisms, including humans. Nearly every cell in a person's body has the same DNA. Most DNA is located in the cell nucleus (nuclear DNA), but a small amount can also be found in the mitochondria. The information in DNA is stored as a code consisting of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA consists of about 3 billion bases, more than 99 percent of which are the same in all people. The order, or sequence, of these bases determines the information available for building and maintaining an organism, similarly to the way in which letters of the alphabet appear in a certain order to form words and sentences. DNA bases pair up (A with T and C with G), to form base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. Together, a nitrogenous base, sugar, and phosphate group are called a nucleotide. Nucleotides are arranged in two long strands that form a spiral called a double helix. The structure of the double helix is similar

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

to that of a ladder, with the base pairs forming the rungs of the ladder and the sugar and phosphate molecules forming the vertical members. An important property of DNA is that it can replicate itself. Each strand of DNA in the double helix can serve as a pattern for duplicating the sequence of bases. This is critical quality during cells division because each new cell must have an exact copy of the DNA present in the old cell.

A gene is the basic physical and functional unit of heredity. Genes are made up of DNA and act as instructions for making molecules called proteins. In humans, genes vary in size from a few hundred to more than two million bases. The Human Genome Project has estimated that humans have approximately 20,000 genes.

Each marker can appear in several different forms, called alleles, each potentially having a different physical expression. For example, three major alleles of the blood gene, A, B, and O interact in order to determine the various ABO blood types. These alleles determine detectable antigens on the surface of red blood cells. Alleles A, B, and O determine antigen A, antigen B, and the absence of either antigen. Every person has two copies of each gene, one inherited from each parent. Most genes are the same in all people, but a small number of genes (less than 1%) differs slightly among people. Alleles are forms of the same gene with small differences in their sequence of DNA bases. These small differences contribute to each person's unique physical features.

In the nucleus of each cell, the DNA molecule is packaged into thread-like structures called chromosomes. Each chromosome is made up of DNA tightly coiled many times around proteins called histones that support its structure. Each chromosome has a constriction point called the centromere, which divides the chromosome into two sections, or "arms." The short arm of the chromosome is labeled the "p arm" the long arm the "q arm". The location of the centromere on each chromosome gives the chromosome its characteristic shape, and can be used to help describe the location of specific genes. The genetic information each person carries in his or her cells is a combination of some genetic information from the mother and some from the father. Human cells contain 46 chromosomes divided into 23 pairs. In each pair, one chromosome came from the mother and the other from the father. Meiosis is the process in

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

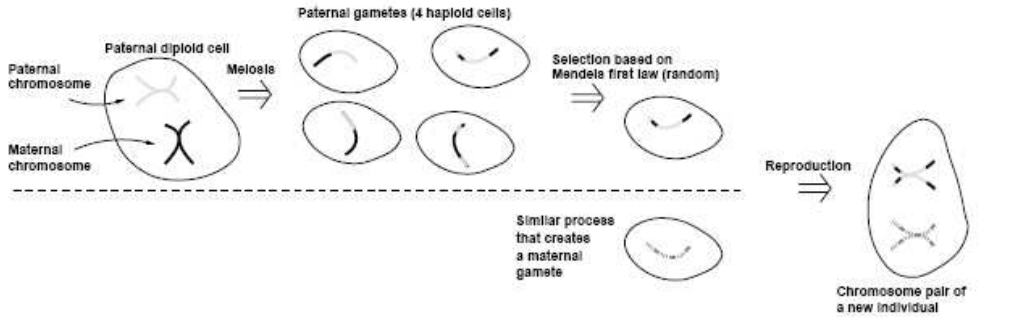

Figure 2.1: Illustration of meiosis for a single chromosome pair of one male and (partially) one female. The result of their meiosis is later combined in reproduction to form the chromosome pair of a new individual.

which a woman produces eggs and a man produces sperm. In meiosis, a normal diploid cell (having 2 copies of each chromosome) undergoes a special form of cell division to create haploid gametes (having one copy of each chromosome). The female gamete is called an egg, the male gamete a sperm. When an egg from the mother and a sperm from the father join, they fuse to form a new diploid cell called a zygote, a process called fertilization. Based on this new arrangement of genetic information, combining gametes from both the mother and the father, the zygote develops into a unique individual. Meiosis not only preserves the genome size of sexually reproducing eukaryotes but also provides three mechanisms to diversify the genomes of the offspring. During the meiosis the two chromosomes of the pair, inherited from the father and from the mother, are most likely recombined to form four new unique chromatides, as depicted in 2.1. This phenomenon is called recombination

2.2 Genetics mapping

Meiosis ensures that the number of chromosomes in a human individual is constant because only one of the four gametes is passed on to a new individual. This gamete is then copied in order to constitute a full chromosome, with two sister chromatides, for the new individual. Furthermore, meiosis and recombination are the reason that the offspring's chromosomes, which constitute a chromo-

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

some pair, are most likely not identical with any of the four chromosome sources of the two parents. Recombinations ensure diversity and uniqueness in the population. The probability that an odd number of crossovers occurs between two loci is referred to as the recombination fraction. In a sense, recombinations do not occur completely at random. The recombination fraction between two loci is related to their physical distance. The probability that a recombination occurs between two loci is called the recombination fraction, and it is denoted by θ. The distance on the genetic map (in units of Morgans) between two loci is defined as the expected number of crossovers occurring on a single chromosome between the loci. In other words, 1 Morgan is the distance over which, on average, one crossover occurs per meiosis. The relation between recombination and crossovers is as follows: an odd number of crossovers between two loci on a gamete corresponds to a recombination, whereas an even number of crossovers corresponds to a nonrecombination. If two loci are in close proximity of each other, at most one crossover occurs between them and therefore the recombination fraction is equal to the average number of crossovers in a gamete. Thus, for small intervals, the recombination fraction directly measures the genetic distance. This relation, however, is not valid for larger intervals. There are several theories, or map functions, that connect recombination fraction and genetic distance in Morgans, as shown in Table 2.1.

Map Function	The Model	Distance Function $x(\theta)$
Morgan	Linear model	$const * \theta$
Haldane	Poisson distribution	$-0.5(ln - 2\theta)$
Kossambi	Model of chiamal interference	$0.5(e^{4m} - 1)/(e^{4m} + 1)$

Table 2.1: Map functions from a recombination probability θ to a distance between two genes x .

Geneticists use maps to describe the location of a particular gene on a chromosome. One type of map uses the cytogenetic location to describe a gene's position. The cytogenetic location is based on a distinctive pattern of bands created when chromosomes are stained with certain chemicals. Another type

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

of map uses the molecular location, a precise description of a gene's position on a chromosome. The molecular location is based on the sequence of DNA building blocks (base pairs) that make up the chromosome. Geneticists use a standardized way of describing a gene's cytogenetic location. In most cases, the location describes the position of a particular band on a stained chromosome. A gene's molecular address pinpoints the location of that gene in terms of base pairs. Different groups of researchers often present slightly different values for a gene's molecular location. Researchers interpret the sequence of the human genome using a variety of methods, which can result in small differences in a gene's molecular address.

Genetic mapping, or linkage mapping, offers firm evidence that a disease transmitted from parent to child is linked to one or more genes. It also provides clues about which chromosome contains the gene and precisely where it lies on that chromosome. Genetic maps have been used successfully to find the single gene responsible for relatively rare inherited disorders, like cystic fibrosis and muscular dystrophy. Maps have also become useful in guiding scientists to the many genes that are believed to interact to bring about more common disorders, such as asthma, heart disease, and more. To produce a genetic map, researchers collect blood or tissue samples from family members where a certain disease or trait is prevalent. Using various laboratory techniques, the scientists isolate DNA from these samples and examine it for the unique patterns of bases seen only in family members who have the disease or trait. These characteristic molecular patterns are referred to as polymorphisms markers. The more DNA markers there are on a genetic map, the more likely it is that one will be closely linked to a disease gene and the easier it will be for researchers to zero in on that gene. Markers themselves usually consist of DNA that does not contain a gene, but they can reveal to the researcher the identity of the person to whom a DNA sample belongs. This makes markers extremely valuable for tracking inheritance of traits through generations of a family. Although there are several different types of genetic markers, the two most frequently used on genetic maps today are known as microsatellite maps (STRs) and SNPs (pronounced "snips"). Both types of markers are easy to use with automated laboratory equipment, so researchers can rapidly map a disease or trait in a

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

large number of family members. A much larger selection of SNPs is available for research than of tandem repeat markers. SNPs are DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered. For example a SNP might change the DNA sequence AAGGCTAA to ATGGCTAA. For a variation to be considered a SNP it must occur in at least 1% of the population. SNPs, which make up about 90% of all human genetic variation, occur every 100 to 300 bases along the 3-billion-base human genome. Two of every three SNPs involve the replacement of cytosine (C) with thymine (T). SNPs can occur in both coding (gene) and noncoding regions of the genome. Many SNPs have no effect on cell function, but scientists believe others could predispose people to disease or influence their response to a drug. Although more than 99% of human DNA sequences are the same across the population, variations in DNA sequence can have a major effect on how humans respond to disease; environmental insults such as bacteria, viruses, toxins, and chemicals; as well as drugs and other therapies. This makes SNPs of great value for biomedical research and for developing pharmaceutical products or medical diagnostics. SNPs are also evolutionary stable - not changing much from one generation to the next - making them easy to follow in population studies. Scientists believe that SNP maps will help them identify the multiple genes associated with such complex diseases as cancer, diabetes, vascular disease, and some forms of mental illness. These associations are difficult to establish with conventional gene-hunting methods because a single altered gene may make only a small contribution to the disease.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Chapter 3

Genetic Linkage Analysis

3.1 Introduction

Linkage analysis is the process of constructing genetic linkage maps. A genetic linkage map is a description of how several loci on a chromosome relate with respect to genetic distance. Hence, a genetic linkage map consists of several loci for which the order and relative distance (recombination fraction) are known. Genetic linkage analysis enables us to use statistical tools to associate the functionality of genes with their location on the chromosome. With the development of highly-informative genetic marker maps multipoint linkage analysis has become a crucial part in linkage analysis studies because of its superiority in pairwise linkage analysis for locating genes and detecting linkage. In the case of linkage analysis in humans, however, scientists face some unique challenges. One challenge is the long generation time in humans - generally 20-25 years. Another is that the choice of mate and the participation of individuals in the study cannot be controlled. For instance, a specific individual cannot be required to participate in the study, regardless of how important he or she may be to the eventual outcome. In addition, exposure to environmental conditions can vary among individual study participants.

Because many variables in human genetics are not under experimental control, identifying regions of interest in the genome that may harbor genes predisposing to human disease becomes a harder to analyze statistical process. The

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

use of mathematical models, formulas, and tools - all based on the basic rules of genetics - helps identify regions of interest. The computational complexity required to perform these calculations increases exponentially because of the large number of markers that participate in the analysis, the high polymorphism of the markers under study, the size of the pedigree, and the number of untyped people in the pedigree. These factors place significant constraints on the space and time requirements of existing programs. Some programs fail to run as the number of markers or their degree of polymorphism increases. Other programs can handle a large number of markers but can analyze only small to medium pedigrees

The data needed to perform linkage analysis is:

- A genetic linkage map
- One pedigree or more
- Information on the inheritance of markers
- Trait status for some or all pedigree members

3.2 LOD score and linkage disequilibrium

The information is used to produce evidence that the inheritance pattern of the trait resembles the inheritance pattern of a marker. Because of crossovers, the inheritance pattern of a trait does not always follow exactly that of a marker. For this reason we need a measure of how closely the inheritance patterns of a trait and a marker resemble each other. Such measures are expressed through scoring functions. The analysis is performed for several pedigrees on a subset of markers across the genome. If the scoring between the trait and a marker indicates potential linkage, that region is further analyzed by including more markers and/or people.

The LOD score, one such scoring function, is a measure of how plausible an observed set of data is, given a model (Ott, 1999). It is the logarithm of the odds of the observed data if linkage is assumed (recombination fraction < 0.5) compared to the null-hypothesis (no linkage) where the recombination fraction

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

is equal to $1/2$

Formally the LOD score is (Ott, 1999):

$$LOD(\theta) = \log_{10}(\frac{L(\theta)}{L(0.5)})$$

where L is the likelihood of a hypothesis H given some observation data F defined as: *probability of the observed data, which in this case, are the known phenotypes (x), given certain values for the unknown recombination fractions (θ)* . The likelihood $L = \sum_g P(x|g)p(g)$ performs a multiple sum over all possible genotype (g) combinations for all members of the pedigree. Maximum likelihood estimation is used to find a θ, that maximizes the LOD score for all markers.

The direct computation of likelihood is exponential both in the number of markers and of individuals. Two main approaches have been developed to compute pedigree likelihood exactly: Elston-Stewart (Elston and Stewart, 1971) and Lander-Green (Kruglyak *et al.*, 1995, 1996; Lander and Green, 1987). The complexity of the Elston-Stewart algorithm is linear in the pedigree size (for suffciently simple pedigrees) but exponential in the number of loci. By contrast, the Lander-Green algorithm is linear in the number of loci but exponential in the number of non-founders in the pedigree (non-founders are individuals whose parents are in the pedigree). Each approach is suitable for a different class of linkage problems, both unavailable to the analysis of a large of medium size pedigrees with large number of loci. These pedigrees are very common and therefore of great interest to research.

Researchers prefer pedigrees as large as possible and as many markers as possible because they are much more informative, less sensitive to input errors, and achieves higher p values. At the same time, multipoint analysis is sensitive to power loss due to misspecification of intermarker distances. Such misspecification is especially problematic when dealing with closely spaced markers, such as SNPs. Moreover, this model assumes linkage equilibrium between markers, which is not true for dense markers.

Linkage equilibrium occurs when a genotype present at one locus is independent of the genotype at a second locus. Linkage disequilibrium (LD) occurs

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

when two chromosome locations (two markers, two genes, two loci) are so close to each other that there is a lack of historical (ancestral) recombination events between them. The lack of crossover has several effects. If one of the locations is where the disease gene is, the marker that is in LD with the gene is "hitchhiked" ("dragged along") by the disease gene.. Thus, persons affected by the disease tend to have a certain marker allele value at this marker. LD is slightly different from "linkage": two linked markers (two markers "in linkage") lack current (as opposed to ancestral) recombination events. In the pedigree data we can still see the crossover between two linked markers, but with a low frequency. At the same time, we rarely see an actual crossover between two markers that are in LD in the pedigree data. Such crossover has occurred rarely and only in the past ("historically"). Haplotype frequencies can be incorrectly inferred when inter-marker LD is unaccounted for, leading to inflated multipoint LOD scores when parental genotypes are missing. Here, the haplotype frequencies are analogous to allele frequencies for a multiallelic marker, and the detrimental effects of allele frequency misspecification on linkage analysis have been well documented (Clerget-Darpoux *et al.*, 1986; Ott, 1977) Thus, when performing linkage analysis on SNPs one should select a sufficient number of markers or use a different approach.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Chapter 4

Error handling in SNP data

4.1 Introduction

Single nucleotide polymorphisms (SNPs) are stably inherited, highly abundant, and distributed throughout the genome. These variations are associated not only with diversity within and among populations, but also with individual responses to medication and susceptibility to diseases (De Wilde *et al.*, 2004). There are approximately 8 million human SNPs in public databases and many of them have been validated. SNPs greatly outnumber short-tandem-repeats (STRs) in the genome, allowing greater flexibility in the marker selection than would otherwise be possible. The large number of available SNPs enables fine mapping in suspected regions or even on the entire genome. SNPs are more stable from the evolutionary point of view because they do not change much from one generation to the next, making them easier to use in population studies. As a result, SNPs are the prevalent set in association studies and in linkage analysis of small families. If parents can be genotyped, a sparse map of SNPs should be sufficient. But when parental genotypes are not available, dense maps are the best candidates for linkage analysis because increasing the marker density significantly increases the amount of inheritance information extracted for some cases, by 20% (Evans and Cardon, 2004a). The Elston-Stewart algorithm cannot be applied on a large number of markers, and the Lander-Green algorithm cannot handle moderate and large families. Moreover, dense SNPs

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

are often in linkage disequilibrium, and none of the linkage algorithms take LD into consideration. But ignoring LD inflates the LOD score (Evans and Cardon, 2004b). SNPs raise additional problems that are not present in polymorphic markers. SNPs have a greater rate of genotype errors, which is an important factor and has a major impact on the power to detect linkage and association, especially where genotype data from parents is not available (Seaman and Holmans, 2005; Abecasis, Cherny, and Cardon, 2001).

Error rates are influenced by a number of factors, but are generally between 0.25% and 1% for microsatellite genotyping (Ewen *et al.*, 2000). Where family information is available, a proportion of genotyping errors can be detected as Mendelian inconsistencies, but this is more difficult for SNP markers with only two alleles (Gordon *et al.*, 1999). The SNPs have low informative content per each marker. Thus multipoint analysis is required, but this is complicated to perform on large families. Even approximate algorithms, such as Markov-Chain Monte Carlo (MCMC), often fail to converge on SNP data (Sieh *et al.*, 2005; Yang *et al.*, 2005).

We developed a tool, SNPdistiller, that makes sense of the data by grouping SNPs, eliminating the noise, and building polymorphic markers from the grouped SNPs. From the genome scan data of the SNP genotyping panels SNPdistiller produces datasets, as suitable as possible for linkage analysis, overcoming the problems listed above. Some of these problems are deriving from panel technology, while others from complication of linkage algorithms. The main innovation of the tool is to gather SNPs into clusters thereby creating new markers. These markers can be treated as polymorphic markers and used to produce less ambiguous results by linkage analysis. Even if only two point analyses are possible because of pedigree complexity, such clusters simulate multipoints by a single point, decreasing the complexity of calculation.

4.2 Error handling

The first problem that arises in working with SNPs is genotype errors produced by the technology of the genotyping panel. There are two types of genotype errors. The first type causes Mendelian inconsistency. It is easy to identify these

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

by checking for Mendelian consistency between parents and offsprings. There are two possible solutions to this problem. One is to run the extended Lange-Goradia algoritm (O'Connell and Weeks, 1999), which eliminates impossible genotypes. The algorithm is optimal on pedigrees, even if loops are presented. The other solution for detecting the inconsistency is to calculate the single point likelihood at each SNP. If there is an inconsistency at some SNP, the likelihood is zero. We used the Lange-Gordia algorithm to be able to identify the problematic genotypes in case the specific SNPs were in an interesting locus.

The second type of genotype error does not cause inconsistency. These errors are highly problematic because they can cause a false positive LOD score or, conversely, cause a LOD score that is too low at the linked locus. An example is a fully recessive disease where two heterozygous parents have affected the child. If the child is homozygous, the LOD score is high; otherwise it is low. Thus, if child had two alleles a/b, and because of genotype error we see a/a, this produces a false positive LOD score, as shown on Figure 4.1. With genotypes we never know for sure whether or not the values are correct. We must use statistical methods to detect such errors. The most common approach is to remove markers with an extreme number of recombination values, then determinate how many recombinations are extreme. We decided instead to count recombinations directly and compare the likelihood of the given data with the same data after withdrawing one genotype. To make this test as little sensitive as possible to noise, SNPdistiller takes the highest amount of SNPs allowed by the complexity of the pedigree. This method resembles the error checking in Merlin (Abecasis, Cherny, Cookson, and Cardon, 2002), but dynamically estimates the size of the dataset for likelihood calculation and is less dependent on user interaction. After calculating likelihood we must decide what result would be treated as an erroneous genotype. The decision was to remove all genotypes that decrease the likelihood of the data. The problem with such a conservative approach is that with a wrong starting point we can remove correct genotypes and leave erroneous ones. To avoid this problem, SNPdistiller removes SNPs from data if it finds more than several problematic genotypes at one SNP. We can remove SNPs without worrying about loosing mapping resolution because current genotype chips provides 250K SNPs, and

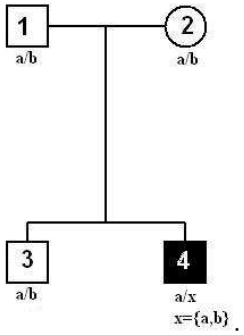

Figure 4.1: Example of genotype error that does not produce Mendelian inconsistency. Both genotypes a/a and a/b are legal, but have a very different influence on LOD score.

the number is growing. It is important to emphasize that the algorithm ignores the disease status of individuals, and therefore does not affect the significance of the LOD score. The detailed algorithm is presented in Algorithm 1.

Algorithm RemoveErroneousGenotypes($TestSNPs$,G)
Input: Genotypes G of the tested SNP and neighbors of the given SNP
Output: G without erroneous genotypes

1: **for each** g in G **do**
2: $r_{linked} = \frac{L(TestSNPs \backslash \{g\}|\theta)}{L(testSNPs|\theta)}$
3: $r_{unlinked} = \frac{L(TestSNPs \backslash \{g\}|\theta{=}0.5)}{L(testSNPs|\theta{=}0.5)}$
4: $r = \frac{r_{linked}}{r_{unlinked}}$
5: **if** $r > 1$ **then**
6: $TestSNPs \leftarrow TestSNPs \setminus \{g\}$
7: **end if**
8: **end for**

Algorithm 1: Algorithm for removing erroneous genotypes for some SNP.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

4.3 Removing uninformative and problematic data

After removing all genotype errors from the data we begin preparing the remaining data for the linkage analysis. The goal is to keep as much informative data as possible while making the complexity of the linkage analysis as simple as possible. First, SNPdistiller removes all SNPs that have only a single allele in the pedigree. Such SNPs do not contribute to linkage analysis but can increase the complexity of the analysis.

Observation 4.3.1 *Markers with single allele presents do not affect the LOD score.*

Rationale Recall that the LOD score is the ratio between the linked and the unlinked models. If there is only one allele present at the pedigree, we cannot recognize the recombinations and therefore it does not affect the LOD score.

Another problem we want to eliminate before proceeding to clustering are alleles with extreme frequencies. The LOD score is sensitive to extreme allele frequencies. In most cases the allele frequencies are taken from dbSNP. But because the frequencies depend on ethnicity, allele frequencies in the dbSNP often do not match the required population. For example, (Takeuchi, Yanai, Morii, Ishinaga, Taniguchi-Yanai, Nagano, and Kato, 2005) found moderate conservation across three ethnic populations in the genetic makeup of SNPs but not in the allele frequencies of SNPs or haplotypes. According to a previous report comparing African Americans and European Americans (Carlson, Eberle, Rieder, Smith, Kruglyak, and Nickerson, 2003), SNP data in the HuSNP set did not show significant correlations of minor allele frequencies (MAFs) between the Japanese and a reference population, and their SNP data in five genes have led to similar observations in Japanese, African Americans, and Caucasians. The misleading allele frequencies can inflate the LOD score because the disease gene tends to occur in the typed members of the pedigree, most often together with the most frequent allele, simply because that allele is common. When it is considered to be less frequent than it really is, its apparently frequent

occurrence with the disease allele appears as evidence for linkage (Ott, 1992).). As a result the LOD score is inflated, causing false positive results.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Chapter 5

Clustering SNP data

5.1 Introduction

It has been proposed to group single nucleotide polymorphisms (SNPs) into haplotype blocks containing a limited number of haplotypes (Gabriel *et al.*, 2002). Within each block, the haplotypes portion is to be tagged by a selected subset of SNPs. Each of proposed selection algorithms has advantages and disadvantages. We begin by reviewing the existing methods, then we present our own.

Several studies have shown that the human genome is structured with segments that are in strong Linkage Disequilibrium (LD) among relatively common SNPs, but between segments recombination has left little LD (Patil *et al.*, 2001). When SNPs are in strong LD, the alleles of a few SNPs on a haplotype suggest the alleles of the other SNPs, which as a result provide redundant information. Consequently, a modest number of common SNPs or other markers, selected from each segment would suffice to define the relevant haplotypes in presumably any population. This hypothesis has led to the HAPMAP project (http://www.hapmap.org), which aims at developing a map of common haplotype patterns throughout the genome in several ethnic populations.

Because of the relationship between SNP blocks and LD, several algorithms try to break SNPs into blocks based on LD and equilibrium within SNPs. To check whether LD is present between SNPs, an r^2 test is used. r^2 measures

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

statistical association, and there is a simple inverse relationship between this measure and the sample size required to detect whether the association between susceptible loci and SNPs has a direct relationship. r^2 takes a value of 1 if only two haplotypes are present. For two bi-allelic loci (alleles A,a at locus 1 and alleles B,b at locus 2), r^2 is defined as

$$r^2 = (p_{AB} * p_{ab} - p_{Ab} * p_{aB})^2 / (p_A * p_a * p_B * p_b)$$

Many of the models assume that there is no recombination within clusters and no linkage disequilibrium between clusters (Abecasis, Cherny, Cookson, and Cardon, 2002; Patil, Berno, Hinds, Barrett, Doshi, Hacker, Kautzer, Lee, Marjoribanks, McDonough, Nguyen, Norris, Sheehan, Shen, Stern, Stokowski, Thomas, Trulson, Vyas, Frazer, Fodor, and Cox, 2001; Zhang, Deng, Chen, Waterman, and Sun, 2002) or that there is some threshold for r^2 (Takeuchi, Yanai, Morii, Ishinaga, Taniguchi-Yanai, Nagano, and Kato, 2005). After finding separation SNPs to blocks, we must find the frequencies of each haplotype within a block. The simplest way to do it is to multiply the frequencies of alleles of which the haplotype is consists. This approach is highly problematic, however, because when two SNPs are in LD their allele are transmitted together. Using the previous example, assuming a strong LD and $p(a) \cong p(b) \rightarrow p(ab) \cong p(a)$ and definitely not $p(ab) = p(a) * p(b)$. The problem within correct estimation of haplotype frequencies is the same as that of the extreme allele frequencies described above.

Another approach is to find haplotype frequencies using the expectation-maximization (EM) algorithm. An EM algorithm is used in statistics for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables. EM alternates between performing an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and a maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found in the E step. The parameters found in the M step are then used to begin another E step, and the process is repeated. Theoretically, the algorithm finds the correct haplotype frequencies, but it requires large datasets. For example, the Merlin algorithm was tried

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

on 3158 individuals in 274 families simultaneously (Abecasis and Wigginton, 2005). Such a large number of individuals is often not available in linkage analysis. Even for dividing SNPs to blocks most of the algorithms need a large number of individuals and are therefore used mostly in association analysis.

5.2 Clustering in SNPdistiller

To divide SNPs into the clusters, we propose to use pedigree and genetic distance information. Based on the pedigree data we identify the SNPs that do have no recombinations between them. Using genetic information, we can learn the expected number of recombinations between two SNPs. We assume that SNPs that are relatively distant from each other are not in LD. For two SNPs to be in same cluster they must satisfy the following criteria:

- The most probable haplotype for this cluster is without recombinations.
- The number of expected recombinations in the cluster for the current pedigree is less than 1.

We use the Superlink (Fishelson and Geiger, 2002) software to find the most probable haplotype, which enables us to run SNPdistiller on a relatively large pedigrees with several SNPs. Because we are interested only in nonrecombinant haplotypes, we set the recombination values to 0. In this way we can reduce the complexity of the haplotype computation, generate results much faster, and run haplotypes analysis on large datasets. If there is recombination, we receive no possible haplotype; otherwise the haplotype is similar to the one we would obtain with the original recombination values (see observation 5.2.1). To find nonrecombinant haplotypes we retain all SNPs that sum of probabilities at founders haplotypes is less then 1 for recombination between them. If complexity allows, the haplotype is found. If complexity is too high or if no nonrecombinant haplotypes exist, SNPditiller decreases the number of SNPs in the proposed haplotype. The process is repeated until a haplotype is found or until only one SNP is left in proposed haplotype

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Algorithm FindLargestHaplo($SNPset$)
Input: A set of SNPs $SNPset$, so the probability for the recombination within the set is less than 1
Output: First largest possible nonrecombinant haplotype
Note: Procedure $FindHaplotype$ finds possible haplotype assuming zero distance between all SNPs. If no such haplotype exists, returns ϕ

```
1: Haplotype ← φ
2: while Haplotype == φ do
3:    Haplotype ← FindHaplotype(SNPset)
4:    if Haplotype == φ then
5:       SNPset ← SNPset \ {leftmostSNP}
6:    end if
7: end while
```

Algorithm 2: Finding largest possible nonrecombinant haplotype

Based on the haplotypes we construct a polymorphic marker, where each haplotype corresponds to some allele. The new allele frequencies are a function of the number of haplotypes. Because we cannot estimate allele frequencies using the EM algorithm, owing to the small number of individuals, we decided to assign equal frequencies to each haplotype. In a segment with n SNPs there should be 2^n haplotypes, but because of LD we actually received m haplotypes. We define the number of different alleles in each polymorphic marker, which represents the cluster $m + 1$. In this way we eliminate extreme frequencies, decreasing the number of false positive results. The cluster position on the chromosome is defined by the average position of the first and last SNP in the cluster. The cluster position is need for multipoint analysis.

Observation 5.2.1 *When all markers are linked, if nonrecombinant haplotype is possible, it is most probably one.*

Rationale The probability of recombination is always < 0.5. Therefore, to maximize likelihood the algorithm always chooses the nonrecombinant option if possible.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

5.3 Producing results

SNPdistiller receives as input two files in post or pre-makeped format. The outputs are ped and data files in post-makeped, which are Superlink compatible input files and can be used directly by Superlink or by Superlink-online (Silberstein *et al.*, 2006) for two-point and more complex analysis. The datafile also contains the description of each marker, including the name and position of the SNPs within it. If some cluster contains only one SNP, it is printed as cluster, but one should be aware of problems of linkage analysis with SNP (described above). This cluster can be easily removed from analysis with superlink-online. In addition, SNPdistiller produces a log file, which includes all the SNPs that were dropped, including a detailed explanation of it - whether the reason was Mendelian inconsistency, unlikely genotype, one allele SNP, or extreme frequency.

5.4 Detailed algorithm

Figure 3 shows the pseudo-code for the algorithm. The algorithm shows all steps from error detection to output processing. At first, SNPdistiller assumes that all SNPs are good and there are no clusters (lines 1-2). Line 4 calls the procedure described in Figure 1. After the "Remove Erroneous SNP" procedure runs, the algorithm proceeds with the remaining SNPs. The algorithm tries to create as large a cluster as possible within the given constraints. The probability of recombination is computed based on the distances between SNPs and the number of offsprings in the pedigree (line 8). The next line tries to create nonrecombinant haplotype by calling the procedure described in Algorithm 2. If it is successful, SNPdistiller calculates allele frequencies for the created cluster. If the algorithm fails to create the cluster within the given constraints, the cluster created in the previous iteration is added to the ClustersSet and a new cluster is initialized with the current SNP (lines 13,14). The algorithm then proceeds to the next SNP, until it iterates over all the SNPs. Line 17 produces the output.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Algorithm SNPdistiller($SNPset$)
Input: All available SNPs $SNPset$ in Superlink or Genehunter format
Output: Set of clusters $ClusterSet$ in Superlink format
Note: Procedure $caclculateFreq$ calculates cluster's alleles frequency based on number of alleles in the cluster or original allele frequency of SNP

1: $ClusterSet \leftarrow \phi$
2: $currentCluster \leftarrow \phi$
3: **for** each snp_i of $SNPset$ **do**
4: $SNPset \leftarrow$ `RemoveErroneousGenotypes`$(SNPset, snp_i)$
5: **end for**
6: **for** each snp_i of $SNPset$ **do**
7: $currentCluster \leftarrow currentCluster \cup \{snp_i\}$
8: $numExpRec \leftarrow$ `RecProb`$(currentCluster) * 2 * NumberNonFounders$
9: $haplotypes \leftarrow$ `FindLargestHaplo`$(currentCluster)$
10: **if** $numExpRec >= 1$ **OR** `NumRec`$(haplotypes) > 0$ **then**
11: $currentCluster \leftarrow currentCluster \cap \{snp_i\}$
12: `caclculateFreq`$(cluster_i)$
13: $ClusterSet \leftarrow ClusterSet \cup \{currentCluster\}$
14: $currentCluster \leftarrow \{snp_i\}$
15: **end if**
16: **end for**
17: $createNewInputFiles()$

Algorithm 3: Detailed SNPdistiller algorithm

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Chapter 6

Evaluation

6.1 Evaluation data

After developing the algorithm we tested it out on various datasets. We used some simulated and some real datasets. Using simulated data, we evaluate the performance of SNPdistiller and of a regular analysis algorithm under a variety of analysis conditions such as pedigree size, SNP density, and LD. Next, we ran SNPdistiler on two real datasets. The first dataset was derived from an INAD study (Khateeb *et al.*, 2006),where the affected locus had been already discovered. The second dataset is still under genetic research.

6.2 Simulated Datasets

We created several datasets to test our algorithm. The advantage of simulated data over real one is that we test sensitivity to some specific parameters while controlling the others. Each experiment was designed to check another input parameter, such as pedigree size, SNP density, and LD. The main task was to verify that SNPdistiller succeeds in determining the affected locus. Another issue had to do with the accuracy of the result: was the maximum LOD score achieved exactly on the defined marker or at some distance from it. In addition, we wanted to determine whether SNPdistiller produces new false positive results or reduces existing ones. For each set we created two segments: one unlinked

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

from the defined locus and another one linked to it. Depending on the set, we choose two-point or multipoint analysis, or both. Next, we compared the results of the LOD score computed directly on the SNPs with the LOD score computed on the clusters produced by SNPdistiller. All datasets were created using the Markerdrop program (Thompson, 1995, 2000; Thompson and George, 2003). Markerdrop simulates marker data at markers linked to a potential trait locus. The user must specify whether the marker data simulation is to be conditional on a trait model or on an inheritance pattern at the trait locus. We chose the marker data simulation to be conditional on a trait model, providing trait locus allele frequencies, genotypic penetrances, and the map position of the trait locus. Phenotypic trait data were provided as the affection status of each individual in the pedigree file. An inheritance pattern at the trait locus was simulated from the trait data, which became the trait model on which the markers were simulated. We used the full recessive model in all experiments. The Markerdrop program do not simulates genotype errors, but allowed us to specify the number of individuals to be typed. In all experiments we typed 100% of the last generation and 80% of the generation before last, a frequent situation in most genetic research.

Experiment 1. Small inbred pedigree

For the first experiment we used a small pedigree that can be analyzed using one of the existing linkage programs based on the Lander-Green algorithm. We wanted to compare in this way the performance of SNPdistiller with standard linkage analysis. Our expectation was to obtain about the same results with and without SNPdistiller. The simulated data included no genotype errors, and therefore an error detection algorithm would not have resulted in any improvement. Clustering would not have contributed either to improved results because the experiment data allowed full multipoint analysis. Therefore, the main objective of the experiment is to check whether SNPdistiller can detect the defined locus with same accuracy as does standard genetic analysis. We simulated genetic data for the inbred pedigree with 12 individuals. The pedigree is shown on Figure 6.1. We used the Morgan Markerdrop program to simulate

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

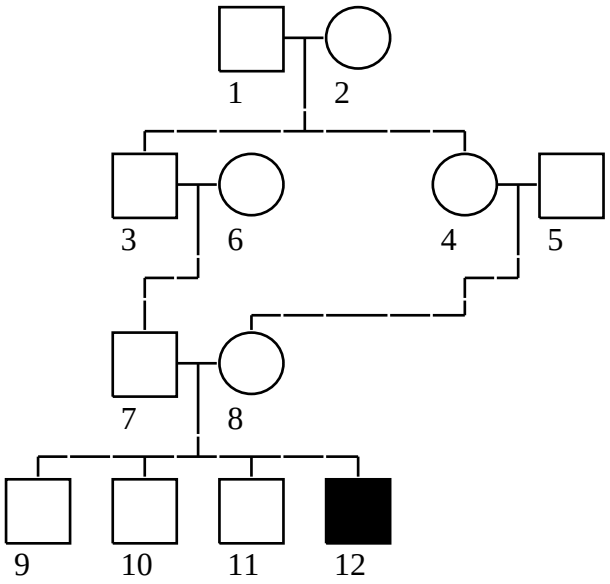

Figure 6.1: Four generation pedigree for simulation algorithm on small pedigree suitable for HMM data

two datasets, one with SNPs every half centimorgans and another one with SNPs every 0.1 centimorgan. For each density data was simulated for two segments, one linked and one unlinked. Each dataset was analyzed separately. For each dataset we ran regular multipoint analyses in addition to the multipoint analysis on the cluster produced by SNPdistiler. Both options produced about the same results. The LOD score of the cluster was slightly lower, by 0.03%. Generally, the results met out expectation. Exact results are shown in Table 6.1. Note that for dense SNPs neither analysis succeeded in pointing exactly to the correct loci. Both analyses pointed 1cM away, but in each analysis the mistakes was in a different direction.

Experiment 2. Medium size pedigree

The purpose of the second experiment was to check the behavior of the algorithm on a moderate size pedigree, where up to 4-point analysis is possible. The advantage of these pedigrees is the informativeness of a single marker. In

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

			Multi point			
Num SNP	Density	Marker	Real		Cluster	
	cM	linked	marker	LOD	marker	LOD
100	0.5	20	18	1.535180	18-20	1.530790
250	0.1	100	56	1.530962	152-159	1.530185
100	0.5	unlinked	out of map	-3.890934	out of map	-3.486398
250	0.1	unlinked	out of map	-4.552607	out of map	-4.710304

Table 6.1: Small Inbred Pedigree Experiment. Note that for dense SNPs neither analysis succeeded in pointing exactly to the correct loci. Both analyses pointed 1cM away, but in each analysis the mistakes was in a different direction.

addition, it is possible to run multipoint analyses to validate high scores, or on the contrary, to reject false positive results. In this experiment we expected two-point analysis on clusters to produce more clear results than regular two-point analysis, and multipoint analysis to provide precise answers about ambiguous two-point results. For this experiment, we used an inbred pedigree with 61 individuals (see Figure 6.2). As in the previous experiment, we simulated two datasets, one with SNPs every half centimorgan and another one with SNPs every 0.1 centimorgan. At both linked chromosomes, the affected locus was linked to 10 cM from the start. We ran two-point analyses both for SNPs and for the cluster and were able to see clearly that at the exact locus position clusters scored about the same LOD as SNPs. But in the adjacent clusters the LOD score was definitely higher than that of the SNPs. This is important in real datasets because the high adjacent LOD score increases the confidence that the point is not a false positive. In dense SNPs the LOD score produced by clusters, was much smoother and more stable from point to point then analysis directly on SNPs. We did obtain one false positive result of 1.4331, but 3-point analysis at this point produced an LOD score of -2.9318. Detailed two-point results are shown in Figure 6.3.

Experiment 3 - large size pedigree

The last experiment with pedigree size was with a pedigree suitable only for two-point analysis. This type of pedigrees is most interesting for the clustering

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

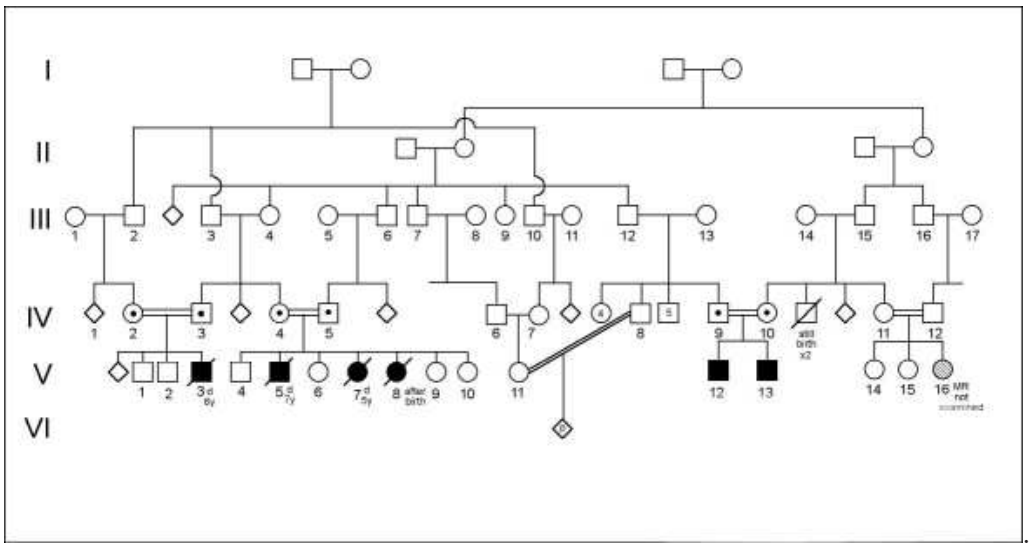

Figure 6.2: Medium size pedigree.

because it allows obtaining close to multipoint analysis results, where only two-point analysis was possible before. For these pedigrees we expected to eliminate at least some of the false positive results, an important issue for these pedigrees because where no multipoint analysis is available one must trust his intuition whether to perform more fine mapping or sequencing on a defined area. SNPdistiller provides an additional tools to make this judgment. We simulated SNP data as previously on an inbred pedigree with 115 individuals (Figure 6.4. Clustering removed some false positive results and also removed some dropes in LOD score in linked are. If it was real data, the reseachers had much easier decision to make about were to look for the affected gene. Detailed two point results are shown at Figure 6.5.

Experiment 4. LD and errors within data

In our last experiment we wanted to check how SNPdistiller contends with LD. We simulated data with the same pedigree structure as in Experiment #2, but this time we did not simulate SNPs with a constant gap between them but grouped them by trios of SNPs. We simulated 4 datasets. Two with a distance of 1.5 cM between adjacent groups and two with a distance of 0.3 cM

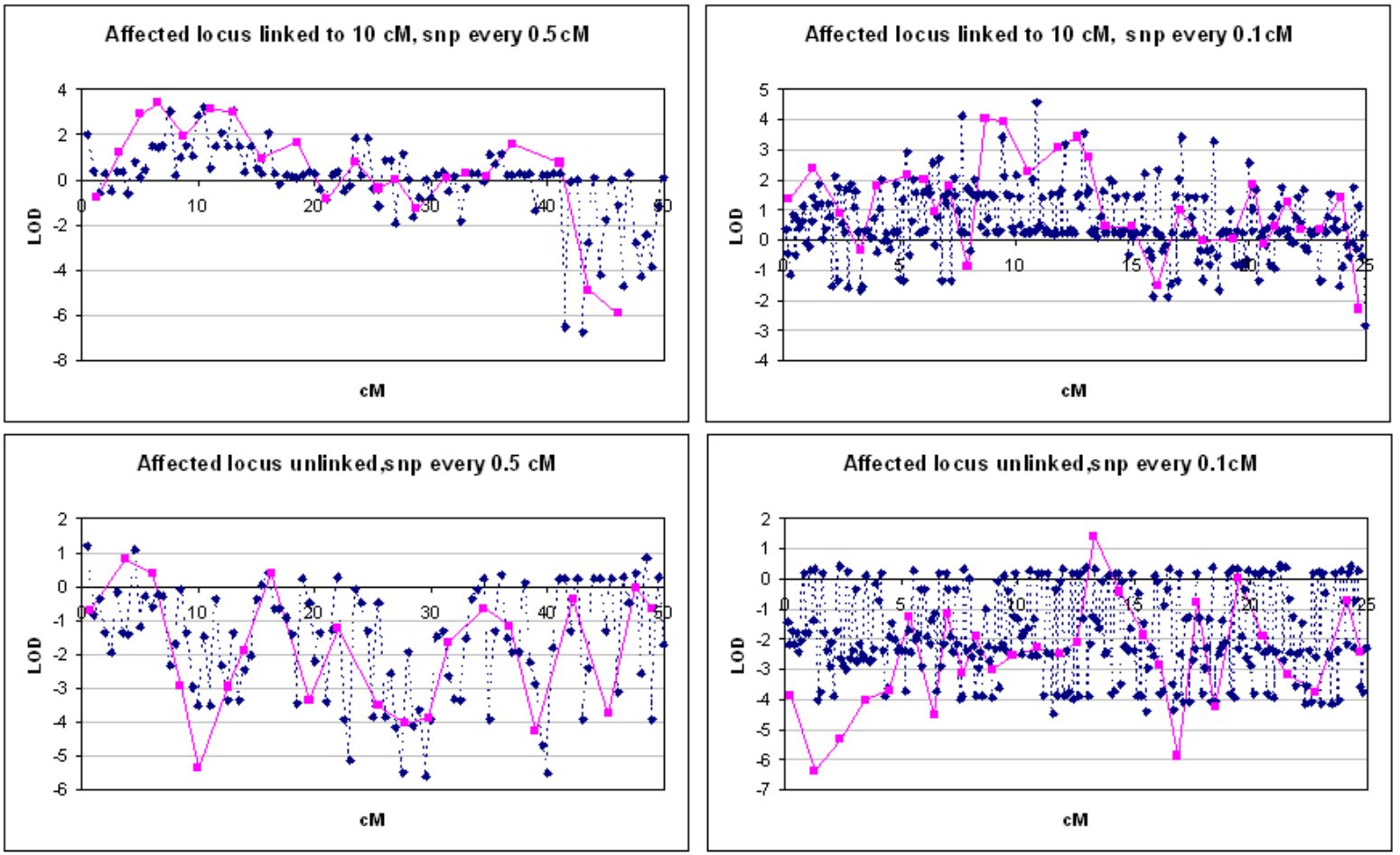

Figure 6.3: Two-point analysis results for medium pedigree. At both linked chromosome the affected loci are linked 10 cM from the start. The dotted line shows the SNP LOD score and the solid lines the cluster LOD score.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

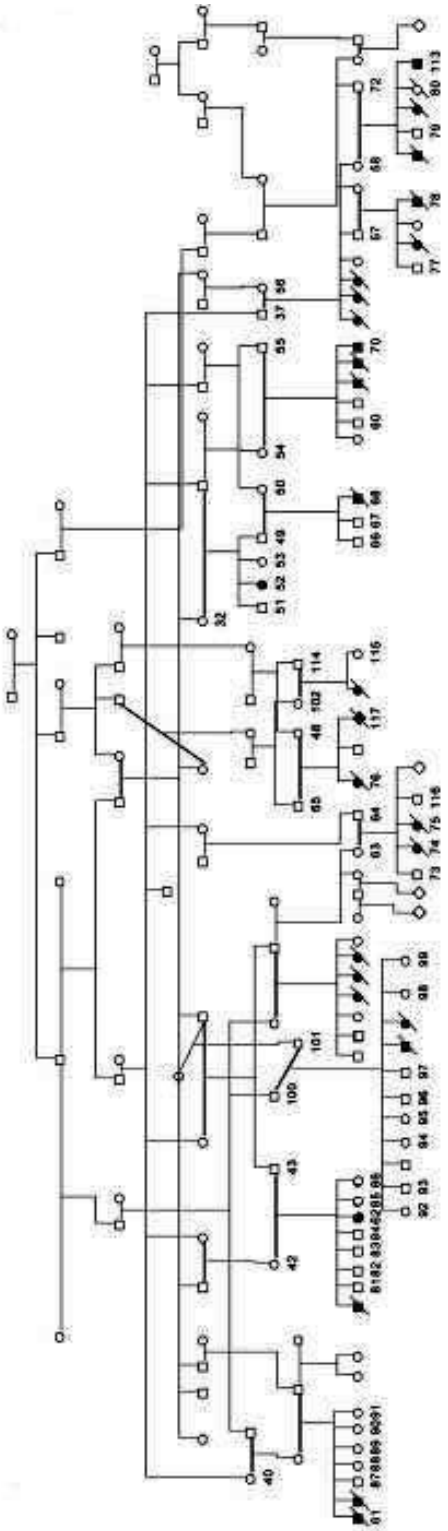

Figure 6.4: Large size pedigree.

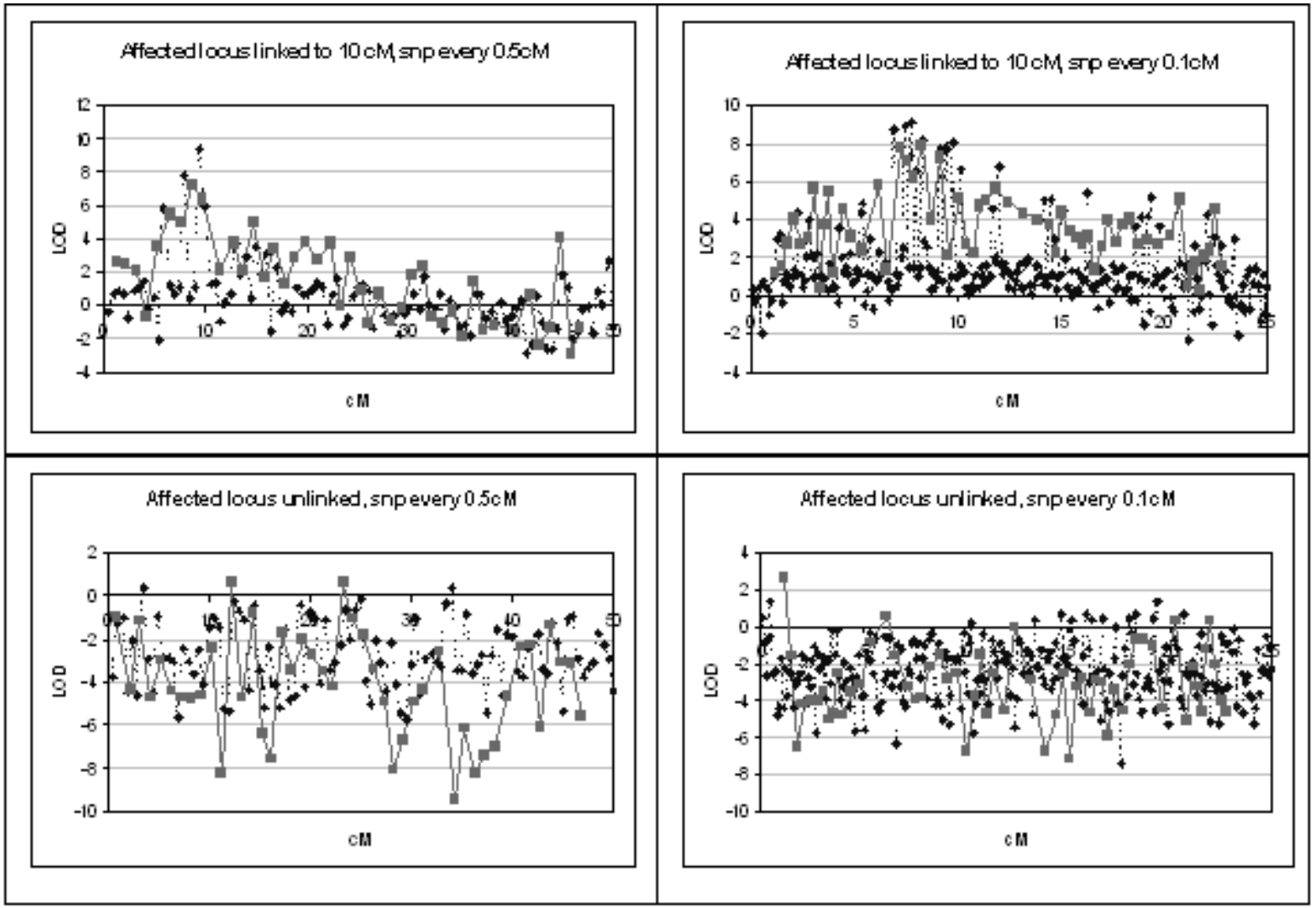

Figure 6.5: Two-point analysis results for large pedigree. At both linked chromosome the affected loci are linked 10 cM from the start. The dotted line shows the SNP LOD score and the solid lines the cluster LOD score.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

between adjacent groups. Again we used two sets with unlinked disease and two sets with linked disease, this time at 9.9 cM. But the input to SNPdistiller indicated that each SNP was 0.5 or 0.1 cM away from the adjacent SNPPs, depends on the dataset. In this way we tried to simulate LD by creating SNPs that are inherited together but have some genetic distance between them. It also simulated inconsistency between input data and real data. After creating the datasets we ran SNPdistiller on all 4 datasets. We performed two-point analysis on the resulting clusters. To compare results, we also performed two-point analysis directly on SNPs. The clusters produced about the same LOD score at the linked locus as SNPs, but in other locations the results were less peak and more unambiguous. The detailed two-point results are shown in Figure 6.6.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

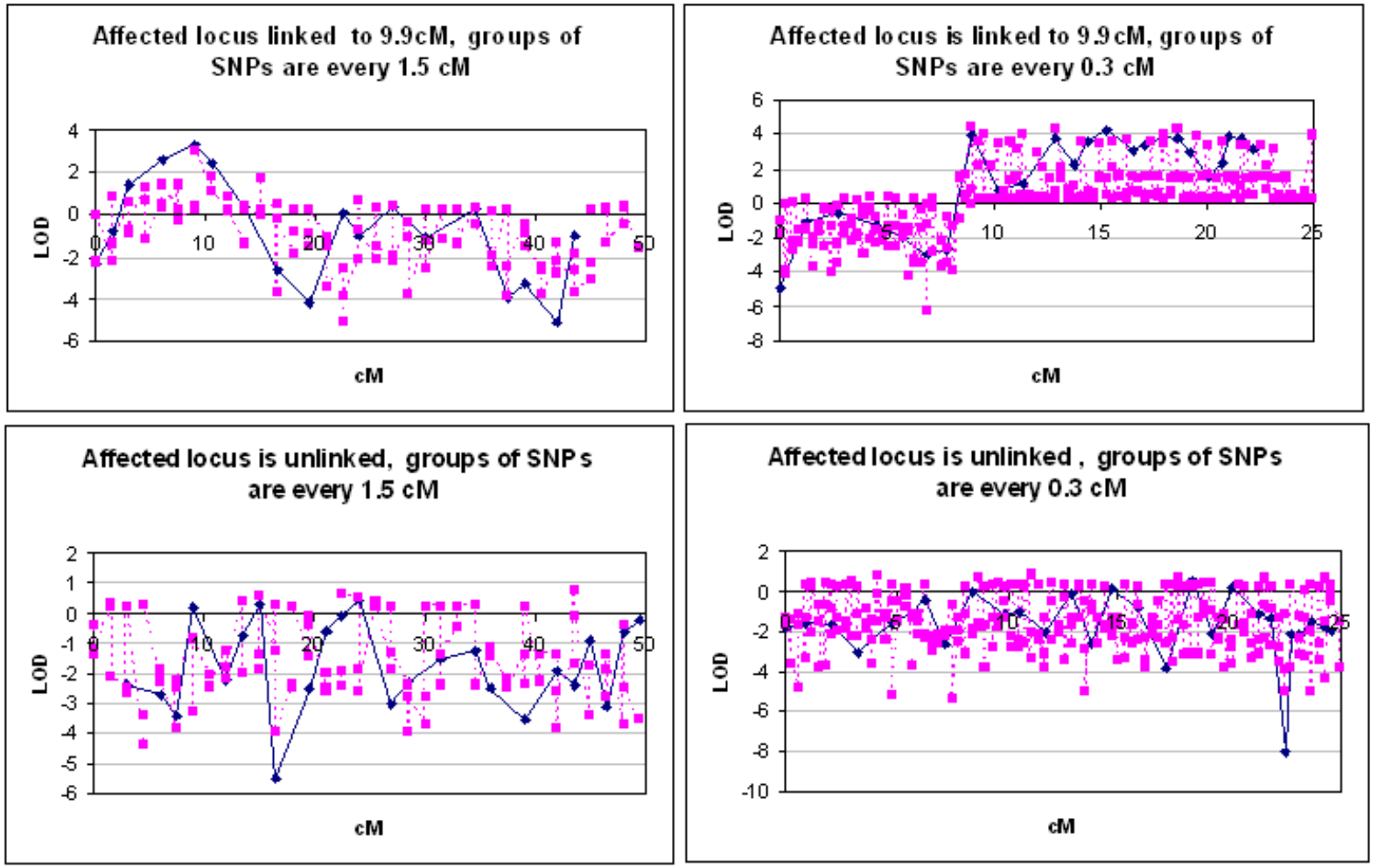

Figure 6.6: Two-point analysis results for experiment #4. At both linked chromosomes the affected loci are linked at 9.9 cM from the start. The dotted line shows the SNP LOD score and the solid lines the cluster LOD score

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

6.3 Real Datasets

PLA2G6 Mutation Underlies Infantile Neuroaxonal Dystrophy

The dataset was originally reported by Khateeb at el (Khateeb, Flusser, Ofir, Shelef, Narkis, Vardi, Shorer, Levy, Galil, Elbedour, and Birk, 2006) in their studies of the mutation that underlies Infantile Neuroaxonal Dystrophy (INAD) using superlink-online. We chose this study because it has two obvious advantages: the pedigree has a typical size for linkage analysis and the researches have already discovered the gene location. The typicality of this dataset made it a good test case for SNPdistiller, and the location of the gene could help us in determinating the accuracy of the algorithm. For this dataset, we performed linkage analysis with and without SNPdistiler and compared the results.

Disease description

Infantile neuroaxonal dystrophy (INAD) (MIM 256600) is a "neurodegenerative disease characterized by pathologic axonal swelling and spheroid bodies in the CNS. Onset is within the first 2 years of life, and the disease culminates in death by age 10 years. Most patients with INAD show a progressive disorder with motor and mental deterioration,cerebellar ataxia, marked hypotonia of the trunk with later bilateral pyramidal tract signs, spastic tetraplegia, hyperreflexia, and early visual disturbances. Seizures are not reported. Electroencephalography shows characteristic high-voltage fast rhythms, with electromyography results consistent with chronic denervation. All patients have abnormal visual evoked potentials. T2-weighted magnetic resonance imaging (MRI) typically shows cerebellar atrophy with signal hyperintensity in the cerebellar cortex and, occasionally, hypointensity in the pallida and substantia nigra Pathological hallmarks are marked neuroaxonal dystrophy, severe cerebellar atrophy, and degeneration of the lateral corticospinal tracts. Axonal endings show spheroid bodies, often detectable in the skin and conjunctivae." (Khateeb *et al.*, 2006)

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Pedigree and genetic data description

Two consanguineous inbred Israeli Bedouin kindred presented with a similar, autosomal recessive, severe progressive neurodegenerative phenotype affecting a total of 8 individuals Figure 6.8. Of the 8 patients, 6 were available for detailed clinical and molecular analyses. Genomewide linkage analysis using 400 polymorphic markers was conducted with five affected individuals of kindred 1. The individuals tested were X:1, X:6, XI:1, XI:2, and X:15, depicted in Figure 6.8A. But fine mapping of those regions in DNA samples of 20 family members ruled out linkage to those loci. Geneticist performed a genomewide linkage analysis of the same five patients and two obligatory carriers (parents of affected individuals), using the Affymetrix 10K SNP arrays. In all five individuals, 226 loci exhibited homozygosity, of which they selected five candidate loci on the basis of two criteria: the region of homozygosity was informative (homozygous in the affected individuals and heterozygous in the obligatory carriers), and it was 12 Mb. Fine mapping of the five loci by testing the 20 available DNA samples of kindred 1 with the use of polymorphic markers ruled out linkage to four of the five loci (data not shown). The fifth locus, on chromosome 22q13.1 (harboring SNPs rs763668 and rs139897), was bordered by recombination events in affected individuals at adjacent SNPs rs132692 and rs926299, implying a homozygosity region of 2.58 Mb.

Experiment description

The purpose of the experiment was to determinate whether SNPdistiller can reduce the number of regions for fine mapping or, alternatively, find affected loci without using fine mapping at all. We began by running SNPdistiller on all SNP data, which produced 1839 clusters (out of 100K SNPs) with 3-7 alleles per cluster. From this point on we used clusters as if they were regular polymorphic markers. First, we calculated the LOD score of the disease locus linked to each cluster. The highest LOD score of 2.1625 was achieved on chromosome 22. The closest score to this LOD was received on chromosome 8, and it was 1.9534. But all adjusted clusters showed negative LOD (-1.4229 and lower). The clusters produced were approximately 3cM apart, so we expected adjacent

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

clusters to support the relatively high LOD scores. On chromosome 11 there were several continuous clusters with LOD scores between 0.3245 and 1.1302. We performed multipoint analyses for these, using as many cluster as possible and we obtained high LOD scores on chromosomes 22 and 11, but only on chromosome 22 it was higher then 3 (see Table 6.2)

SNPdistiller statistics

For each run SNPdistiller creates a log file that contains SNPdistiller run statistics. Out of a total of 10K SNPs 20 SNPs had Mendelian errors and 1901 had only one allele in a given pedigree. Another 9886 were marked as unlike genotypes, and therefore were set to untyped. Because of unlike genotypes, total of 130 SNPs were dropped and not used in the clustering process.

Result comparison

SNPdistiller performed 3 basic steps. First, it eliminated all errors and non-informative data. This step is necessary in all linkage analysis and it used to be performed manually until now. Second, it removed all possible inconsistencies. The final step was clustering. After each step we compared the LOD score with the two-point and multipoint analyses. We chose to compare results on chromosome 22 because we knew where the peak was supposed to be (Figure 6.7)

Result discussion

We want to achieve two goals by running SNPdistiller. First, ensure the correctness of the algorithm by estimating the loci linked to the disease. Second, simplify the regular process of genomewide linkage as much as possible, given that manual search for SNPs with some property, such as homozygousity or informativeness, over thousands of SNPs can be time consuming and open to human error. The experiment has achieved both objectives. Multipoint analysis clearly shows the defined locus. We also removed many false positive results in two-point analysis and have almost not added any new ones. After running the multipoint analysis we were left with only one chromosome. Using regular

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

	LOD Score	
Chr	Two Point	Multi Point
22	2.16	3.81
11	1.13	2.4
8	1.95	-3.0112

Table 6.2: LOD score

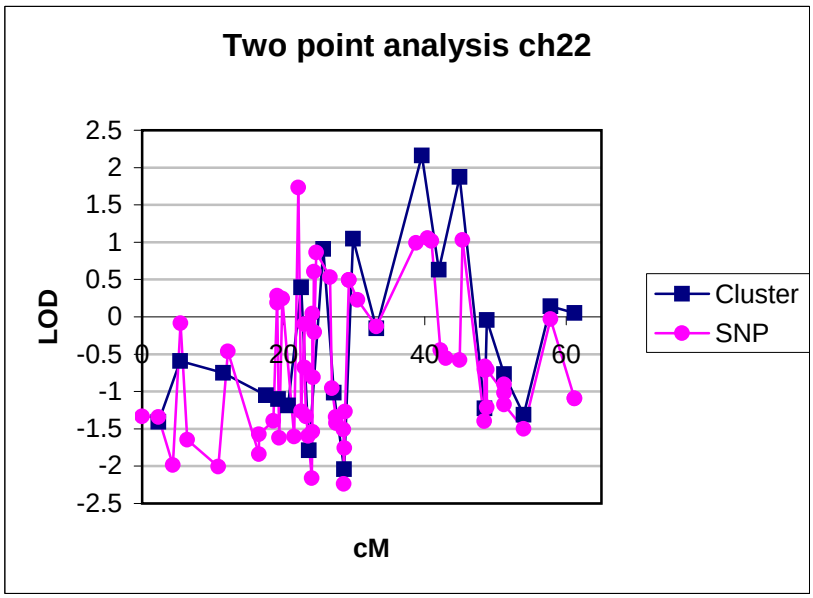

Figure 6.7: Comparing two-point analysis on the original SNP data and the produced clusters. As can be seen from the graph, clusters produce much more consistent results then SNP does

SNP analysis number of regions obtained was 226, eventually reduced to 5 by manual labor. Subsequently fine mapping is required in 5 regions. The multi-point results are also much clearer on the cluster produced whether directly on SNPs. Thus, by adding SNPdistiller to the regular genowide linkage analysis we reduced and partly automated the search for design loci and were able to find the defined locus without fine mapping.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

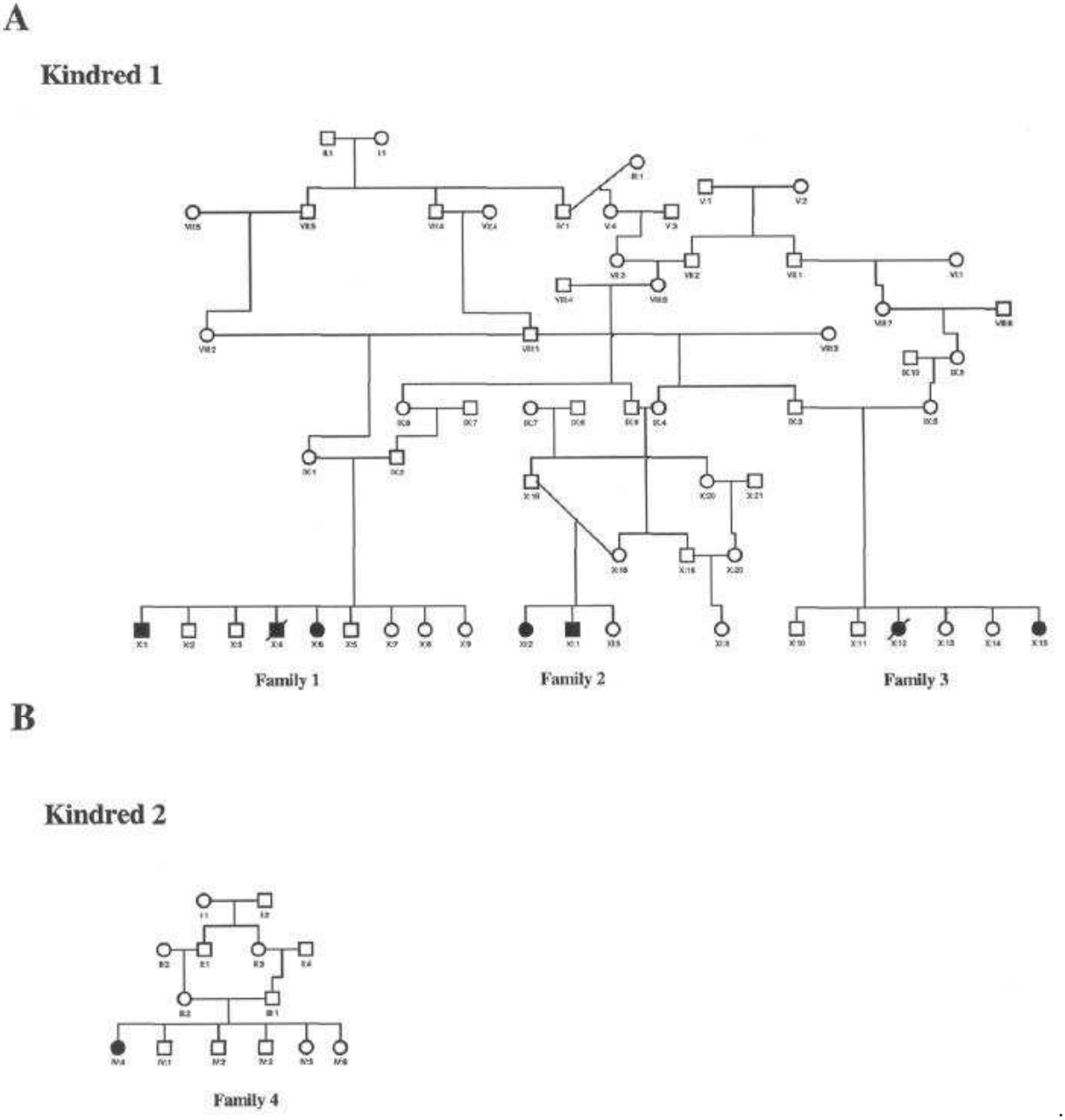

Figure 6.8: Pedigrees of the investigated families. A, Kindred 1, in which 5 affected individuals and 15 unaffected individuals were available for genetic analysis. B, Kindred 2, in which one affected individual and two parents were available for genetic analysis. A high rate of consanguinity and an autosomal recessive pattern of inheritance can be observed.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Research data

After verifying that we can locate disease loci using SNPdistiller, we performed genetic linkage analysis using SNPdistiller on real dataset with unknown disease loci. We used a dataset that is currently under research at one of the Israeli hospitals.

Pedigree and genetic data description

The affected pedigree is an inbred pedigree (see Figure 6.9) with two affected individuals having the same phenotype. The inheritance mode of the disease was assumed autosomal recessive, based on the pedigree. At first, no affected individuals were available for linkage studies. A genomewide linkage analysis, with the use of 400 polymorphic markers, was conducted on the siblings and parents of the affected individuals. No unequivocal results were received. After performing a multipoint analysis, six different loci in six different chromosomes scored LODs between 1.05 and 1.18. Eventually, the blood of one of the affected individuals was available for study. The sample was genotyped and we performed multipoint analysis again. This time only four different loci scored over 1, but under 1.53, producing a very low scores for multipoint analysis. At this point was decided to perform a genomewide linkage analysis of the same patients, using the Affymetrix 250K SNP array. In addition, the aunt and uncle of the affected individuals were also typed. The child of this couple have similar but not exactly the same phenotype as the affected individuals, but was not available for study.

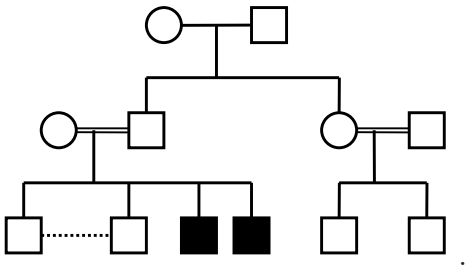

Figure 6.9: Structure of the investigated family. All individuals in the second generation are related.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Experiment description

We looked for affected loci that probably came from a common ancestor. We did not want to miss the loci during the analysis, so we performed the analysis using regular SNPs data in addition to the clusters produced by SNPdistiller. First we performed a two-point analysis. After receiving the results of two-point analysis we performed a multipoint analysis on regions that produced relatively high and consistent LOD scores.

Experiment results

No cluster scored more then 3 in the two-point analysis. Both SNPs and the clusters produced relatively high LOD scores on twelve chromosomes, but the area high LOD score was continuous only on one of the chromosome (chromosome A), without the LOD score dropping. On another chromosome (chromosome B) the LOD score was high, but some clusters between the high LOD clusters, scored -3. To ensure that we did not miss any interesting area, we performed multipoint analysis on both chromosomes. Multipoint analysis on chromosome A produced an LOD score of 1.1858. Multipoint analysis on chromosome B scored 3.0761 for cluster #1181, which includes two SNPs. All high LOD scores were located in a region of about 7 cM in length.

Discussion of results

After obtaining a high LOD score for chromosome A we looked at the raw data. We found that the affected person was almost fully homozygous in this region, with 507 out of 524 typed SNPs being homozygous. 17 SNPs in this region were untyped. The entire region was bounded by heterozygote SNPs both from below and above. If we compare the SNPdistiller results with those of the regular SNP study we notice that two-point analysis in this area for chromosome A also produced a positive LOD score although a lower one. The highest LOD score produced by multipoint analysis in this area was 2.5, but if we would have added several more SNPs to the analysis it would probably have reached a score of 3.

The advantage of SNPdistiller in this case was to reduce output size and simplify the results for better understanding. Instead having to examine 250K points, we needed to look only at about 20K points.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Chapter 7

Future Research and Conclusions

7.1 Conclusions

This thesis was motivated by the increasing need to prepare SNP data for multipoint likelihood computations on general pedigrees with a large number of SNPs, supporting most disease models that are supported by linkage analysis programs. In this thesis, we reviewed the main concepts behind processing SNP data for linkage analysis.

The main contributions of the thesis are the algorithms that automate the process of preparing SNP data for linkage analysis, making large size data of 250K or more suitable for the existing linkage analysis programs. In addition, the SNPdistiller software package was created as part of the research. The software automates the entire process, from error detection to the creation of the input suitable for the linkage analysis packages. Most exiting packages require many additional inputs from users and do not support all stages. The crucial contribution of our algorithms stems from its interaction with the linkage software and the automatic adjustment to the complexity of the data, enabling the usage of the maximum available data and, at the same time, staying within possible for computation problems bounds, rather than using user supplied parameters, as it is done in (Abecasis *et al.* (2002))

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

In this thesis we presented algorithms for cleaning the SNP data by removing both Mendelian errors and statistically unlike genotypes. The main component is the algorithm for clustering SNPs into polymorphic markers for general pedigrees, ignoring the disease model, so changes in the disease assumption or the development of new linkage analysis algorithms do not reduce the quality of the data. We have implemented these algorithms in a new computer program called SNPdistiller.

Another contribution of this work is the location of suspicious regions in case of disease, which was not possible using existing software.

Some research problems remain open. In the current version of SNPdistiller, the clustering algorithm uses the Superlink Haplotype algorithm. Some gains in speed reduction in complexity are achieved by setting recombination values to zero, but currently the performance of the algorithm is limited by the performance of the haplotype algorithm. Improving the haplotyping algorithm will automatically improve the clustering algorithm as well. Another option was to use approximate algorithms, for example Simwalk's (Sobel and Lange (1996)), but we still needed to extract all erroneous genotypes.

7.2 Future Directions

This thesis formulated a complete flow of SNP data processing for linkage analysis. It demonstrated the benefits of algorithms that self-adjust to the complexity of the data. This work has laid the foundation for a fully automatic algorithm for SNP data processing and creates several opportunities for future work in the SNP data field.

Input data format

The current version of SNPdistiller receives files in linkage format as input (see Appendix). An easier solution for the geneticist would be to import directly Affymetrix output files, thereby decreasing the possibility of inserting genotype errors in the translation from one format to another.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

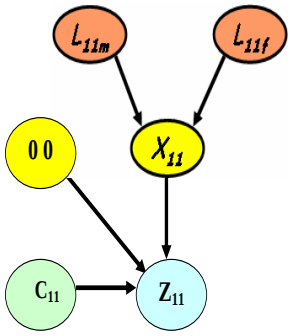

Figure 7.1: Modified Bayes network for a single person. L_{ijm}s the maternal allele at locus i of person j. The values of this variables are the possible alleles l_i at locus i. L_{ijf} is the paternal allele at locus i of person j. The values of this variables are the possible alleles l_i at locus i. X_{ij} is an unordered allele pair at locus i of person j. The values are pairs of ith-locus alleles (l_i,l_i). C_{ij} is the confidence selector. The values are derived from Affymetrix SNP data. Z_{ij} is a variable that receives its value either from X_{ij} or 0̈ 0̈depending on the value C_{ij} value

Error detection

For each genotype Affymetrix provides the confidence in the sample read. This information can be used to find genotypes errors. Currently, the order of the people in the pedigree and the physical SNP order affect which SNPs are excluded from future analysis. To prevent the retaining of erroneous genotypes and the removal of correct ones, we remove an SNP from the analysis if the number of errors exceeds some threshold. This could be changed by ordering the genotype analysis based on measurements with the lowest to the highest confidence. Another option is to integrate the confidence value directly into the Bayes network during likelihood computation by adding one selector for each existing phenotype, which would choose between the given phenotype and an unknown option. Defining 3 new variables for each typed phenotype: C_{ij} confidence selector. The values are derived from Affymetrix SNP data. Dummy variable with constant values "0 0". And Z_{ij} is variable that obtains its value either from X_{ij} or "0 0" depending on the value of C_{ij}. The transformed network per person is given in Figure 7.1.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Clustering

The current clustering algorithm is limited by the complexity of the data. When the SNP data is very dense and the pedigree data relatively large, some of the clusters can be broken into several smaller clusters. This would not be necessary if we had infinite computation power, but the haplotype algorithm is very complex and can process only a moderate number of markers. The possible solution could lie with transitive haplotyping. If markers A, B, and C do not show recombination between them and neither do markers B, C, and D, then A, B, C, and D belong to the same clusters. The disadvantage of this solution is the exponentional increase in run time. The advantage is that the number of clusters produced depends only on the real number of clusters and not on the number of SNPs in the data. This results in two-point analysis results that are more readable.

Another optional solution to this problem is based on the clustering algorithm derived from the unsupervised learning problem. The main problem is how to define good "distance" metrics. We propose r^2s tatistics between two SNPs, but this requires a great deal of future research and adjustment of the data. The main advantage of such a solution, in addtion to previously mentioned advantages is enabling clustering with a small number of recombinations inside the cluster, which is closer to real case situations.

Parallelization

Most stages of the algorithm can be easily parallelized. At every given point one or several SNPs are processed out of thousands of SNPs. Thus the task can be parallelized, significantly reducing the running time. Parallelization may be accomplished as follows:
Possible parallelization

- Detect mendelian errors: each SNP is independent of other SNPs
- Finding erroneous genotypes: each SNPs depends maximum on K neighbors. K can be found based on pedigree size at the beginning of the SNPdistiller run.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

- Clustering. Same as before, but the task is somewhat complex because of the need to find the clusters exactly. A different heuristic can be used for dividing SNPs for parallelizing the process.

The Computational Biology Laboratory at the Technion is currently working on parallezation task.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Appendix A

SNPdistiller manual

The SNPdistiller program is used for preparing SNP data for linakge analysis The program allows to clean and preprocess data with an arbitrary number of loci, either sex-linked or autosomal. The pedigrees can be of arbitrary size and can contain inbreeding loops. In addition to SNPs, affection status and phenotypes can be given as input, but will be ignored. The program does not assume that input data is SNP and can be run on any marker data.

A.1 Input format

SNPdistiller receives two files as input. Parameter file is a standard LINKAGE/Superlink data file in with any program code format. Pedigree file is a standard LINKAGE data file BEFORE (Genehunter format) or AFTER (Superlink format) running MAKEPED of the LINKAGE-package. The flag specifying whether it Superlink or Genehuneter format should be provided as first parameter

1 - Genehunter format

2 - Superlink format

Example for Genehunter format command line: SNPdistiller 1 datafile.dat pedfile.dat

Example for Superlink format command line: SNPdistiller 2 datafile.dat pedfile.dat We would here present the Superlink format. The input is divided

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

into two parts, each one described in a different file. One input file contains the pedigree and genotypic data. The second input file contains a description of the loci being analyzed, the recombination rates and the loci order. Section A.1.1 describes the types of genotypes and phenotypes that can be interpreted by superlink. Sections A.1.2and A.1.3 contain a detailed description of the pedigree file and locus file, respectively. A large part of the description in this section is based on Superlink user guides and can be found at http://bioinfo.cs.technion.ac.il/superlink

Possible Genotypes and Phenotypes

SNPdistiller can interpret two possible phenotype types: 1. Affection status(optional data). A numeric code which indicates the presence or absence of a disease (or other qualitative phenotype). Sometimes, a numeric code which represents a liability (or risk) class is also included as part of the phenotype. 2. Numbered alleles. This phenotype is used to represent codominant alleles at a single locus. It is used to describe markers. The phenotype is made up of two allele numbers which correspond to an unordered genotype.
These phenotypes correspond to three types of loci that can be analyzed by proceeded to SNPdistiller:

1. Affection status. This type of locus has 2 alleles, a normal allele and a disease allele. It corresponds to an affection status phenotype. This data is ignored by SNPdistiller and exists only for user convenience.

2. Complex Affection. Also ignored. For more detailes on the format see Superlink manual

3. Numbered alleles. This type of locus can have any number of alleles. As stated above, the alleles are codominant. The corresponding phenotype for this locus type is the numbered alleles phenotype. Although SNPs assumed to have only two alleles, SNPdistiller can work with any data.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Pedigree File Format

This file describes the input pedigree(s). It contains both phenotypic information, and pedigree information required in order to traverse the pedigree when calculating the likelihood. Each line in the file describes an individual in one of the pedigrees. The structure of each line is described in Table A.1

Column	Description
1	Pedigree number
2	Individual number (id)
3	Number (id) of father
4	Number (id) of mother
5	Number (id) of first child
6	Number (id) of next sibling from same father
7	Number (id) of next sibling from same mother
8	Sex (1 encodes Male, 2 encodes female)
9	Ignored (you can simply enter 0). Left because of compatibility to other progrmas
10 $\Rightarrow$ last column	Phenotypic data for the loci being analyzed. Phenotypic data for the loci being analyzed. The phenotypic information appears according to the input order

Table A.1: General description of the structure of a line in the pedigree file.

The description of each single phenotype is as follows:

1. For an affection status phenotype:
   - In the first column appears the disease status, where:
     0 = unknown
     1 = unaffected
     2 = affected
   - In the second column appears the liability class that this person belongs to, if there is more than 1 liability class. If there is only one liability class, the phenotype is composed of only one column, which includes the disease status.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

2. For a numbered alleles phenotype, there are 2 columns which hold the alleles of the person at this marker. An unknown phenotype is coded as 0 0. For a sex-linked locus, males have a single allele. With the allele 1, for example, the phenotype can be coded as either 1 0 or 1 1.

parents must be present in the pedigree, even if one is unknown. If one parent is unknown, an id number must still be created, and a record for the parent must appear in the pedigree file.

The first offspring number along with the next paternal sibling and next maternal sibling numbers create lists of pointers that make it possible to pass from one child to the next. These numbers need to be set in such a way that all the children are included in the list of pointers. The entry for the next paternal sibling of the last child is 0 and the same for the next maternal sibling entry. The first offspring can be chosen arbitrarily from ones children.

Locus File Format

This file describes the loci and the different parameters necessary for the analyzing programs. It is divided into four major parts:

1. General information on the loci and the loci order.
2. Description of the loci.
3. Recombination information.
4. Program-specific information.

When explaining the structure of the locus file, two concepts of locus order will be used. The first is input order, or the order in which the phenotypes corresponding to the loci appear in the pedigree file (section A.1.2). The second concept is chromosome order, or the physical order assumed for the loci on the chromosome. The input order is fixed for a given set of input files, but the chromosome order can be changed to test various hypotheses. Various parameters, such as recombination rates, gene frequencies, penetrances, etc., are specified in the locus file. The specified values refer to the initial values

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

of these parameters. The analysis programs can modify some of these values for specific purposes, e.g. maximum likelihood estimation. Following is a more detailed description of each of these parts. This part consists of three lines.

- The first line of the file contains information on the following parameters:
  - No Loci (no. of loci in analysis) - can be any number greater than 1.
  - Risk-Locus - 0 if risk is not to be calculated. Ignored
  - Sex-Linked - a numeric code that indicates if the data is sex-linked (code 1) or autosomal (code 0).
  - Program-Code. If this parameter is not specified, then it is considered to be 0.Ignored

  The format of this line is: *No-Loci Risk-Locus Sex-Linked Program-Code No-Complex-Affection-Loci*
- The second line contains information on the following parameters (all of which are ignored by SNPdistiller):
  - Mutation-Locus - a mutation locus
  - Mutation-Male and Mutation-Female - male and female mutation rates
  - Hap-Freq - Haplotype frequencies (if 1).

  The format is: *Mutation-Locus Mutation-Male Mutation-Female Hap-Freq*
- The third line specifies the chromosome order of the loci (i.e., the physical order assumed for the loci).

Loci Description

The loci are described in the order in which they appear in the pedigree file (i.e., input order and not chromosome order). The description varies according

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

to the locus type. The following types of loci are possible (each one is assigned a different numeric code):

- 1 = affection status
- 3 = numbered alleles
- 4 = complex affection

Type 0 (Quantitative variable) and type 2 (Binary factors), which are possible in the linkage fastlink programs, are not implemented in superlink. Type 4 does not exist in th linkage/fastlink programs, however, it exists in the tlinkage programs. The format for each locus type is as follows:

- **Numbered Alleles (coded 3):** locus type (namely, 3) and the number of possible alleles for this locus. The second line consists of the gene frequencies.
- **Affection Status (coded 1) and Complex Affection (coded 4):** are not relevant for the SNPdistiller.

Recombination Information

The next two-three lines of the locus file provide recombination information.

1. **sex difference, interference**.

   The format is: Sex-difference Interference

   The following sex-difference options are possible:

   - 0 → no sex-difference.
   - 1 → constant sex-difference (the ratio of female/male genetic distance is the same in all intervals).
   - 2 → variable sex-difference (the female/male genetic distance ratio can be different in each interval).

   - 0 → no interference.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

- 1 → interference without a mapping function.
- 2 → Kosambi map function

2. **recombination values** - between each pair of consecutive loci according to the assumed chromosome order.

Program-Specific Information

Ignored by SNPdistiller, but output files will preserve the program code

A.2 Output format and options

By default, SNPdistiller will create cluster from given input files. The output of the SNPdistiller will include next files

1. Pedfile and datafile in Superlink format. The files will include markers which represent the cluster which were created. If affected loci was given as input, it would be also at output.

2. err.txt. This file include data on all genotypes and markers which were droped. It also include information why some genotype or marker was droped. Whether it was non informative, Mendelian error or unlikely genotype

If one does not want to create cluster, but want to get SNPdistiller results at some early point, it can be done by specifying addition input parameter

1. → SNPdistiller will exit after checking Mendelian errors

2. → SNPdistiller will exit after removing all non informative markers

3. → SNPdistiller will exit after removing all markers which have 0 likelihood

4. → SNPdistiller will exit after removing unlikely genotypes

5. → SNPdistiller will exit after clustering(this is default)

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

A.3 Useful Code Constants

Fine tuning of the algorithm can be done by changing some constants inside the code and recompiling the program. All related constants are defined in *.h files

- CLUSTER_FREQ - defines whether all alleles in cluster will have equal frequencies (2) or based on original alleles frequencies(1)
- LOW_FREQ_LIMIT and HIGH_FREQ_LIMIT - defines extreme allele frequencies. Bellow and above this numbers marker will be dropped.
- MIN_NUMBER_ALLELES_PER_CLUSTER - clusters with less alleles then given number will be dropped. Currently setted to 2, meaning all clusters are ok.
- MAX_SM_PER_CLUSTER - defines maximum cluster size in cM. Defined by distance between first SNP in cluster and last one.

In addition code written so, different parts of it can be easily replaced by other function. It is convenient in case we would decide to replace haplotype or other algorithm

Some of the data was tooked from The Robert S. Boas Center for Genomics and Human Genetics and BRICS RS-02-7 Ingolfsdottir et al.: A Formalization of Linkage Analysis

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Bibliography

Abecasis, G. R. and Wigginton, J. E. (2005). Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am J Hum Genet*, **7**7(5), 754–767.

Abecasis, G. R., Cherny, S. S., and Cardon, L. R. (2001). The impact of genotyping error on family-based analysis of quantitative traits. *Eur J Hum Genet*, **9**(2), 130–134.

Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2002). Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, **3**0(1), 97–101.

Carlson, C. S., Eberle, M. A., Rieder, M. J., Smith, J. D., Kruglyak, L., and Nickerson, D. A. (2003). Additional snps and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat Genet*, **3**3(4), 518–521.

Clerget-Darpoux, F., Bonaïti-Pellié, C., and Hochez, J. (1986). Effects of misspecifying genetic parameters in lod score analysis. *Biometrics*, **4**2(2), 393–399.

De Wilde, S., Carey, I. M., Bremner, S. A., Richards, N., Hilton, S. R., Strachan, D. P., and Cook, D. G. (2004). A comparison of the recording of 30 common childhood conditions in the doctor's independent network and general practice research databases. *Health Stat Q*, **.**(22), 21–31.

Elston, R. C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Hum Hered*, **2**1(6), 523–542.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Evans, D. M. and Cardon, L. R. (2004a). Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. *Am J Hum Genet*, **7**5(4), 687–692.

Evans, D. M. and Cardon, L. R. (2004b). Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. *Am J Hum Genet*, **7**5(4), 687–692.

Ewen, K. R., Bahlo, M., Treloar, S. A., Levinson, D. F., Mowry, B., Barlow, J. W., and Foote, S. J. (2000). Identification and analysis of error types in highthroughput genotyping. *American Journal of Human Genetics,*, **6**7, 727–736.

Fishelson, M. and Geiger, D. (2002). Exact genetic linkage computations for general pedigrees. *Bioinformatics*, **1**8 Suppl 1, 189–198.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science*, **2**96(5576), 2225–2229.

Gordon, D., Heath, S. C., and Ott, J. (1999). True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. *Hum Hered*, **4**9(2), 65–70.

Khateeb, S., Flusser, H., Ofir, R., Shelef, I., Narkis, G., Vardi, G., Shorer, Z., Levy, R., Galil, A., Elbedour, K., and Birk, O. S. (2006). Pla2g6 mutation underlies infantile neuroaxonal dystrophy. *Am J Hum Genet*, **7**9(5), 942–948.

Kruglyak, L., Daly, M. J., and Lander, E. S. (1995). Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *Am J Hum Genet*, **5**6(2), 519–527.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet*, **5**8(6), 1347–1363.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Lander, E. S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A*, **8**4(8), 2363–2367.

Mah, J. T. and Chia, K. S. (2007). A gentle introduction to snp analysis: resources and tools. *J Bioinform Comput Biol*, **5**(5), 1123–1138.

O'Connell, J. R. and Weeks, D. E. (1999). An optimal algorithm for automatic genotype elimination. *Am J Hum Genet*, **6**5(6), 1733–1740.

Ott, J. (1977). Linkage analysis with misclassification at one locus. *Clin Genet*, **1**2(2), 119–124.

Ott, J. (1992). Strategies for characterizing highly polymorphic markers in human gene mapping. *Am J Hum Genet*, **5**1(2), 283–290.

Ott, J. (1999). Analysis of human genetic linkage,third edition. *The Johns Hopkins University Press*.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P., and Cox, D. R. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **2**94(5547), 1719–1723.

Seaman, S. and Holmans, P. (2005). Effect of genotyping error on type-i error rate of affected sib pair studies with genotyped parents. *Hum Hered*, **5**9, 157–164.

Sieh, W., Basu, S., Fu, A. Q., Rothstein, J. H., Scheet, P. A., Stewart, W. C., Sung, Y. J., Thompson, E. A., and Wijsman, E. M. (2005). Comparison of marker types and map assumptions using markov chain monte carlo-based linkage analysis of coga data. *BMC Genet*, **6** Suppl 1.

Silberstein, M., Tzemach, A., Dovgolevsky, N., Fishelson, M., Schuster, A., and Geiger, D. (2006). Online system for faster multipoint linkage analysis via

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

parallel execution on thousands of personal computers. *Am J Hum Genet*, **7**8(6), 922–935.

Sobel, E. and Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet*, **5**8(6), 1323–1337.

Takeuchi, F., Yanai, K., Morii, T., Ishinaga, Y., Taniguchi-Yanai, K., Nagano, S., and Kato, N. (2005). Linkage disequilibrium grouping of single nucleotide polymorphisms (snps) reflecting haplotype phylogeny for efficient selection of tag snps. *Genetics*, **1**70(1), 291–304.

Thompson, E. A. (1995). Monte carlo in genetic analysis. *Technical report*, **D**epartment of Statistics, University of Washington(294).

Thompson, E. A. (2000). Statistical inferences from genetic data on pedigrees. *NSF-CBMS Regional Conference Series in Probability and Statistics*, **6**.

Thompson, E. A. and George, W. A. (2003). Multipoint linkage analyses for disease mapping in extended pedigrees: A markov chain monte carlo approach. *Statistical Science*, **1**8, 515–531.

Yang, X. R., Jacobs, K., Kerstann, K. F., Bergen, A. W., Goldstein, A. M., and Goldin, L. R. (2005). Linkage analysis of the gaw14 simulated dataset with microsatellite and single-nucleotide polymorphism markers in large pedigrees. *BMC Genet*, **6** Suppl 1.

Zhang, K., Deng, M., Chen, T., Waterman, M. S., and Sun, F. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci U S A*, **9**9(11), 7335–7339.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

התורשה והנתונים הגנטיים, ועקב כך ניתן להשתמש באלגוריתם על מגוון רחב של עצי תורשה.
לצורך בדיקת הכלי ביצענו שני סוגים של ניסויים. בניסוי ראשון הגרלנו נתונים שמטרתם לבדוק ביצועי אלגוריתמים במצבים שונים – כגון גודל המשפחה, צפיפות הSNPים וכו'. בניסוי השני, השתמשנו בנתונים אמיתיים שהתקבלו ממחקרים גנטיים. בניסוי הראשון לקחנו עצי תורשה מגדלים שונים, והגרלנו SNPים בעלי תכונות מסוימות. השתמשנו בעצים בגדלים המייצגים בעיות שונות באנליזת תאחיזה – החל מעץ תורשה קטן המאפשר ריצה של תוכנות תאחיזה על כל הSNPים בו זמנית ועד לעצי תורשה גדולים שבהם רק בדיקות תאחיזה דו נקודתיות אפשריות. בנוסף דימינו מצב בו חלק מSNPים נמצאים בחוסר שווי משקל תאחיזתי (linkage disequilibrium). עבור כל עץ הגרלנו שני קבוצות נתונים. בקבוצה הראשונה התכונה נמצאת בתאחיזה למפה, ובקבוצה השנייה התכונה אינה נמצאת בתאחיזה. במהלך הניסוי הרצנו בדיקות תאחיזה פעמיים, פעם אחת השתמשנו בSNPים ופעם אחת במקבצים. ובסוף עשינו השוואה של תוצאות. בכל הניסויים המקבצים מיפו את התכונה בהצלחה. בחלק מהניסויים התוצאה על מקבצים הייתה יותר ברורה מהתוצאה באמצעות SNPים. בנוסף ראינו שלא נוצרות תוצאות חיוביות כוזבות חדשות וגם שנעלמו חלק מהתוצאות החיוביות כוזבות שהתקבלו בשיטות ללא מקבצים. בשלב שני, ניסינו את האלגוריתם על שתי קבוצות של נתונים אמיתיים. בקבוצה הראשונה מיקום התכונה היה ידוע. בניסוי זה מופו 8 אנשים ע"י עשרת אלפים SNPים. במקור, גנטיקאים נזדקקו למיפוי עדין יותר לאחר מיפוי ע"י SNPים, כי הSNPים לא נתנו תוצאה חד משמעית, אלא רק 5 מיקומים מועמדים. לעומת זאת, לאחר יצירת מקבצים ע"י SNPdistiller, בבדיקות תאחיזה דו נקודתיות התקבלו רק שני מקומות מועדים. בדקנו תאחיזה רב נקודתית במקומות אלה, ורק המיקום המדויק של התכונה קיבל ניקוד גבוה. קבוצת הנתונים השנייה נמצאת עדיין במחקר גנטי ולכן מיקום התכונה אינו ידוע. לרשותנו היה מיפוי של 15 בני משפחה בעזרת 250 אלף SNPים ו400 סמנים גנטיים פולימורפיים. גנטיקאיים ביצעו בדיקות תאחיזה דו נקודתיים על כל הSNPים ובדיקות תאחיזה דו ורב נקודתיות על סמנים גנטיים פולימורפיים. התוצאה הייתה מאוד לא חד משמעית, כאשר הרבה מקומות קיבלו ציון גבוה יחסית. כלומר, מועמדים להיות בתאחיזה לתכונה. בדיקות תאחיזה עם שימוש במקבצים שנוצרו ע"י SNPdistiller הראו רק מיקום אחד משוער של התכונה.
בכל הניסויים ראינו, שהאלגוריתם מאפשר ניתוח יותר פשוט ומדויק של SNPים ע"י סינון הנתונים משגיאות ואיסוף שלהם לתוך קבוצות יותר אינפורמטיביות, כאשר כל השלבים מתבצעים אוטומטית, ללא צורך בהתערבות מצד המשתמש. המגבלה העיקרית של האלגוריתם היא חישוב של הפלוטיפ מקסימאלי, אבל ההגבלה הזאת עדיין מאפשרת לטפל בכמות SNPים גדולה. חלק מהשלבים של האלגוריתם ניתנים לשיפורים עתידיים ע"י אינטראקציה עם מסדי נתונים קיימים ושיפור או שינוי האלגוריתם המאחד SNPים לתוך קבוצות. עוד שיפור המתבקש לאלגוריתם הוא המיקבול שלו. בסך הכול האלגוריתם מציג שיטה לאוטומציה של עיבוד SNPים לצורך בדיקת תאחיזה כאשר האלגוריתם מתעלם מפנוטיפ המחלה או התכונה המבוקשת ונתונים גנטיים שלה ולכן אין חשש לאפיון לא נכון של המקבצים, עקב אפיון לא נכון של המחלה.

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

תקציר

ביואינפורמטיקה הינה שימוש של טכניקות וכלים חישוביים לחקר ביולוגיה מולקולארית, גנטיקה או לניתוח נתונים קליניים. תחום הביואינפורמטיקה גדל מאוד בשנים אחרונות בעקבות מיזם הגנום וכתוצאה משימוש במחשבים חזקים יותר. כעת ניתן להשתמש בחומרה ותוכנה מתקדמות ליצירת מתודולוגיות חדשניות לשימוש בנתונים גנטיים הנמצאים במסדי נתונים שונים, שנוצרו במחקרים בינלאומיים ועל ידי כך, להפיק נתונים רלוונטיים למחקר ספציפי מכלל המידע הנאסף. המטרה האולטימטיבית היא ליצור תוכנות מחשב אשר יכולות לספק במהירות מידע על הגנים מתאימים, שינויים בתוך הגנים האלה וניבוי של מבנה ופונקציונאליות של החלבונים המקודדים, בכדי להפיק מידע על מחלות מורכבות. התזה מתמקדת בעיבוד והכנה של מידע על וריאנטים בנוקלאוטיד בודד עבור כלים לניתוח תאחיזה גנטית.

אנליזה של תאחיזה גנטית שימושית למיפוי מחלות ותכונות גנטיות, מאפשרת שימוש בכלים סטטיסטים לשיוך תפקודיות של גנים למיקומם בכרומוזום. אנליזה זו משתמשת במודל הסתברותי של הורשה ובנתונים המגיעים מעצי תורשה לצורך מציאת גנים האחראים לתכונה הנחקרת. לצורך האנליזה פותחו מפות של סמנים גנטיים. בדיקות של תאחיזה רב נקודתית הפכו להיות כלי מכריע באנליזת תאחיזה, בגלל העליונות שלהם על בדיקות תאחיזה דו נקודתיות בחיפוש ומציאת גנים וגילוי תאחיזה בין תכונות לגנים. אבל בדיקות רב נקודתיות מאוד מורכבות מבחינה חישובית והסיבוכיות של החישובים גדלה אקספונציונלית עם כמות הסמנים, גודל המשפחה, מספר האנשים הלא מאופיינים במשפחה ופולימורפיות של סמנים גנטיים. כל הגורמים אלה מגבילים את התוכנות הקיימות בשימוש שלהם במשאבי מחשב. חלק מהתוכנות יכולות להתמודד עם משפחות גדולות, אבל עם מספר קטן של סמנים. תוכנות אחרות יכולות להתמודד עם מספר גדול של סמנים גנטיים, אבל רק במשפחות קטנות.

SNPים (Single Nucleotide Polymorphism) הם וריאציה בנוקלאוטיד בודד בתוך הכרומוזום. אומדנים עכשוויים מעריכים שSNPים מופיעים בתדירות של 100 - 300 בסיסים. מהערכה זו נובע שבגנום אנושי קיימים, בקירוב, 10 - 30 מיליון SNPים פוטנציאלים, כאשר מעל 4 מיליון SNPים כבר זוהו ומופו. בגלל המספר העצום של SNPים אי אפשר להשתמש בכולם בו זמנית באנליזת תאחיזה. יחד עם זה, מספר קטן של SNPים נושא רק מידע מועט, עקב הדמיון הרב בין מרקרים צמודים. בנוסף, השגיאות המתרחשות בזמן הגנוטיפיזציה גורמות לSNPים להיות עוד פחות מועילים לצורך אנליזה. לכן, יש להשתמש ביתרון של SNPים, שהוא הכמות הגדולה שלהם, אבל לגרום לנתונים המופקים להיות יותר אינפורמטיביים ופחות רגישים לשגיאות.

בתזה הנוכחית אנו מציגים אלגוריתם המנסה להשיג מטרה זו והכלי הממש אותו, SNPdistiller. האלגוריתם מטפל בכל התהליך של הכנת SNPים לאנליזת תאחיזה, החל מניקוי נתונים משגיאות אחרי גנוטיפיזציה ועד ליצירת קלט מתאים לכלי בדיקה קיימים לתאחיזה גנטית. בשלב ראשון, SNPdistiller מסיר גנוטיפים לא סבירים ושגויים. לכל SNP, SNPdistiller בודק שימור חוקי ההורשה ללא התחשבות בSNPים השכנים. אם גנוטיפ כלשהוא סותר את חוקי ההורשה,ה SNP מוצא מהאנליזה. הכלי ממשיך בחיפוש אחר גנוטיפים לא סבירים. לכל גנוטיפ בעץ תורשה הכלי משווה את סבירות הנתונים עם ובלי הגנוטיפ הספציפי, כאשר נלקח בחשבון מספר גדול ככל האפשר, של הSNPים השכנים. אם סבירות הנתונים עם הגנוטיפ נמוכה מאשר בלי הגנוטיפ הנ"ל, כנראה שיש שגיאה באפיון של הגנוטיפ. לפיכך, גנוטיפים שהוגדרו "לא סבירים" לא נכנסים לאנליזה. הכלי ממשיך בארגון SNPים לתוך מקבצים המחקים סמנים פולימורפיים ואינפורמטיביים. שני SNPים יהיו באותו מקבץ, אם הם נמצאים במרחק סביר אחד מהשני ובעץ התורשה הנחקר אין בינם רקומבינציה. ע"י כך, בכל מקבץ יכול להיות בין אחד ועד מספר כלשהוא SNPים. בפועל כמות הSNPים בתוך מקבץ מוגבלת ע"י הכוח החישובי של האלגוריתם המאגד אותם. SNPdistiller משתמש באלגוריתם למציאת הפלוטיפים, כאשר לאלגוריתם הוכנסו שינויים קטנים. הפלט של הכלי הוא שלושה קבצים. שני הקבצים הראשונים מהווים קלט לתוכנות תאחיזה ומכילים נתונים על המקבצים ועץ תורשה. קובץ שלישי מכיל נתונים על ריצת SNPdistiller - באיזה SNPים הכלי השתמש לצורך יצירת מקבצים, באיזה לא השתמש ומדוע. בכל השלבים הכלי עושה התאמה אוטומטית למורכבות עץ

המחקר נעשה בהנחיית פרופ' דן גייגר בפקולטה למדעי המחשב
אני מודה לטכניון על התמיכה הכספית הנדיבה בהשתלמותי

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

הכנת מידע על וריאנטים בנוקלאוטיד בודד לבדיקות תאחיזה גנטית

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
מגיסטר למדעים במדעי מחשב

אנה צמח

הוגש לסנט הטכניון - מכון טכנולוגי לישראל
שבט תשס"ט חיפה ינואר 2009

Technion - Computer Science Department - M.Sc. Thesis MSC-2009-02 - 2009

הכנת מידע על וריאנטים בנוקלאוטיד בודד לבדיקות תאחיזה גנטית

אנה צמח