Available online at www.sciencedirect.com
ScienceDirect
International Journal of Approximate Reasoning
48 (2008) 499–525
INTERNATIONAL JOURNAL OF APPROXIMATE REASONING
www.elsevier.com/locate/ijar

RC_Link: Genetic linkage analysis using Bayesian networks ☆

David Allen *, Adnan Darwiche

Computer Science Department, University of California, Los Angeles, CA 90095, United States

Received 22 November 2006; received in revised form 14 June 2007; accepted 3 October 2007
Available online 10 October 2007

Abstract

Genetic linkage analysis is a statistical method for mapping genes onto chromosomes, and is useful for detecting and predicting diseases. One of its current limitations is the computational complexity of the problems of interest. This research presents methods for mapping genetic linkage problems as Bayesian networks and then addresses novel techniques for making the problems more tractable. The result is a new tool for solving these problems called *RC_Link*, which in many cases is orders of magnitude faster than existing tools.

Keywords: Bayesian networks; Genetic linkage analysis; Pedigree; RC_Link; Probabilistic inference

1. Introduction

Ordering genes on a chromosome and determining the distance between them is useful in predicting and detecting diseases. Detecting where disease genes are located and what other genes are near them on a chromosome can lead to determining which people have a high probability of occurrence, even before symptoms appear, allowing for earlier treatment. Genetic linkage analysis is a statistical method for this mapping of genes onto a chromosome and determining the distance between them [27].

Recently it has been shown that Bayesian networks are well suited for modeling and reasoning about this domain [15,2,22,14]. In this paper, we present a new tool called *RC_Link* which can model genetic linkage analysis problems as Bayesian networks and do inference efficiently, in many cases orders of magnitude faster than existing tools.

This paper will first give background on the domain of genetic linkage analysis and Bayesian networks in Section 2 and then in Section 3 will discuss how the linkage problems are encoded as Bayesian networks. Sections 4–6 will then present many of the techniques *RC_Link* uses to make the problems more tractable. These techniques can also be used by other existing tools to further improve their performance. Finally, Section 7 will present experimental results comparing *RC_Link* with existing tools and then offer some concluding remarks in Section 8.

☆ Available at http://reasoning.cs.ucla.edu/rc_link.

* Corresponding author. Tel.: +1 310 206 5201.
E-mail addresses: dlallen@cs.ucla.edu (D. Allen), darwiche@cs.ucla.edu (A. Darwiche).

0888-613X/$ - see front matter
doi:10.1016/j.ijar.2007.10.003

2. Background

Genetic linkage analysis is an important method for mapping genes onto chromosomes and helping to predict disease occurrences prior to the appearance of symptoms. Research over the past few years has shown that Bayesian networks are well suited for doing linkage analysis computations and many tasks which were considered intractable a few years ago are now solvable, allowing the biology, genetics, and bioinformatics researchers to further study their data and draw new conclusions.

2.1. Genetic linkage analysis

Many algorithms used for genetic linkage analysis are extensions of either the Elston–Stewart algorithm [13] or the Lander–Green algorithm [20]. The first algorithm does well with many people and few genes, while the second algorithm works well for fewer people and many genes. Quite a few genetic linkage analysis tools have been produced, most notably FASTLINK [9,31,6], GENEHUNTER [19], VITESSE [26], and SUPERLINK [15]. In order to understand the genetic linkage analysis tasks these tools are solving we will first briefly review some relevant Biology background.

Human cells contain 23 pairs of chromosomes, which are sequences of DNA containing the genetic makeup of an individual and are inherited from a person's parents. Each pair consists of one chromosome inherited from the person's father and one from their mother. Locations on these chromosomes are referred to as *loci* (singular: *locus*). A locus which has a specific function is known as a *gene*. These functions, which can be a result of a combination of multiple genes, can include such things as determining a person's blood type, hair color, or their susceptibility to a disease. The actual state of the genes is called the *genotype* and the observable outcome of the genotype is called the *phenotype*. A *genetic marker* is a locus with a known DNA sequence which can be found in each person in the general population. These markers are used to help locate disease genes. Fig. 1 displays a chromosome, its DNA makeup, and identifies one gene.

Each parent contains their 23 pairs of chromosomes, however they each only pass a total of 23 chromosomes on to their children, one chromosome from each pair, resulting in the child having 23 pairs. It is possible for the transferred copy to be entirely a duplicate of the chromosome from the parent's father or from the parent's mother (the offspring's grandfather or grandmother), however more likely it contains nonoverlapping sequences from both. The locations on the chromosome where the sequences switch between the two parents are known as crossover or recombination events. The *recombination frequency*, θ, (also called the *recombination fraction*) between two consecutive genes is defined as the probability of a recombination event occurring between them.[1] Therefore, if two genes are unlinked, or uncorrelated, they will have $\theta = 0.5$ (meaning the state of the first will not influence the state of the second), whereas linked genes will have $\theta < 0.5$. This frequency is related to the physical distance between them, for example if two genes are close together there may be little chance for a recombination to occur, however if two genes are far away the probability of recombination increases. In Fig. 2 a chromosome is depicted along with the location of three (ordered) genes. It furthermore depicts the recombination frequencies, θ_1 and θ_2, between the two consecutive pairs.

Therefore, given a large population of people, their inheritance structure (i.e. a family tree, also called a *pedigree*, an example of which is in Fig. 3), and partially known genotype and/or phenotype information (e.g. genetic marker readings and disease affection status), genes can be mapped onto the chromosomes based on how frequently recombination events occur between pairs of genes. More formally, let P represent a population of related individuals, let $\mathbf{e}$ be the known evidence on genotypes and/or phenotypes, and let $\hat{\boldsymbol{\theta}}$ be a vector containing the recombination frequencies between each pair of consecutive genes (Hence if we have n genes, then $\hat{\boldsymbol{\theta}}$ will have $n-1$ θ_i values). We can then compute $Pr(\mathbf{e}|P,\hat{\boldsymbol{\theta}})$, which is the likelihood of the known data for the given population and recombination frequencies.

A common task in genetic linkage analysis is then to compute this likelihood for multiple $\hat{\boldsymbol{\theta}}$ vectors, selecting the one with the maximum likelihood. When doing analysis between different populations, the numerical

[1] This frequency between two loci is sometimes measured in units called *centimorgans*, where 1% recombination is equal to 1 centimorgan.