Chapter 6
Gene Mapping by Pattern Discovery

Petteri Sevon, Hannu T. T. Toivonen, and Päivi Onkamo

Summary

The objective of gene mapping is to localize genes responsible for a particular disease or trait. We consider association-based gene mapping, where the data consist of markers genotyped for a sample of independent case and control individuals. In this chapter we give a generic framework for nonparametric gene mapping based on pattern discovery. We have previously introduced two instances of the framework: haplotype pattern mining (HPM) for case–control haplotype material and QHPM for quantitative trait and covariates. In our experiments, HPM has proven to be very competitive compared to other methods. Geneticists have found the output of HPM useful, and today HPM is routinely used for analyses by several research groups. We review these methods and present a novel instance, HPM-G, suitable for directly analyzing phase-unknown genotype data. Obtaining haplotypes is more costly than obtaining phase-unknown genotypes, and our experiments show that although larger samples are needed with HPM-G, it is still in many cases more cost-effective than analysis with haplotype data.

6.1 Introduction

The first step in discovering genetic mechanisms underlying a disease is to find out which genes, or more precisely, which polymorphisms, are involved. Gene mapping, the topic of this chapter, aims at finding a statistical connection between the trait under study and one or more chromosomal regions likely to be harboring the disease susceptibility (DS) genes. Chromosomal regions that cosegregate with the trait under study are searched for in DNA samples

from patients and controls. Even though the coding parts of the genes—the exons—cover only a small fraction of the human genome, the search cannot be restricted to them: polymorphisms affecting disease risk may reside in the introns or promoter regions quite far from the exons, having an effect on the expression level or splicing of the gene.

All the important simple monogenic diseases have already been mapped, or at least it is well known how it can be done. The general interest is shifting toward complex disorders, such as asthma or schizophrenia, where individual polymorphisms have rather weak effects. There may be epistatic interaction between several genes, and some mechanisms may be triggered by environmental factors. Complex disorders are also challenging clinically: it is of primary importance that the diagnoses are based on identical criteria. Systematic noise caused by inconsistent definitions for symptoms could severely hinder the search for the genetic component of the disorder. The mutation does not always cause the complex disorder (lowered penetrance), or the same disorder may be caused by other factors (phenocopies). There are other stochastic processes involved, such as recombinations and mutations, and genealogies are usually known only a few generations back. For these reasons, only probabilistic inferences can be made about the location of the DS genes.

In this chapter we present haplotype pattern mining (HPM), a method of gene mapping that utilizes data mining techniques. The chapter is organized as follows. First, we review the basic concepts in genetics and gene mapping in section 6.2. Next, we give an abstract generic algorithm for HPM in section 6.3 and present and evaluate three instances of that in section 6.4. Finally, we give a summary of related work in section 6.5 and close with a discussion in section 6.6.

6.2 Gene Mapping

Markers. *Markers* provide information about genetic variation among people. They are polymorphic sites in the genome, for which the variants an individual carries can be identified by laboratory methods. The location of a marker is usually called a *locus* (pl. *loci*). The variants at a marker are called *alleles*. We will use small-integer numbers to denote alleles throughout the chapter. The array of alleles in a single chromosome at a set of markers is called a *haplotype*.

Example 6.2.1. Let M1, M2, M3, and M4 be markers located in this order along chromosome 1. Let the alleles at these marker loci in a given instance of chromosome 1 be 1, 3, 2, and 1, respectively. The haplotype for this chromosome over all the markers is [1 3 2 1], and the haplotype over markers M2 and M4, for instance, is [3 1].