Gene Mapping by Pattern Discovery

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

Springer-Verlag
Berlin Heidelberg NewYork
London Paris Tokyo
HongKong Barcelona
Budapest

DRAFT
Accepted for publication in 'Data Mining in Bioinformatics'
Jason Wang, Mohammed Zaki, Hannu Toivonen, and Dennis Shasha (Eds.), Springer

Chapter 1. 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 consists of markers genotyped for a sample of independent case and control individuals. In this chapter we give a generic framework for non-parametric 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 in 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.

1.1 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

DRAFT
Accepted for publication in 'Data Mining in Bioinformatics'
Jason Wang, Mohammed Zaki, Hannu Toivonen, and Dennis Shasha (Eds.), Springer

to be harbouring the disease susceptibility (DS) genes. Chromosomal regions that co-segregate 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—only cover 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 as well, having 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 towards 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 for the disorder. With complex disorders the mutation does not always cause the 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 only known 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 for 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 1.2. Next, we give an abstract generic algorithm for HPM in Section 1.3 and present and evaluate three instances of that in Section 1.4. Finally, we give a summary of related work in Section 1.5 and close with a discussion in Section 1.6.

1.2 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 1.2.1. 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].

Marker data is only used for making inferences about the genealogical history of the chromosomes in the sample, the actual disease predisposing polymorphisms are not typically expected to be among the markers. If two chromosomes have the same allele at a marker, the allele may be identical by descent (IBD), inherited from a relatively recent common ancestor. It is possible that two copies of same allele have different mutation histories, in which case the two alleles are said to be identical by state (IBS). On the other hand, different alleles at a marker in a pair of chromosomes do not completely exclude the possibility of a recent common ancestor; the marker may have mutated recently in one of the two lineages, or there might have been a genotyping error.

Linkage. The concept of *linkage* is crucial for gene mapping. In meiosis the human chromosomes are organized as homologous pairs lined up next to each other. In a random recombination process, these aligned chromosomes may form crossovers and exchange parts. Recombination can be modeled with reasonable accuracy as a Poisson process. The number of crossovers over a given genetic distance d follows Poisson distribution with mean d, and the distance between two consecutive cross-overs follows exponential distribution with intensity parameter d. As a consequence, loci close to each other in the same chromosome are closely linked, and crossovers in between are rare. Genetic distances between loci are measured in Morgans (M): one Morgan is the distance at which the expected number of crossovers in a single meiosis is one. The relationship between genetic distance and physical distance measured in basepairs (bp) is such that on the average roughly 1 Mb corresponds to 1 cM, but the ratio varies a lot throughout the genome.

Linkage disequilibrium. Because of recombinations, in a hypothetical infinite randomly mating population all markers would eventually be in *linkage equilibrium*, totally uncorrelated. The opposite phenomenon—*linkage disequilibrium* (LD)—may arise from many different sources; random drift due to finite population size, recent population admixture, population substructure, etc. From gene mapping perspective, utilizable LD in present population results from chromosomes sharing fragments where no cross-overs have taken place since the most recents common ancestor. Genetic bottlenecks, where an initially small population has gone through a relatively long period of slow growth followed by rapid expansion, are an important source for this type of LD. As the initial population is quite small, only a handful of copies of a mutation, the founder mutations, may have entered the bottleneck in different founder haplotypes. The effect of drift is at its strongest during the period of slow growth, skewing the distribution of the founder mutation frequencies. Consequently, only few of the founder mutations are likely to be present in the current population in significant

DRAFT
Accepted for publication in 'Data Mining in Bioinformatics'
Jason Wang, Mohammed Zaki, Hannu Toivonen, and Dennis Shasha (Eds.), Springer

numbers. Small isolated founder populations such as Kainuu in North-Eastern Finland or the French-Canadians are examples of recent bottlenecks. The whole Caucasian population is thought to have gone through a bottleneck approximately 50,000 years ago migrating out of Africa [4]. LD decays over time, as the chromosomes get more fragmented and conserved regions get shorter (Figure 1.1). The expected length of a region conserved over g generations is 2 M/g. LD resulting from the "out of Africa" bottleneck can still be observed over a 100 kb range.

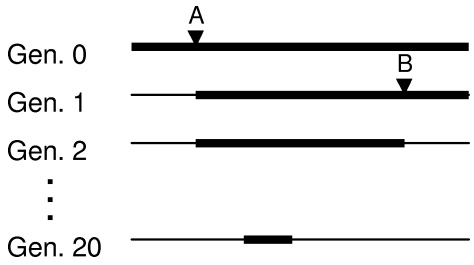


Fig. 1.1. The evolution of a chromosomal region over 20 generations. The thicker line represents fragments from the original chromosome at generation 0. In the first two meioses, crossovers at locations A and B have replaced the ends of the chromosome by material from other chromosomes. After 20 generations only a short fragment of the original chromosome has remained intact.

For an investigator, linkage disequilibrium is both a friend and an enemy. Because of the confounding effect, nearby polymorphisms are correlated, and other markers can be used as surrogates for the disease susceptibility mutation. Therefore a reasonably dense map of markers covering the genomic region under study can be sufficient for gene mapping. Furthermore, without LD all polymorphic loci would be independent of each other, leading to an unbearable multiple testing problem. On the other hand, LD makes it extremely hard to tell which polymorphism is behind the trait. Recent studies [15] show that in Caucasian populations the genome consists of blocks of 20–100 kb, where there are effectively only a handful of different haplotypes in each, and no crossovers can be observed. It may be impossible to map polymorphisms inside a block, yet a single block can contain several genes.

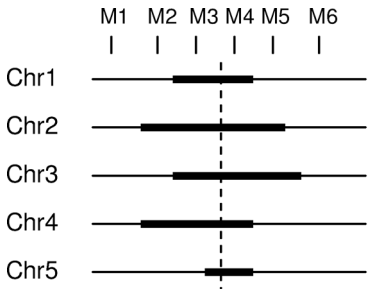
Gene mapping paradigms. Family studies using marker data from extended pedigrees or sib-pairs are based on detecting crossovers using a sparse marker map. Roughly, the idea is to predict the location of the DS gene to be where the marker alleles co-segregate with the trait value. However, due to the relatively small number of crossovers observable in pedigrees the resolution of such studies is not particularly good. Therefore family-based linkage analysis is used as the first step of a mapping project, to guide which regions to focus on in subsequent analyses.

DRAFT
Accepted for publication in 'Data Mining in Bioinformatics'
Jason Wang, Mohammed Zaki, Hannu Toivonen, and Dennis Shasha (Eds.), Springer

Case–control studies of independent individuals can in principle take advantage of a much larger number of historical crossovers in the (unknown) genealogy leading to the observed sample. It is only possible to get indirect evidence of these crossovers, in the form of shared patterns apparently inherited from a common ancestor, adding to the uncertainty of the analysis.

The concept of IBD generalizes to chromosomal regions: A region is IBD in a homologous pair of chromosomes if no crossovers have occurred in either of the lineages since the most recent common ancestor. As a result, haplotypes for any set of markers within the IBD region are identical save for marker mutations. Multi-marker haplotypes are more informative than single alleles, and consequently haplotype sharing is more convincing evidence of IBD status.

All the chromosomes bearing a mutation inherited from a common ancestor also share a varying amount of the surrounding region IBD (Figure 1.2). All case–control methods are based on detecting haplotypes corresponding to these IBD regions, and their association to the trait. In the proximity of the DS gene, LD can be increased artificially via the selection of the study subjects. If the affected are over-represented in the sample, the set of haplotypes will be enriched with the haplotype bearing the DS mutation. This is particularly useful if the causal mutation is rare.



	M1	M2	M3	M4	M5	M6
H1	1	3	3	2	1	1
H2	4	2	3	2	2	3
H3	2	1	3	2	2	2
H4	3	2	3	2	1	3
H5	1	1	2	2	1	1

Fig. 1.2. A set of chromosomes that are IBD at the location denoted by the vertical dashed line also share a varying length of the surrounding sequence (on the left). This sharing is reflected by the corresponding haplotypes (on the right).

Acquisition of data. The two most common types of markers are single nucleotide polymorphisms (SNP) and short tandem repeats (STR), also known as microsatellites, where the number of repeats of a short sequence, typically 2-4 bases, varies. STRs are the more informative of the two, the number of alleles may be more than a dozen. The number of alleles in SNPs is two, but SNPs are much more frequent in the genome, and thus enable denser marker maps and are more suitable for fine mapping. SNPs are also more stable than STRs. Mutation rates for SNPs are estimated at 10^{-8} per meiosis, whereas for STRs they can be as high as 10^{-3}.

The human genome is organized in 22 pairs of homologous chromosomes (autosomes), and a pair of sex chromosomes. A marker residing in an autosome or in the pseudo-autosomal region of the sex chromosomes has two instances in any individual. The process of reading the alleles at a marker is called *genotyping*, and the pair of alleles is the *genotype* at the marker. Current laboratory techiques produce *phase-unknown* genotypes; there is no telling which of the two alleles is of paternal or maternal origin. Term genotype also applies to any set of markers; a multi-marker genotype is the array of the single marker (phase-unknown) genotypes.

Since laboratories produce phase-unknown genotype data, haplotypes are not readily available for analysis. Haplotypes can be inferred based on genotypes from relatives. The most common procedure for obtaining case–control haplotype data is to genotype family trios consisting of the parents and a child. Assuming that the genotypes are known for all three, the phases of the alleles of the child can be determined in all cases but the one in which all the three have similar heterozygous genotype at the marker.

Example 1.2.2. Assume that the phase-unknown genotypes over two markers in a family trio are

	M1	M2
father	1,2	1,2
mother	2,3	1,2
child	2,3	1,2

For the first marker we can infer the alleles that the child has inherited from the mother(3) and the father(2), but for the second marker there is no way to determine the phases.

Additionally, the non-transmitted parental alleles are also determined. As a result, four independent haplotypes can be obtained from a trio; the two transmitted and the two non-transmitted pseudo-haplotypes. Note that the non-transmitted pseudo-haplotypes are the complements of the transmitted haplotypes with respect to the parental genotypes, and do not necessarily correspond to any real chromosomes.

At present time, the cost of genotyping in a large scale mapping study is considerable. The need to detect DS genes in relatively small samples motivates the development of more powerful methods for *in silico* analysis of marker data.

1.3 Haplotype patterns as a basis for gene mapping

In this section we present a general framework, Haplotype Pattern Mining (HPM), for gene mapping based on haplotype patterns. HPM tests each marker for association based on haplotype sharing around it. HPM looks for

patterns in the marker data that could be informative of the location of a DS gene. Since the information is essentially contained in haplotypes reflecting IBD sharing in a part of a chromosome, the patterns are haplotypes over subsets of the marker map that are likely to correspond to such IBD regions.

In the following subsections we first present the generic HPM algorithm in terms of three components; language $\mathcal{L}$ of haplotype patterns, qualification predicate q over $\mathcal{L}$, and marker scoring function s. Then we give a detailed description for each of the components.

1.3.1 Outline of the algorithm

The input for HPM consist of marker data (either a set of haplotypes, a set of phase-unknown genotypes, or a combination of both), and the associated trait values. Optionally the input may also include a set of explanatory covariates such as body mass index, age, sex, blood measurements, etc. Formally, let $M = \{1, \ldots, m\}$ be the marker map, and D be a $n \times m$ matrix of marker data; its columns correspond to markers, and rows correspond to observations, which may be haplotypes or genotypes. If the ith observation is a haplotype, then $D_{ij} \in \mathcal{A}_j \cup \{0\}$, otherwise $D_{ij} \in (\mathcal{A}_j \cup \{0\})^2$. $\mathcal{A}_j$ is the set of alleles at marker j, and 0 denotes a missing allele value. With genotype data the order of the alleles in a pair is insignificant. Let Y be the vector of trait values associated with the haplotypes and genotypes. The trait may be dichotomous or quantitative. In case of haplotypes derived from a trio, one can use the trait value of the child for the transmitted haplotypes and the trait value of the respective parent for the non-transmitted haplotypes. Let X be the matrix containing additional covariates.

The generic HPM works as follows. First, all potentially interesting haplotype patterns are searched for. Let $\mathcal{L}$ be a language of haplotype patterns, and q be a qualification predicate over $\mathcal{L}$: $q(\boldsymbol{p})$ is true iff $\boldsymbol{p}$ is a potentially interesting pattern in the given data set. Practical choices for q set a lower bound for the number of occurrences of a pattern in the data set. Second, a score is calculated for each marker based on the relevant subset of potentially interesting patterns. For a given marker, only patterns that are likely to reflect IBD sharing at the marker are taken into account. Let $s : 2^{\mathcal{L}} \times \mathrm{Perm}(Y) \rightarrow \mathbb{R}$ be a scoring function. $\mathrm{Perm}(Y)$ denotes the set of all permutations of vector Y. The score for marker j given trait vector Y is $s(Q \cap R_j, Y)$, where Q is the set of potentially interesting patterns and $R_j \subseteq \mathcal{L}$ is the set of patterns that are relevant at marker j. Finally, the statistical significance of the scores is measured, resulting in a P value for each marker, and an overall P value corrected for testing over multiple markers. This necessitates the definition of a null hypothesis, and a means for comparing the observed scores to the distribution of the scores under the null hypothesis.

HPM does not model the process generating the trait values or marker data. Therefore we can only test the association between the trait and features

of the marker data. The null hypothesis "*The values of the trait vector are independent of the haplotypes and genotypes*" can be tested by randomizing the relationship between the two using a permutation test. We require q to be invariant with respect to permutations of the trait vector. This way we can enumerate the set of patterns satisfying q once, and use the set for calculating the marker-wise scores in the permuted data as well.

Algorithm Generic-HPM
Input: Pattern language $\mathcal{L}$, qualification predicate q, scoring function s, marker data D, trait vector Y and possibly covariates X.
Output: Marker-wise scores y_j and P values P_j for each marker j, a corrected overall P value.

Method:

1. Find all potentially interesting patterns: $Q = \{\boldsymbol{p} \in \mathcal{L} \mid q(\boldsymbol{p})\}$
2. Compute the score for each marker j : $y_j = s(Q \cap R_j, Y)$
3. For $i \in \{1, \ldots, r\}$, where r is the number of iterations in the permutation test, do
4. Generate a randomly permuted trait vector $Y^{(i)} \in \mathrm{Perm}(Y)$
5. Compute the score for each marker j : $y_j^{(i)} = s(Q \cap R_j, Y^{(i)})$
6. Compute marker-wise P values for each marker by contrasting the observed scores to the samples drawn from the null distributions
7. Compute an overall corrected P value for the best finding

Fig. 1.3. Algorithm for generic HPM. Details are given in the text.

The algorithm for generic HPM is given in figure 1.3. The marker-wise P value can be used for predicting the location of the DS gene. The marker with the lowest P value is a natural choice for a point estimate. The corrected P value is good for assessing whether there is a DS gene in the investigated region in the first place or not.

1.3.2 Haplotype patterns

Haplotype patterns serve as discriminators for chromosomal regions that are potentially shared IBD by a set of chromosomes in the data set. Language $\mathcal{L}$ of haplotype patterns consists of haplotypes over subsets of the marker map, with a few constraints. Marker maps with over hundred markers are not uncommon today, in the near future maps of several thousands of markers can be expected. The number of possible haplotypes grows exponentially with the number of markers in the map. It is not possible to consider all the possible haplotypes in the analysis, but on the other hand, not all haplotype patterns are biologically conceivable. Meaningful patterns correspond to IBD sharing between chromosomes, so markers included in a pattern should form a contiguous block. Allowing a restricted number of wildcards within a pattern may be desirable, as there may be marker mutations breaking an otherwise

DRAFT
Accepted for publication in 'Data Mining in Bioinformatics'
Jason Wang, Mohammed Zaki, Hannu Toivonen, and Dennis Shasha (Eds.), Springer

IBD region, or there may be markers having a lot of missing or erroneous allele values. Additionally, haplotypes extending over very long genetic distances are highly unlikely to survive over many generations and meioses, and therefore the set of patterns to be considered can be restricted with an upper limit for the genetic distance between the leftmost and rightmost markers that are assigned with an allele.

Let $\boldsymbol{p} = [p_1 \cdots p_m]$ be a haplotype pattern, where $p_j \in \mathcal{A}_j \cup \{*\}$, where $\mathcal{A}_j$ is the set of alleles at marker j, and $*$ is a wildcard symbol which matches any allele in the data. Pattern $\boldsymbol{p}$ overlaps marker j, or marker j is within pattern $\boldsymbol{p}$, if j is between the leftmost and rightmost markers bound in $\boldsymbol{p}$ (inclusive). Length of $\boldsymbol{p}$ can be defined as either 1) the genetic distance between the leftmost and rightmost marker bound in $\boldsymbol{p}$, or 2) the number of markers between and including the leftmost and rightmost marker bound in $\boldsymbol{p}$. We define language $\mathcal{L}$ of patterns as set of such vectors $\boldsymbol{p} = [p_1 \cdots p_m]$, where length$(\boldsymbol{p}) \leq \ell$ and either 1) the number of wildcards $(*)$ within $\boldsymbol{p}$ is at most w, or 2) the number of stretches of consecutive wildcards within $\boldsymbol{p}$ is at most g, and the length of such stretches is at most ℓ_{G}. Pattern parameters ℓ, w, g, and ℓ_{G} are given by the user.

Haplotype i matches pattern $\boldsymbol{p}$ iff for all markers j holds: $p_j = *$ or $p_j = D_{ij}$. The frequency of pattern $\boldsymbol{p}$, freq$(\boldsymbol{p})$, is the number of haplotypes matching $\boldsymbol{p}$. With genotype data things are more complicated; a match is certain only if at most one of the markers assigned with an allele in the pattern is heterozygous in a genotype. A match is possible if at least one of the alleles at each marker in the genotype matches the corresponding allele in the pattern. One possibility for handling the uncertain cases is optimistic matching, where a genotype matches a pattern if any of the possible haplotype configurations matches it: genotype i matches pattern $\boldsymbol{p}$ iff for all markers j holds: $p_j = *$ or $p_j = g_1$ or $p_j = g_2$, where $(g_1, g_2) = D_{ij}$. In Section 1.4.3 we will show that this simplistic approach works surprisingly well. More elaborate schemes are possible, e.g. genotypes can be weighted by 2^{1-n}, where n is the number of heterozygous markers in the genotype which are also assigned with an allele in the pattern.

Example 1.3.1. Let $\boldsymbol{p} = [* \ * \ 1 \ * \ 2 \ *]$ be a haplotype pattern over markers $(1, \ldots, 6)$. $\boldsymbol{p}$ overlaps markers 3, 4 and 5 and is matched by for example haplotype [3 2 1 4 2 0] and genotype [(1,1) (1,2) (1,1) (2,4) (1,2) (2,3)]. Genotype [(1,1) (1,2) (1,2) (2,4) (1,2) (2,3)] may match $\boldsymbol{p}$, depending whether allele 1 at marker 3 and allele 2 at marker 5 are from the same chromosome or not. With optimistic matching, we consider this possible match as a match.

In the instances of HPM we have used, the qualification predicate is based on a minimum frequency: $q(\boldsymbol{p}) \equiv \mathrm{freq}(\boldsymbol{p}) \geq f_{\min}$, where the minimum frequency is either given by the user or derived from other parameters and some summary statistics of the data such as sample size and the number of disease-associated and control observations.

1.3.3 Scores

The purpose of the scoring function is to produce a test statistic for each marker, measuring total association of the marker to the trait over all haplotype patterns that are relevant at the marker. The higher the score, the stronger the association is. We define the set R_j of relevant patterns at marker j as the set of patterns overlapping marker j.

A very simple—yet powerful—scoring function, used in [21, 22], counts the number of strongly disease-associated patterns overlapping the marker:

$$s(Q_j, Y') = |\{\boldsymbol{p} \in Q_j \mid A(\boldsymbol{p}, Y') \geq a_{\min}\}|, \tag{1.1}$$

where $Q_j = Q \cap R_j$, and $A(\cdot)$ is a measure for pattern–trait association or correlation. The association threshold $a_{\min}$ is a user-specified parameter. Table 1.1 illustrates the procedure.

Another scoring function, used in [17, 14], measures the skew of the distribution of pattern–trait-association in the set of overlapping patterns. The skew is defined as a distance between the set of P values test of pattern–trait association tests for the patterns in Q_j and their expected values if there was no association:

$$s(Q_j, Y') = \frac{1}{k} \sum (P_i(Y') - U_i) \log \frac{P_i(Y')}{U_i}, \tag{1.2}$$

where $k = |Q_j|$, $P_1(Y'), \ldots, P_k(Y')$ is the list of P values sorted into ascending order, and $U_1, \ldots, U_k$ are the expected ranked P values assuming that there is no association and that patterns are independent, $U_i = \frac{i}{k+1}$.

Both scoring functions described above consider each pattern as an independent source of evidence. In reality, the patterns are far from independent, but the assumption of independence is a useful approximation. An ideal scoring function would take the structure in Q_j into account.

In all current instances of HPM the scoring function measures the pattern–trait association independently for each pattern. A pattern whose occurrence correlates with the trait is likely to do well in discriminating the chromosomes bearing the mutation. What is a meaningful test for this correlation depends on the type of data. With a dichotomous trait, e.g. affected–unaffected, association can be simply tested using Z-test (or χ^2-test) or Fisher's exact test for a 2 by 2 contingency table, where the rows correspond to the trait value and the columns to the occurrence of the pattern:

	M	N	$\sum$
A	n_{AM}	n_{AN}	n_{A}
U	n_{UM}	n_{UN}	n_{U}
$\sum$	n_{M}	n_{N}	n

Let us assume that there are n_{M} observations that match pattern $\boldsymbol{p}$, and n_{N} observations that do not match $\boldsymbol{p}$, and that there are n_{A} affected

Table 1.1. This table illustrates the computation of marker-wise scores with association threshold $Z_{\min} = 3$. The patterns are ordered by the strength of association. Note that the wildcards within a pattern are included in the score for that marker.

Pattern	M1	M2	M3	M4	M5	M6	Z
$\boldsymbol{p}_1$	*	*	2	*	1	*	5.8
$\boldsymbol{p}_2$	*	1	2	1	3	*	4.4
$\boldsymbol{p}_3$	*	2	2	*	1	*	4.0
$\boldsymbol{p}_4$	1	2	2	*	1	*	3.4
$\boldsymbol{p}_5$	*	1	2	1	3	3	2.8
Score	1	3	4	4	4	0	

and n_{U} unaffected observations in total. Let the frequencies in the 2 by 2 contingency table, where the rows correspond to the trait value (A or U), and the columns to matching (M) or not matching (N) $\boldsymbol{p}$, be n_{AM}, n_{AN}, n_{UM} and n_{UN}, respectively. The value of the test statistic

$$Z = \frac{(n_{\mathrm{AM}} n_{\mathrm{UN}} - n_{\mathrm{UM}} n_{\mathrm{AN}})\sqrt{n}}{\sqrt{n_{\mathrm{M}} n_{\mathrm{N}} (n_{\mathrm{AM}} + n_{\mathrm{AN}})(n_{\mathrm{UM}} + n_{\mathrm{UN}})}} \tag{1.3}$$

is approximately normally distributed. One- or two-tailed test can be used. One-tailed test is appropriate if one is only interested in patterns with a positive correlation to the trait. Assuming that there are no missing alleles in the data, it is possible to derive a lower bound for pattern frequency given the association threshold:

$$f_{\min} = \frac{n_{\mathrm{A}} n x}{n_{\mathrm{C}} n + n x}, \tag{1.4}$$

where x is the association threshold for χ^2 statistic, or the Z threshold squared (see [21] for details). No pattern with a frequency lower than $f_{\min}$ can be strongly associated. Even if there are missing alleles, this lower bound can be used—it is not imperative that all the strongly associated patterns satisfy q.

With a quantitative trait one can use the two-sample t-test for identical means between the group of chromosomes matching the pattern and those not matching it. The number of degrees of freedom (number of chromosomes minus two) is usually large enough to justify the use the Z-test instead of the t-test.

If explanatory covariates are included in the data, one can formulate a linear model

$$Y_i = \alpha_1 X_{i1} + \ldots + \alpha_k X_{ik} + \alpha_{k+1} I_i + \alpha_0, \tag{1.5}$$

where Y_i is the trait value for chromosome i, X_{ij} is the value of the jth covariate for the ith observation, and I_i is an indicator variable for the occurrence of the tested pattern. Its value is 1 if the pattern matches the ith observation, otherwise 0. The significance of the pattern as an explanatory variable can be tested by comparing the best fit model to the best fit model where $\alpha_{k+1} = 0$.

Missing alleles in the observations are dealt with in a conservative manner: If an allele is missing at a marker bound in pattern $\boldsymbol{p}$, and there is a mismatch in any other marker, then the observation is counted as a mismatch. Otherwise we cannot know for sure whether $\boldsymbol{p}$ occurs in the observation, and to avoid any bias we ignore the observation when calculating the association for pattern $\boldsymbol{p}$.

1.3.4 Searching for potentially interesting haplotype patterns

Let $\preceq$ be a generalization relation in $\mathcal{L}$: $\boldsymbol{p} \preceq \boldsymbol{p}'$ if any observation matching $\boldsymbol{p}'$ also matches $\boldsymbol{p}$. q is monotonous in $\preceq$ if $\boldsymbol{p} \preceq \boldsymbol{p}' \wedge q(\boldsymbol{p}') \Rightarrow q(\boldsymbol{p})$, which is true for $q(\boldsymbol{p}) \equiv \mathrm{freq}(\boldsymbol{p}) \geq f_{\min}$. With monotonous q, set Q of patterns satisfying q can be efficiently enumerated using data-mining algorithms [2], or standard depth-first search (implementation for HPM given in [22]). Otherwise, one can introduce a monotonous auxiliary predicate q_{m} such that $q(\boldsymbol{p}) \Rightarrow q_{\mathrm{m}}(\boldsymbol{p})$. The set of patterns satisfying q_{m} can be enumerated as described above, and each of these patterns can then be individually tested for q.

With some choices for q and s it is possible that pattern p does not contribute to the score of any marker in any permutation of Y even if $q(\boldsymbol{p})$ holds. Marginal speed-up can be achieved, if q in Step 1 of the algorithm is replaced with $q' : q'(\boldsymbol{p}) \equiv q(\boldsymbol{p}) \wedge \exists j, Y' \in \mathrm{Perm}(Y) : \boldsymbol{p}$ contributes to $s(Q \cap R_j, Y')$.

Example 1.3.2. Let us assume Z-test is used with a dichotomous trait, $q(\boldsymbol{p}) \equiv \mathrm{freq}(\boldsymbol{p}) \geq f_{\min}$, and $s(Q', Y') = |\{\boldsymbol{p} \in Q' \mid Z(\boldsymbol{p}, Y') \geq Z_{\min}\}|$. Maximum value attainable for Z can be calculated based on the numbers of matching and non-matching observations. If the maximum value is below the association threshold $Z_{\min}$, the pattern is rejected. Given n, n_M, n_{N}, n_{A} and n_{U}, the largest Z value is achieved, when n_{AM} and n_{UN} are maximized: If $n_{\mathrm{M}} \geq n_{\mathrm{A}}$ then $n_{\mathrm{AM}} = n_{\mathrm{A}}, n_{\mathrm{UM}} = n_{\mathrm{M}} - n_{\mathrm{A}}, n_{\mathrm{AN}} = 0$, and $n_{\mathrm{UN}} = n_{\mathrm{N}}$, else if $n_{\mathrm{M}} \geq n_{\mathrm{C}}$ then $n_{\mathrm{AM}} = n_{\mathrm{M}}, n_{\mathrm{UM}} = 0, n_{\mathrm{AN}} = n_{\mathrm{N}} - n_{\mathrm{C}}$, and $n_{\mathrm{UN}} = n_{\mathrm{C}}$, otherwise $n_{\mathrm{AM}} = n_{\mathrm{M}}, n_{\mathrm{UN}} = n_{\mathrm{N}}$, and $n_{\mathrm{AN}} = n_{\mathrm{UM}} = 0$. If negative associations are considered, the minimum value of the Z statistic has to be calculated as well. This can be done analogously, by swapping A and U in the formulae. Similar procedure is possible for Fisher's exact test.

1.3.5 Evaluating statistical significance

With real data the allele frequencies and marker spacing vary across the marker map. Consequently, the distribution of scores varies as well, and the

DRAFT
Accepted for publication in 'Data Mining in Bioinformatics'
Jason Wang, Mohammed Zaki, Hannu Toivonen, and Dennis Shasha (Eds.), Springer

scores as such are not necessarily a good indicator of the location of the DS gene. Instead, the significances of the marker-wise scores should be evaluated. HPM computes empirical P values for the markers using a permutation test. Figure 1.4 illustrates a succesful localization with simulated data.

Let $y_j^{(1)}, \ldots, y_j^{(r)}$ be the sample from the score distribution for marker j under the null hypothesis, and let y_j be the observed score at the marker. The empirical P value for the marker is then

$$\hat{P} = \frac{|\{i \in \{1, \ldots, r\} \mid y_j^{(i)} \geq y_j\}|}{r}.$$

As always with permutation tests, the number of iterations should be sufficiently large for the P value estimates to be accurate. $\hat{P} \sim \frac{1}{r}\mathrm{Bin}(r, P)$, and its standard deviation is $\sqrt{\frac{1}{r}P(1-P)}$. As a rule of thumb, at the desired significance level at least 50 iterations should have a score greater than the critical value, e.g. at $\alpha = 0.05$ at least 1,000 iterations should be performed.

The marker-wise P values are not corrected for testing over multiple markers, and they should be understood as a means of ranking the markers only. However, a single corrected P value for the best finding can be obtained with another permutation test using the smallest marker-wise P value as the test statistic. This P value can also be used to answer the question whether there is a DS gene in the investigated region in the first place or not. The two nested permutations can be carried out efficiently at the cost of a single test (see [18] for details).

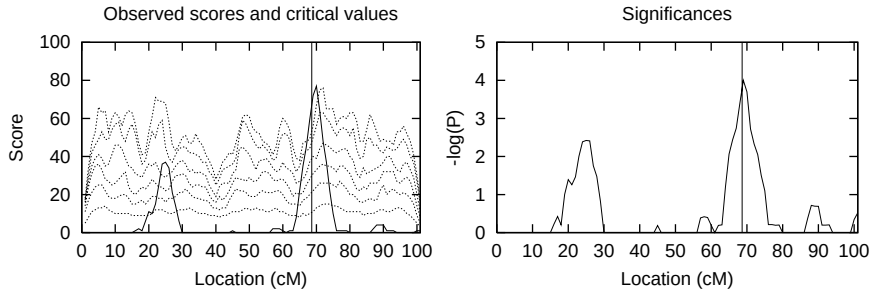


Fig. 1.4. The graph on the left shows the scores (solid line) and critical values at significance levels $\alpha = 0.001, 0.002, 0.005, 0.01, 0.02$ and 0.05 (dotted lines) over 101 evenly spaced markers. The graph on the right shows the negated logarithms (base 10) of the corresponding P values. The vertical line denotes the correct location of the DS gene.

1.4 Instances of the generalized algorithm

We present three instances of the generalized HPM algorithm. The original version for haplotype data and dichotomous trait [21, 22] and QHPM for quantitative traits and covariates [17, 14] have been previously published. In this section we introduce a third instance—HPM-G for phase-unknown genotype data.

We demonstrate the performance of the three instances in various settings using simulated data. We used the Populus simulation package [21, 13] for generating realistic data sets for the analyses. In each of the simulations a small isolated founder population was generated, growing exponentially from initial 100 people to 100,000 over 20 generations. In each setting, a single 100 cM chromosome was simulated. The marker maps consisted either of 101 microsatellite markers or 301 SNP markers equidistantly spaced over the chromosome. Denser map was used with SNP markers because a single SNP marker is much less informative than a microsatellite marker. Each simulation was repeated 100 times in order to facilitate power analysis. We are interested in the localization power as a function of the tolerated prediction error. For example, in Figure 1.5A the 60% curve at 2 cM shows that for 70% of the replicates the predicted location was no more than 2 cM off the correct location. At the scale of the data sets, a mapping result is considered acceptable if it narrows down the 100 cM chromosome into a 20 cM or smaller region.

We did not apply permutation tests in the power analyses, but used the scores as a basis for the localization instead: the point estimate for the gene location is the marker with the highest score. This way we were able to carry out the power analyses in much less time. Because there was no variation in the marker density over the chromosome and the alleles in the initial population were drawn from the same distribution for each marker, the score distributions are likely to be quite similar for all markers. We have previously shown that on this kind of data it does not make much difference whether the localization is based on the P values or the scores [21].

1.4.1 Original HPM for haplotype data and dichotomous trait

In the original version of HPM for haplotype data and dichotomous trait, we use the simple scoring function counting the number of strongly associated patterns, described in Equation 1.1. The χ^2-test is used for measuring pattern–trait-association and only positively associated patterns are considered. The frequency threshold is derived from association threshold x using Equation 1.4.

Marker map consisted of microsatellite markers each with one common allele with frequency 0.4 and four alleles with frequency 0.15. The frequency of the disease predisposing mutation was approximately 2% in each data set. The ascertainment of data was conducted as follows: 100 trios with an affected

child were randomly chosen from the final population. The haplotypes were reconstructed from the trios and all uncertain alleles were set to zero denoting a missing value. The parameters for HPM were the same for all experiments: maximum lenth for patterns was 7 markers, association threshold was 9 and one gap of up to two markers was allowed. The execution time was less than one second without the permutation test, and about 20 seconds with 1,000 permutations for a single replicate on a Pentium4 at 1.4GHz.

First, we simulated data sets with different phenocopy rates ranging from 60% to 90%. The results in Figure 1.5A show that the localization power reaches its maximum at phenocopy rate between 60%–70%, and decreases steadily with increasing phenocopy rate as expected.

Next, we assessed the effect of missing data by randomly removing 2%, 5% and 10% of the marker genotypes in the data with 80% phenocopy rate prior to haplotype reconstruction. This procedure resulted in approximately 8%, 15% and 25% of missing alleles in the haplotype data. Due to haplotyping ambiguities, $\sim 4\%$ of the alleles were missing even if there had not been any missing genotypes in the trios. The results in Figure 1.5B show that up to 15% there is practically no loss in power, which demonstrates remarkable tolerance for missing data.

To put the results into perspective, we compare HPM to TDT of Genehunter2 [8]. With TDT we considered haplotypes up to four markers in length (maximum in Genehunter2), and used the centerpoint of the best haplotype as the point estimate. Results (Figure 1.5C) show that at phenocopy rate of 80% there is virtually no difference between the methods, but at higher rates HPM is clearly superior.

Finally, we showcase the method on a real Type I diabetes data set [3, 21]. There are 25 markers spanning a region of 14 Mb in the data. Two DS genes are known to reside in the region, very close to each other. We downsampled the original data set consisting of 385 sib-pair families to 100 trios (half the data size we used in [21]). The results obtained with 100,000 permutations are shown in Figure 1.5D. The marker closest to the genes gives the second best P value 0.00014. The corrected overall P was 0.0015, indicating that the observed association is highly unlikely to be a coincidence.

1.4.2 QHPM for quantitative trait and covariates

The diagnostics of a complex disease is often based on a combination of symptoms and quantitative measurements. For example, a possible diagnosis is $(X_1 \geq A \wedge (X_2 \geq B \vee S))$, where X_1 and X_2 are values of two quantitative subtraits, and S is a proposition for a symptom. Different patients may have completely different genetic contributors and pathogenesis. It may be easier to find the quantitative trait loci (QTLs) affecting each of the subtraits independently than trying to map all the DS genes directly based on the affection status only.

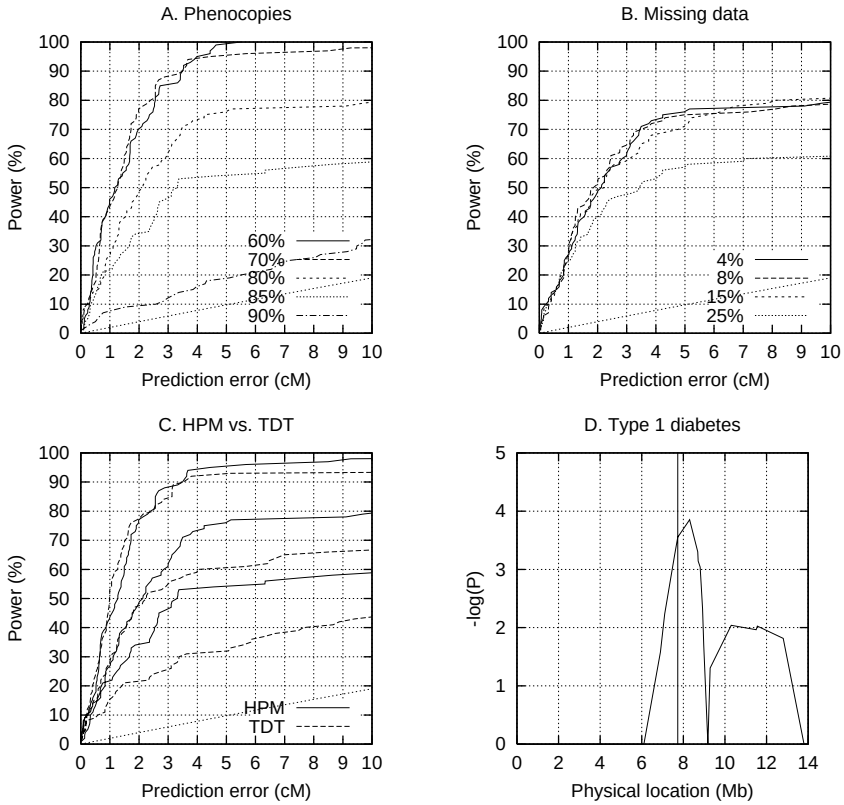


Fig. 1.5. Performance of HPM. A) Effect of phenocopy rate on localization accuracy. B) Effect of the amount of missing alleles on localization accuracy. C) A comparison between HPM and the multipoint TDT of Genehunter2. Phenocopy rates were at 80%, 85% and 90%. The dotted curve on the bottom of every power graph denotes the power of random guessing. D) Succesful localization on real Type 1 diabetes data. The vertical line shows the correct location.

The original HPM can only cope with a dichotomous trait. Generally, dichotomization of a quantitative variable wastes much of the information. Additionally, the power of the analysis is sensitive to the cut-off point. There may be other information about the subjects available, environmental and other non-genetic factors, e.g. smoking or nutritional habits, and measurements that are not related with the diagnosis criteria. To be able to fully utilize the available data, a method should be capable of

- using a quantitative trait as the response variable, and
- using the other measurements as explanatory covariates.

DRAFT
Accepted for publication in 'Data Mining in Bioinformatics'
Jason Wang, Mohammed Zaki, Hannu Toivonen, and Dennis Shasha (Eds.), Springer

QHPM is a version of HPM designed to meet the abovementioned criteria. It uses the linear model given in Equation 1.5 for measuring pattern–trait association, and the scoring function given in Equation 1.2. We next assess the performance of QHPM on simulated data and compare it to QTDT [1], an accommodated version of TDT. The results have been previously published in [14].

The simulations were carried out in the manner described in Section 1.4.1, except that there were only four alleles for each marker; one common allele with initial frequency 0.4, and three alleles with frequency 0.2. The disease predisposing mutation was inserted to six randomly chosen chromosomes in the initial population. Liability for the disease was calculated using formula

$$L = Ag + e_1 + e_2 + r_1 + C,$$

where g is an indicator variable for the presence of the mutation in the individual, e_1 and e_2 are environmental factors, and r_1 is a random component. e_1, e_2 and r_1 are drawn from standard normal distribution for each individual. The strength of the genetic effect is determined by A. The probability of being affected was given by the *expit* function

$$P(\text{Affected}) = \frac{e^L}{1 + e^L}.$$

Two models were considered; an easy model with $A = 5$, and a difficult model with $A = 2$. The value of C was adjusted so that the prevalence of the disease is 5%. Additionally, five different quantitative variables were calculated from formula

$$Q_j = jg + e_1 + e_2 + r_2,$$

where $j \in \{1, \ldots, 5\}$ determines the strength of the genetic effect, and r_2 is a random component drawn from the uniform distribution in [0,1]. The sample was ascertained based on the affection status; 200 trios with an affected child were randomly selected from the final population.

The maximum length of patterns was set to 7 markers, and a single one marker gap was allowed. Minimum pattern frequency $f_{\min}$ was 10. The results in Figure 1.6 show, that QHPM clearly outperforms QTDT with the difficult model. With the easy model, QHPM has a slight edge with Q_5 and Q_3, whereas with Q_2 QTDT gives better results. Q_1 turned out to be too difficult for mapping; neither method could do better than random guessing either with the easy or the difficult model.

1.4.3 HPM-G for phase-unknown genotype data

Haplotype data is not always easy to obtain, typically the haplotypes are inferred based on genotypes of family members. The most cost-effective way

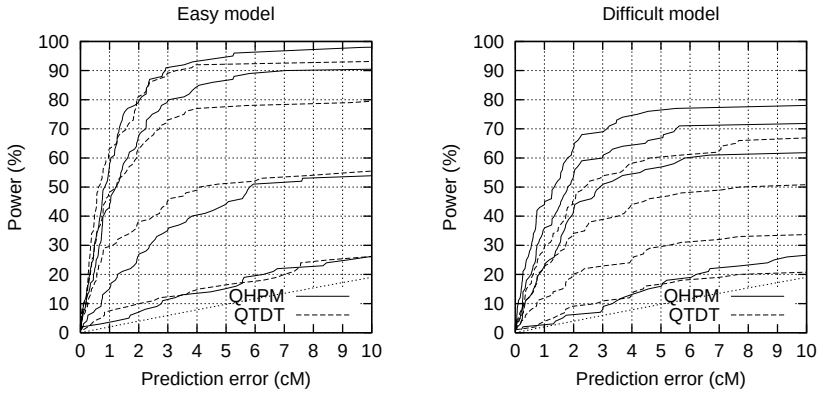


Fig. 1.6. The localization powers of QHPM (solid lines) vs. QTDT (dashed lines) are illustrated for A) the easy model and B) the difficult model. The curves correspond to quantitative traits Q_5, Q_3, Q_2 and Q_1 in top-down order.

to obtain haplotypes for a case–control study is to genotype family trios, from each of which four independent haplotypes can be extracted. The efficiency of genotyping is 2/3, as there in fact are six haplotypes in a trio, and two of them are read twice. The parents would need to be recruited; however, they may be deceased, or not willing to participate. Genotyping of these additional individuals is laborious and elevates the study expenses. Moreover, the phases cannot always be determined in a trio. Using phase-unknown genotype data directly for mapping, no extra individuals need to be genotyped, and no data is missing due to haplotyping ambiguities. Additionally, recruiting problems are alleviated and there is more freedom in selecting the cases and controls, including the ratio between the two classes.

The abstract formulation of HPM allows us to easily adapt it for genotype data. HPM for genotype data (HPM-G) is identical to the original version, with the exception that optimistic pattern matching is used. All the matches in the real haplotypes are found, but also a large number of spurious matches, which introduce noise to the marker-wise scores. Consequently, the number of frequent patterns found by HPM-G is typically an order of magnitude larger than by HPM.

In order to compare HPM-G to HPM we simulated both microsatellite and SNP data sets in the way described in Section 1.4.1. The data sets were ascertained with equal costs of genotyping, assuming that the haplotypes for HPM are reconstructed from family trios. The haplotype data sets consisted of 200 disease-associated and 200 control haplotypes, derived from 100 trios. The data set for HPM-G consisted of 150 affected and 150 control genotypes. 300 individuals need to be genotyped in both cases.

The parameters used in Section 1.4.1 were used as a basis for parameter settings. With SNP data, the maximum length of patterns was increased to 19 markers, to give equal maximum genetic length of 6 cM. We used an 50% elevated association threshold for HPM-G, as the expected number of mutation carriers in genotype data sets was 50% higher than that in the haplotype data sets. The execution time of HPM-G was about 4 seconds with microsatellite data, or $6\frac{1}{2}$ minutes with SNP data, for a single replicate (Pentium4, 1.4GHz). With 1,000 permutations the execution times are approximately 4 minutes and 6 hours, respectively. The execution time of HPM with SNP data was 6 seconds without permutation test, and 3 minutes and 40 seconds with 1,000 permutations.

We compared the two methods at four different phenocopy rates both with microsateelite and SNP data. From the results shown in Figure 1.7A we can conclude that with microsatellite data HPM-G can tolerate slightly higher phenocopy rates than HPM with equal genotyping costs. With SNP data the methods are evenly matched (Figure 1.7B), but the execution time of HPM-G is much higher. This is due to the fact that with SNP data the number of spurious matches grows considerably.

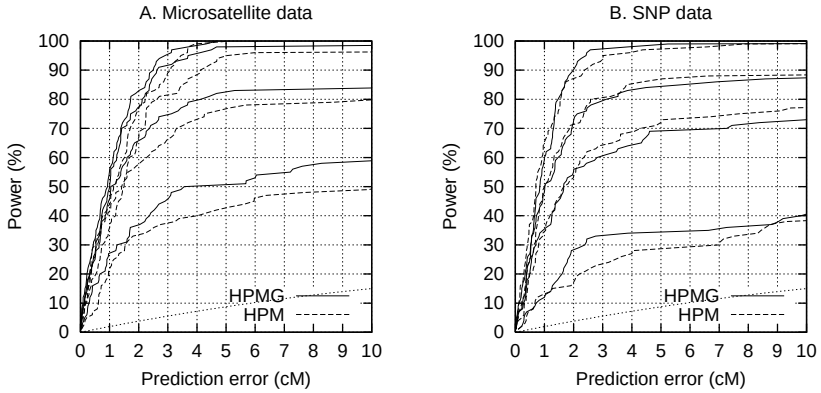


Fig. 1.7. Comparison of HPM-G and HPM with different phenocopy rates—70%, 80%, 85% and 90% in top-down order. A) Localization accuracy on microsatellite data. B) Localization accuracy on SNP data.

1.5 Related work

Fine-scale mapping of disease genes by linkage disequilibrium has been researched intensively since 90's. Lazzeroni [10] gives a good overview of the work until 2000. The earliest work relied on methods which measure

association between the disease status and one marker at a time, or, in other words, the LD between a marker locus and the implicit disease locus [5, 7]. The disease gene is then predicted to be close to the locus with the highest association. Composite likelihood methods by Devlin et al. [6] and Terwilliger [20] consider several markers at a time, but do not utilize any haplotype information.

Service et al. [16] and McPeek and Strahs [11] were among the first to suggest LD-based haplotype analysis methods. The model by Service et al. analyzes LD of the disease to three markers at a time and estimates the disease locus with respect to the three marker loci. McPeek and Strahs are closer to the HPM approach: their method is based on an analysis of the length of haplotype sharing among disease chromosomes. Zhang and Zhao have extended the method to handle phase-unknown genotype data as well [23]. These methods, like most of previous haplotype-based methods, are statistically elegant but computationlly demanding. They tend to be exponential in the number of markers, sometimes in the number of haplotypes.

The implicit assumption of independent haplotypes in the methods mentioned above may be very unrealistic in some populations. Parametric methods by Lam et al. [9] and Morris et al. [12] and non-parametric TreeDT by Sevon et al. [18] model the genealogical relationships among the observed haplotypes.

F-HPM, a variant of HPM, has been suggested independently by Zhang et al. [24]. It extends HPM to use pedigree data and quantitative traits by using a quantitative pedigree disequilibrium test proposed by the same authors.

Linkage analysis is an alternative for LD analysis in gene mapping. The idea, roughly, is to analyze pedigree data and to find out which loci are inherited with the disease. Due to the lower effective number of recombinations, linkage analysis is less suitable for fine mapping than LD analysis. Transmission/disequilibrium tests (TDT) [19] are a well-established way of testing both association and linkage in a sample where LD exists between the disease locus and nearby marker loci.

1.6 Discussion

Gene mapping, the problem of locating disease-predisposing genes, is one of the early steps in many medical genetics studies that ultimately aim at prevention and cure of human diseases. The completion of the human DNA sequence gives a lot of useful information about the genome, in particular about polymorphisms, whether potentially disease-predisposing or useful just as markers in gene mapping studies. Availability of the human DNA sequence does not remove the gene mapping problem, however: we cannot tell from the DNA sequence alone which gene or polymorphism is associated with which trait.

DRAFT
Accepted for publication in 'Data Mining in Bioinformatics'
Jason Wang, Mohammed Zaki, Hannu Toivonen, and Dennis Shasha (Eds.), Springer

From a data mining perspective, the data sets are small. They are, however, growing fast in dimensionality (number of markers) so mapping methods need to be scalable in that respect. Discovery of new knowledge is also an important aspect, even if our discussion has concentrated on predicting the gene location. Geneticists are interested in the patterns that show strong correlation with a disease, and they often investigate them manually, e.g., by constructing possible genealogies to test the plausibility of a DS gene. Strongly disease-correlated patterns or suitable disjunctions of them can sometimes also be useful as putative gene tests before the gene is actually located.

We described Haplotype Pattern Mining, a flexible and generic algorithm for non-parametric gene mapping. It is based on searching for genetic patterns that are strongly associated to the trait under study, and on mapping the disease gene to the genetic region with the most evidence for trait association. HPM incorporates several characteristic components of a typical data mining task:

- definition of an application-specific pattern language,
- searching for frequent patterns, and
- evaluating the strength of rules of form $pattern \rightarrow trait$.

In principle, HPM falls into the category of predictive data mining applications. There is a single variable, the trait, that we attempt to explain using the marker data and possibly other covariates. However, instead of having the classification or regression accuracy as the objective we are more interested in the patterns that are used for prediction and where they are located.

Even though data sets are expected to grow as the laboratory techniques evolve, the pattern search step will probably not become an issue with HPM in the near future. The computational burden mainly results from the subsequent analysis of the pattern set. With a large set of patterns, the permutation test procedure may be quite time consuming. We already saw that with phase-unknown SNP genotype data the execution times were several hours. Ideas for more efficient handling of patterns, e.g., closed patterns, could be utilized to speed up the permutation test.

An advantage of HPM is that it is model-free, as it does not require any—potentially misleading—explicit assumptions about population or mode of inheritance. Experiments show that HPM tolerates high degrees of missing data and high phenocopy rates. By introducing HPM-G for phase-unknown genotype data we have significantly extended the scope of HPM: it can now handle dichotomous or quantitative traits, covariates, SNP and microsatellite markers, and haplotype or genotype data in any combinations. HPM has a clear advantage over many parametric methods: as a by-product HPM gives an explicit list of disease associated patterns accompanied by a variety of statistics. This output is found very informative for the geneticists.

DRAFT
Accepted for publication in 'Data Mining in Bioinformatics'
Jason Wang, Mohammed Zaki, Hannu Toivonen, and Dennis Shasha (Eds.), Springer

Gene mapping is an iterative process: starting with the whole genome, the search successively narrows down the region potentially harbouring the DS genes. New markers are added and possibly new patients are recruited at each iteration. The first stage—the genome scan—is customarily conducted as a family study, using linkage analysis, resulting in candidate regions of 20–30 cM. HPM is best suited to the next stage, where the candidate regions are further reduced down to only few centiMorgans. However, our results on simulated data sets indicate that with a dense enough marker map, HPM could actually be used for a full genome-wide search, at least in populations where LD is expected to extend over several centiMorgans. This may become feasible in near future as genotyping becomes less expensive, and the costs of extra genotyping may become insignificant compared to the costs and difficulties associated with recruitment of families for linkage analysis. Experiments reported in [17] suggest that HPM could be applied to fine-mapping as well—however, proper assessment of the potential for fine-mapping is yet to be done.

Acknowledgements

The authors have developed HPM together with Vesa Ollikainen, Kari Vasko, Heikki Mannila, and Juha Kere. Many thanks to Vesa Ollikainen for providing us with the simulated data sets and some of the analysis results for the experiments with HPM and QHPM.

Glossary

Allele A variant form of a marker or a gene.

Basepair, kb, Mb Two complementary bases forming a single step in a double-stranded DNA or RNA molecule. Length of DNA (or RNA) sequences is measured in basepairs (bp). 1 kb = 1000 bp, 1 Mb = 1000 kb.

Crossover Reciprocal breakage and reunion of two homologous chromosomes. Before reunion the partial chromosomes exchange partners.

Gene A stretch of DNA coding for a protein.

Gene mapping Process which aims at locating a gene affecting a given trait.

Genotype The genetic code of an individual. Specifically, a marker genotype is the pair of alleles at the marker, and a (phase-unknown) multi-marker genotype is a vector of (unordered) allele pairs over the set of markers.

Haplotype A vector of alleles in a single chromosome over a set markers or genes.

Identical by descent, IBD Two alleles or haplotypes are identical by descent, if they have been inherited from a common ancestor unchanged.

Identical by state, IBS Two alleles or haplotypes are identical by state, if they cannot be distinguished by laboratory methods.

Linkage Nearby markers tend to be transmitted together. Linkage between two loci can be expressed quantitatively by recombination fraction (the probability of the loci being separated in a single meiosis).

Linkage disequilibrium, LD Nonrandom association of nearby markers.

Locus (pl. loci) The location of a specific marker or gene in a chromosome.

Marker A polymorphic stretch of DNA for which the variants can be reliably detected.

Morgan, M, cM Genetic distance between two loci, measured in Morgans (M), is defined as the expected number of crossovers between the loci in a single meiosis. 1 M = 100 cM. On average, 1 cM is roughly 1 Mb, but the ratio varies a lot throughout the genome.

Penetrance The probability of the occurence of a phenotype given a genotype.

Phase The parental origin of an allele, maternal or paternal.

Phenocopy A phenotype of non-genetic origin that appears similar to that of genetic origin.

DRAFT
Accepted for publication in 'Data Mining in Bioinformatics'
Jason Wang, Mohammed Zaki, Hannu Toivonen, and Dennis Shasha (Eds.), Springer

Phenotype An observable characteristic or trait of an individual, e.g. presence of a disease.

Prevalence The relative frequency of a disease in a population.

Recombination The interchange of genetic material between two homologous chromosomes during meiosis. In humans this occurs by crossing over.

References

1. G. R. Abecasis, L. R. Cardon, and W. O. Cookson. A general test of association for quantitative traits in nuclear families. *The American Journal of Human Genetics*, 66:279–292, 2000.
2. R. Agrawal, H. Mannila, R. Srikant, H. T. T. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In: U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds) *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, California, pages 307–328, 1996.
3. S. Bain, J. Todd, and A. Barnett. The British Diabetic Association—Warren repository. *Autoimmunity*, 7:83–85, 1990.
4. M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29:229–232, 2001.
5. B. Devlin, and N. Risch A comparison of linkage disequilibrium measures for fine scale mapping *Genomics*, 29:311–322, 1995.
6. B. Devlin, N. Risch, and K. Roeder. Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics*, 36:1–16, 1996.
7. S. W. Guo. Linkage disequilibrium measures for fine-scale mapping: a comparison. *Human Heredity*, 47:301–314, 1997.
8. L. Kruglyak, M. J. Daly, M. P. Reeve-Daly, and E. S. Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *The American Journal of Human Genetics*, 58:1347–1363, 1996.
9. J. C. Lam, K. Roeder, and B. Devlin. Haplotype fine-mapping by evolutionary trees. *The American Journal of Human Genetics*, 66:659–673, 2000.
10. L. C. Lazzeroni. A chronology of fine-scale gene mapping by linkage disequilibrium. *Statistical Methods in Medical Research*, 10:57–76, 2001.
11. M. S. McPeek, and A. Strahs. Assesment of linkage disequilibrium by the decay of haplotype sharing, with applications to fine-scale genetic mapping. *The American Journal of Human Genetics*, 65:858–875, 1999.
12. A. P. Morris, J. C. Whittaker, and D. J. Balding. Fine-scale mapping of disease loci via shattered coalescent modelling of genealogies. *The American Journal of Human Genetics*, 70:686–707, 2002.
13. V. Ollikainen. Simulation techniques for disease gene localization in isolated populations. (Ph.D. thesis) University of Helsinki, Report A-2002-2, 2002.
14. P. Onkamo, V. Ollikainen, P. Sevon, H. T. T. Toivonen, H. Mannila, and J. Kere. Association analysis for quantitative traits by data mining: QHPM. *Annals of Human Genetics*, 66:419–429, 2002.
15. D. E. Reich, S. F. Schaffner, M. J. Daly, G. McVean, J. C. Mullikin, J. M. Higgins, D. J. Richter, E. S. Lander, and D. Altshuler. Human genome sequence variation, and the influence of gene history, mutation and recombination. *Nature Genetics*, 32:135–142, 2002.

DRAFT
Accepted for publication in 'Data Mining in Bioinformatics'
Jason Wang, Mohammed Zaki, Hannu Toivonen, and Dennis Shasha (Eds.), Springer

16. S. K. Service, D. W. Temple Lang, N. B. Freimer, and L. A. Sandkuijl. Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *The American Journal of Human Genetics*, 64:1728–1738, 1999.
17. P. Sevon, P. Onkamo, V. Ollikainen, H. T. T. Toivonen, H. Mannila, and J. Kere. Mining the associations between phenotype, genotype, and covariates. Genetic Analysis Workshop 12, *Genetic Epidemiology*, 21 (Suppl. 1):S588–S593, 2001.
18. P. Sevon, H. T. T. Toivonen, and V. Ollikainen. TreeDT: Gene mapping by tree disequilibrium test. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 365–370, 2001, (Extended version available at http://www.cs.helsinki.fi/TR/C.html).
19. R. S. Spielman, R. E. McGinnis, and W. J. Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetis mellitus (IDDM). *The American Journal of Human Genetics*, 52:506–516, 1993.
20. J. D. Terwilliger. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *The American Journal of Human Genetics*, 56:777–787, 1995.
21. H. T. T. Toivonen, P. Onkamo, K. Vasko, V. Ollikainen, P. Sevon, H. Mannila, M. Herr, and J. Kere. Data mining applied to linkage disequilibrium mapping. *The American Journal of Human Genetics*, 67:133–145, 2000.
22. H. T. T. Toivonen, P. Onkamo, K. Vasko, V. Ollikainen, P. Sevon, H. Mannila, and J. Kere. Gene mapping by haplotype pattern mining. In *Proceedings of the IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, pages 99–108, 2000.
23. S. Zhang, and H. Zhao. Linkage disequilibrium mapping with genotype data. *Genetic Epidemiology*, 22:66–77, 2002.
24. S. Zhang, K. Zhang, J. Li, and H. Zhao. On a family-based haplotype pattern mining method for linkage disequilibrium mapping. In *Proceedings of the 7th Pacific Symposium on Biocomputing*, pages 100–111, 2002.

DRAFT
Accepted for publication in 'Data Mining in Bioinformatics'
Jason Wang, Mohammed Zaki, Hannu Toivonen, and Dennis Shasha (Eds.), Springer

Biographies

Petteri Sevon received his M.Sc. degree in computer science at the University of Helsinki, Finland, in 2000. He is currently a Ph.D. student under supervision of Prof. Hannu Toivonen. His research interests include data mining and statistical genetics. He has three years of experience in practical genetic analyses with Prof. Juha Kere's research groups at the Finnish Genome Center, Helsinki and Karolinska Institutet, Huddinge, Sweden.

Petteri has published five refereed papers on methods for genetic analysis and their applications. He holds four patent applications.

Hannu Toivonen is a professor of computer science at the University of Helsinki, Finland. He received his M.Sc. and Ph.D. degrees in computer science from the University of Helsinki in 1991 and 1996, respectively. Hannu's research interests include data mining and computational methods for data analysis, with applications in genetics, ecology, and mobile communications. Prior to his current position, he has worked for six years at Nokia Research Center.

Hannu has published over 50 refereed papers on data mining and analysis and holds over 10 patent applications. He coauthored the Best Applied Research Award paper in KDD-98, and he is ranked among the 1000 most cited computer scientists by CiteSeer. He regularily serves on the program committees of all major data mining conferences. He was a program committee co-chair for ECML/PKDD conferences in 2002, and he is a founding co-chair of the KDD workshop series Data Mining in Bioinformatics (2001–).

Päivi Onkamo is post doc researcher in Helsinki Institute for Information Technology, in the University of Helsinki, Finland. She received her M.Sc. on the topic of Molecular evolution of Artiodactyls in 1995. Later on, she has combined genetics with biometry, concentrating in the field of genetic epidemiology of complex human diseases. She received her Ph.D. in 2002 on genetic epidemiology of Type 1 diabetes.

Päivi has published 15 original articles on various aspects of genetic epidemiology. She holds 2 patent applications. As coauthor of a presentation on applying data mining methods to gene mapping, she received Award of the best presentation by graduate student in the annual meeting of International Genetic Epidemiology Society in 2000. Currently, she continues her work with application of various computer scientific methods to genetic problems in the group of Hannu Toivonen and Heikki Mannila.

DRAFT
Accepted for publication in 'Data Mining in Bioinformatics'
Jason Wang, Mohammed Zaki, Hannu Toivonen, and Dennis Shasha (Eds.), Springer

Contact information

Corresponding author: Petteri Sevon
tel. +358 9 19144429
e-mail Petteri.Sevon@cs.helsinki.fi

Hannu T.T. Toivonen
tel. +358 9 19144252
e-mail Hannu.Toivonen@cs.helsinki.fi

Päivi Onkamo
tel. +358 9 19144273
e-mail Paivi.Onkamo@cs.helsinki.fi

Address for all authors:
Department of Computer Science
P.O.Box 26 (Teollisuuskatu 23)
FIN-00014 University of Helsinki
Finland

fax +358 9 19144441

All authors are also associated with Helsinki Institute for Information Technology

DRAFT
Accepted for publication in 'Data Mining in Bioinformatics'
Jason Wang, Mohammed Zaki, Hannu Toivonen, and Dennis Shasha (Eds.), Springer