Gene Mapping by Haplotype Pattern Mining

Hannu T.T. Toivonen[1,2], Päivi Onkamo[2], Kari Vasko[3,2], Vesa Ollikainen[4],
Petteri Sevon[5], Heikki Mannila[1,6], Juha Kere[4]

[1]*Nokia Research Center;* [2]*Rolf Nevanlinna Institute, University of Helsinki;*
[3]*Center for Scientific Computing;* [4]*Finnish Genome Center, University of Helsinki;*
[5]*Department of Computer Science, University of Helsinki;* [6]*Helsinki University of Technology*[*]

Abstract

Genetic markers are being increasingly utilized in gene mapping. Discovery of associations between markers and patient phenotypes — such as a disease status — enables the identification of potential disease gene loci. The rationale is that in diseases with a reasonable genetic contribution, diseased individuals are more likely to have associated marker alleles near the disease susceptibility gene than control individuals.

We describe a new gene mapping method, Haplotype Pattern Mining (HPM), that is based on discovering recurrent marker patterns. We define a class of useful haplotype patterns in genetic case-control data, give an algorithm for finding disease-associated haplotypes, and show how to use them to identify disease susceptibility loci.

Experimental studies show that the method has good localization power in data sets with large degrees of phenocopies and with lots of missing and erroneous data. We also demonstrate how the method can be used to discover several genes simultaneously.

1. Introduction

Gene mapping is a challenging statistical and computational task. Genetic markers along chromosomes provide data that can be used to discover associations between patient phenotypes and chromosomal regions, i.e., potential disease gene loci. The growing number of available genetic markers, anticipated to reach hundreds of thousands in the next few years, offers new opportunities but also amplifies the computational complexity of the task.

Linkage disequilibrium (LD), the non-random association of marker alleles and haplotypes to the disease, has recently gained attention as a means to detect disease genes. In diseases with a reasonable genetic contribution, affected individuals are likely to have higher frequencies of associated marker alleles near the disease susceptibility (DS) gene than control individuals. LD mapping is based on comparing allele or haplotype frequencies between the affected and the control individuals and on predicting the gene locus based on the significance of the haplotype–disease association. Current methods look at either one marker at a time, or they try to analyze haplotypes with fairly involved statistical models and estimation methods.

We have developed a new LD-based approach for gene mapping, called Haplotype Pattern Mining (HPM). Instead of looking at isolated markers or sequences of consecutive markers, as most current methods do, we search for haplotype patterns consisting of nearby markers. The localization of the DS gene is then based on the set of disease-associated patterns.

Haplotype pattern mining can be implemented computationally efficiently. In this paper we complement our previous work [13] by presenting a search algorithm for haplotype patterns. We further extend the previous work by new experimental results with realistic, simulated data sets, including the discovery of several genes simultaneously.

2. Problem

Genetic markers provide an economical, sparse view of chromosomes. A marker is a short polymorphic chromosomal region for which the allele — i.e., the variant — can be tested. In disease studies, intermarker distances are typically in the order of millions of base pairs, providing savings in terms of money, time, and data, if compared to using more detailed information about the DNA.

Even sparsely located markers can be very informative. For instance, given an ancestor with a disease gene, the de-

[*]Authors' e-mails: `Hannu.TT.Toivonen@nokia.com`, `Paivi.Onkamo@rni.helsinki.fi`, `Kari.Vasko@csc.fi`, `Vesa.Ollikainen@cs.helsinki.fi`, `Petteri.Sevon@cs.helsinki.fi`, `Heikki.Mannila@nokia.com`, `Juha.Kere@helsinki.fi`

scendants that inherit the gene are also likely to inherit the alleles of nearby markers. The exact probability of inheriting any combination of markers depends, of course, on the gene location with respect to the markers, the population history or the coalescence history, and marker mutations; all of these are unknown.

The gene mapping problem we consider can be described as follows. For simplicity, we present the problem for one chromosome only.

Definition 1 A *marker map* M is a vector $(m_1, \ldots, m_k)$ of markers. Each chromosome is represented by a *haplotype vector* H over map M, where $H = (h_1, \ldots, h_k)$ and $h_i \in$ alleles(m_i) for all $i, 1 \leq i \leq k$; alleles(m_i) is the domain of the ith marker. ■

The number of alleles is typically small, up to ten for microsatellite markers and exactly two for SNP's (single nucleotide polymorphisms).

Our framework consists of a case–control setting, where the individuals have been haplotyped. Each individual contributes a chromosome pair, so the number of chromosomes is twice the number of individuals. We ignore the fact that chromosomes come in pairs and simply consider the input data as consisting of a set of disease-associated haplotypes (from the cases) and a set of control haplotypes (from the controls or pseudo-controls). Finally, we assume a DS gene to be present, based on linkage analysis, for instance. Thus the question here is not whether there is a DS gene or not in the area being studied, but where the gene is.

Definition 2 (The LD-based gene mapping problem) The input consists of a marker map M, a set $A = \{A_1, \ldots, A_p\}$ of disease-associated haplotypes over map M, and a set $C = \{C_1, \ldots, C_q\}$ of control haplotypes over map M. The task is to predict the location of the disease susceptibility gene on the map M. ■

A key to the problem is the more likely association of nearby markers to the disease status. The problem is far from trivial, however, due to the stochasticity of the coalescence process, and also due to missing information and haplotyping errors.

3. Haplotype Pattern Mining

The Haplotype Pattern Mining (HPM) method is based on the simple observation that linkage disequilibrium is likely to be strongest around the DS gene and, consequently, the gene locus is likely to be where most of the strongest associations are. The method predicts the DS locus to where strongly disease-associated haplotype patterns are. Genetically the method is model free, as there is no inheritance model for the disease. In statistical terms, the method can be described by a non-parametric model.

In this section we give a non-operational specification of the HPM method after Toivonen et al [13]. In the following section we introduce an algorithm for implementing the HPM method.

We examine linkage disequilibrium by looking for haplotype patterns that consist of a set of nearby markers, not necessarily consecutive ones.

Definition 3 Given a marker map $M = (m_1, \ldots, m_k)$, a *haplotype pattern* P on M is a vector $(p_1, \ldots, p_k)$, where $p_i \in Dom(m_i) \cup \{*\}$ for all $i, 1 \leq i \leq k$, where * is the 'don't care' symbol. The haplotype pattern P *occurs* in a given haplotype vector (chromosome) $H = (h_1, \ldots, h_k)$ if either $p_i = h_i$ or $p_i = *$ for all $i, 1 \leq i \leq k$. ■

Example 1 Consider a marker map of 10 markers. The vector

$$P_1 = (*, 2, 5, *, 3, *, *, *, *, *),$$

where $1, 2, 3, \ldots$ are marker alleles, is an example of a haplotype pattern. This pattern occurs, for instance, in a chromosome with haplotype $(4, 2, 5, 1, 3, 2, 6, 4, 5, 3)$. ■

HPM is based on recognizing disease-associated haplotype patterns that have potentially been derived from a common ancestor, by searching for haplotype patterns that are likely to be identical by descent. Gaps are allowed in the patterns to better accommodate for mutations, errors, missing data, and recombinations.

Example 2 Assume that a continuous chromosomal region including markers 2–5 is inherited from a common founder by a number of individuals, and that a marker mutation early in the coalescence history of the disease chromosome has changed the allele in marker 4 for a large fraction of current chromosomes. The haplotype shared by these individuals can be expressed as a haplotype pattern of the form of P_1 above. ■

Example 3 Assume that a continuous chromosomal region including markers 2–5 is inherited from a common diseased founder by a number of individuals. Errors in genotyping marker 4 may lead to a situation where a continuous haplotype pattern is not as significantly associated to the disease status as the one with a gap.

Errors of another type can be introduced in the construction of marker maps, by inferring a wrong order of markers. Assuming the physical order of markers 4 and 5 is actually the reverse, situations may occur where pattern P_1 is observed for continuous shared regions. ■

The gaps caused by marker mutations and errors are short, whereas missing information can span several consecutive markers, depending on how the data has been collected. Long gaps could be introduced by double recombinations, but they are rare on the genetically short distances where patterns can be observed in the first place. Since long patterns are not likely to exist, at least not in significant amounts, it can be useful for performance reasons to restrict the length of the patterns that are used in gene localization.

Definition 4 Assume a haplotype pattern $P = (p_1, \ldots, p_k)$. The *(genetic) length* of P is the maximum genetic distance (in Morgans) between any two markers m_i, m_j with $p_i \neq * \neq p_j$.

Gaps are maximal subsequences of 'don't care' symbols, excluding the tails of the pattern: a *gap* in P is a contiguous sequence $p_u, \ldots, p_v$ of alleles, where (1) $p_i = *$ for all $i, u \leq i \leq v$ (the gap consists of 'don't care' symbols), (2) $u > 1$ and $v < k$ (the gap is not at either end of the pattern), and (3) $p_{u-1} \neq *$ and $p_{v+1} \neq *$ (the gap is bounded by alleles, rather than 'don't care' symbols). The length of the gap is $v - u + 1$. ■

The HPM algorithm we present takes as parameters the maximum number and maximum length of gaps, as well as the maximum length of patterns to be considered.

We use the signed χ^2 statistic, denoted $\pm\chi^2$, to measure the marker–disease association. A signed version of the measure is used in order to discriminate disease association from control association, i.e., from protective haplotypes.

Definition 5 The *signed χ^2 measure* $\pm\chi^2(P)$ of a haplotype pattern P is the standard χ^2 measure where the sign is positive if the relative frequency of P is higher in cases than in controls, and negative otherwise. Given a positive *association threshold* x, we say that P is *strongly associated* with the disease if $\pm\chi^2(P) \geq x$.

Given the data — markers M, a set A of disease-associated haplotypes and a set C of control haplotypes on M — and an association threshold x, we denote the collection of all strongly disease-associated patterns by $\mathcal{P}$, i.e., $\mathcal{P} = \{P \text{ is a haplotype pattern on } M \mid \pm\chi^2(P) \geq x\}$. ■

(Since the variables are simple indicators — case vs. control, matches vs. does not match pattern P — the χ^2 statistic is equal to the square of the correlation coefficient between these indicator variables, multiplied by the number of chromosomes considered. The signed version is negative if the correlation coefficient is.)

If pattern parameters are specified — a maximum genetic length, a maximum number of gaps, or a maximum length for gaps — the set $\mathcal{P}$ is refined by requiring that these additional restrictions are also fulfilled by the patterns in $\mathcal{P}$.

Haplotype patterns close to the DS locus are likely to have stronger association than haplotypes further away; consequently the locus is likely to be where most of the strongest associations are.

Definition 6 The *marker frequency* $f(m_i)$ of marker m_i (with respect to M, A, C, x as above) is the number of strongly disease-associated patterns that contain marker m_i, possibly in a gap:

$$f(m_i) = |\{P = (p_1, \ldots, p_k) \in \mathcal{P} \mid \text{there exist } t \leq i \text{ and } u \geq i \text{ such that } p_t \neq * \neq p_u\}|. \quad ■$$

The idea is that each haplotype pattern roughly corresponds to a continuous chromosomal region, potentially identical by descent, where gaps allow for corruption of marker data. While markers within gaps are not used in measuring the disease association of the pattern, the whole chromosomal region of the pattern is thought to be relevant.

The marker frequency gives a score for each marker. On the condition that we assume a DS gene to be present, e.g., based on linkage analysis, we would predict the gene to be somewhere close to the markers with largest frequencies. As a point prediction we simply give the locus of the most frequent marker.

Definition 7 The HPM *point prediction* of DS gene locus is the location of the marker m_i that maximizes $f(m_i)$, or, alternatively, the *significance-based point prediction* is the location of the marker that maximizes $p(m_i)$, where $p(m_i)$ is the statistical significance of $f(m_i)$ under the null hypothesis that 'all the chromosomes are drawn from the same distribution', i.e., that there is no gene effect. ■

This prediction method does not, of course, imply that we assume the DS locus to really overlap with the marker; we simply predict at the granularity of marker density. Consequently, the optimal point predictions of our method are within one half of the intermarker distance from the true loci.

The significance-based approach is useful for several reasons: uneven marker spacing, different allele distributions of markers, unevenly distributed missing and erroneous data, and background linkage disequilibrium all can twist the observed marker frequencies from what they would be in an ideal situation.

The use of marker significances can be illustrated as follows. Consider the frequency of a fixed marker as the test statistic. If it is very unlikely that at least that large a frequency occurs by chance, then it seems likely that the DS locus is genetically close to that marker. The significance-based approach predicts the DS gene to be in the vicinity of the marker with the smallest p value. Consecutive markers are dependent, and thus a large number of mutually dependent p values are produced. This is not a problem, however, since we do not use the p values for hypothesis testing, but only for ranking markers.

4. Algorithm

The algorithmically interesting task in the HPM method is determining the set $\mathcal{P}$ of all strongly disease-associated haplotype patterns. They can be computed efficiently as follows. First, given an association threshold x, a lower bound can be derived for the frequency of strongly disease-associated haplotype patterns. Second, given such a frequency threshold, all patterns exceeding the threshold can be enumerated efficiently with a fairly straight-forward depth-first search method.

The $\pm\chi^2$ test statistic for the disease association of a pattern P is defined as follows. Given a 2×2 contingency table of the numbers of disease-associated (A) and control (C) chromosomes matching a pattern (P) or not (N), the statistic is defined as $\pm\frac{(\pi_{AP}\cdot\pi_{CN}-\pi_{AN}\cdot\pi_{CP})^2\cdot\pi}{\pi_A\cdot\pi_C\cdot\pi_P\cdot\pi_N}$, where π_{ij} is the number of chromosomes with properties i and j, π_i the number of chromosomes with property i, π the total number of chromosomes, and the sign is positive if P is more frequent in cases than in controls, otherwise it is negative. We next show how a lower bound can be computed for the pattern frequency.

Given a strongly disease-associated pattern P, we have $\pi_{AP} \geq \frac{\pi_A\cdot\pi\cdot x}{\pi_C\cdot\pi+\pi_A\cdot x}$ for π_{AP}, the frequency of P among the disease-associated chromosomes, where π_A is the number of disease-associated chromosomes, π_C the number of control chromosomes, and x a lower bound for the $\pm\chi^2(P)$ disease association.

This claim can be proofed as follows. Since the pattern P is disease-associated, we have $\pi_{AP}\cdot\pi_{CN} > \pi_{AN}\cdot\pi_{CP}$. The test statistic is positive, and it is maximized when the pattern matches no controls, i.e., $\pi_{CP} = 0$, implying all pattern matching chromosomes are disease-associated ($\pi_{AP} = \pi_P$) and no controls match the pattern ($\pi_{CN} = \pi_C$). Then $\frac{(\pi_{AP}\cdot\pi_{CN}-\pi_{AN}\cdot\pi_{CP})^2\cdot\pi}{\pi_A\cdot\pi_C\cdot\pi_P\cdot\pi_N} = \frac{(\pi_{AP}\cdot\pi_C)^2\cdot\pi}{\pi_A\cdot\pi_C\cdot\pi_{AP}\cdot(\pi-\pi_P)} = \frac{\pi_{AP}\cdot\pi_C\cdot\pi}{\pi_A\cdot(\pi-\pi_{AP})}$ and $\frac{\pi_{AP}\cdot\pi_C\cdot\pi}{\pi_A\cdot(\pi-\pi_{AP})} \geq x \Rightarrow \pi_{AP} \geq \frac{\pi_A\cdot\pi\cdot x}{\pi_C\cdot\pi+\pi_A\cdot x}$.

The same approach works for finding protective haplotypes, by considering patterns P with $\pm\chi^2(P) \leq -x$. The derivation of the lower bound for the frequency among controls is similar to the case above; a the lower bound for π_{CP} is obtained by simply swapping π_A and π_C in the above result. Obviously, both disease-associated and protective haplotypes can be found when $|\pm\chi^2(P)| \geq x$. Then the smaller of π_{AP} and π_{CP} can be used as a lower bound for π_P, making the implementation slightly simpler.

Algorithm HPM below finds all haplotype patterns whose frequency exceeds the computed lower bound, and evaluates the association measure on them. By the above result the algorithm is guaranteed to first find a superset of all strongly disease-associated patterns and then, by evaluating the strength of the association in the set, the exact set of strongly disease-associated patterns.

The algorithm works in a generate-and-test fashion. It starts from short haplotype patterns P and recursively builds and tests larger and larger patterns in the *depthFirst* procedure. The efficiency of the algorithm is based on the lower bound for the frequency of haplotype patterns, as it effectively prunes search branches that can not contain strongly disease-associated haplotype patterns.

Algorithm HPM

Input: Marker map $M = (m_1, \ldots, m_k)$, set A of disease-associate haplotypes on M, set C of control haplotypes on M, association threshold x, maximum pattern length ℓ, maximum number of gaps g, and maximum gap size s.

Output: Prediction for a DS locus: the marker with the highest marker frequency in those haplotype patterns over M that are strongly associated with the disease with respect to A, C, and x, and that adhere to ℓ, g, and s.

Method:

$\mathcal{P} := \emptyset$ // Set of strongly associated haplotype patterns
π_A := number of disease-associated chromosomes
π_C := number of control chromosomes
$\pi := \pi_A + \pi_C$
$lb := \pi_A * \pi * x/(\pi_C * \pi + \pi_A * x)$ // A lower bound
$P = (p_1, \ldots, p_k) := (*, \ldots, *)$ // Variable for patterns
for $i := 1$ to k **do**
    for each $a \in$ alleles(m_i) **do**
        $p_i := a$
        depthFirst$(P, i, i, 0, 0)$ // Test P and its extensions
        $p_i := *$ // Reset p_i
for $i := 1$ to k **do**
    compute marker frequency $f(m_i)$ in the set $\mathcal{P}$ of patterns
output the marker which has the highest marker frequency

procedure depthFirst(P, start, i, nr-of-gaps, gap-length)
// Test haplotype pattern P and all patterns that can be
// generated by extending P from the right.
if $\pm\chi^2(P) \geq x$ **and** $p_i \neq *$ **then**
    insert P to the set $\mathcal{P}$ // Strongly associated pattern!
if $i = k$ **or** genetic-distance$(m_{start}, m_{i+1}) > \ell$ **then**
    return // Extended patterns would be too long
if frequency of P in disease-associated chromosomes $< lb$ **then**
    return // Extended patterns can't be disease-associated
// Create and test legal extensions of current pattern P:
for each $a \in$ alleles(m_{i+1}) **do**
    $p_{i+1} := a$ // Give marker $i+1$ all possible values
    depthFirst(P, start, $i+1$, nr-of-gaps, 0)
if $p_i \neq *$ **and** nr-of-gaps $< g$ **and** $s \geq 1$ **then**
    $p_{i+1} := *$ // Introduce a new gap starting at marker $i+1$
    depthFirst(P, start, $i+1$, nr-of-gaps $+ 1, 1$)
if $p_i = *$ **and** gap-length $< s$ **then**
    $p_{i+1} := *$ // Extend the current gap over marker $i+1$
    depthFirst(P, start, $i+1$, nr-of-gaps, gap-length $+ 1$)
$p_{i+1} := *$ // Reset p_{i+1}
return

Missing values are handled (not shown in the algorithm) by ignoring, for each pattern individually, those chromosomes that have missing data in the 'active' (non-*) part of the pattern. The effect is that the number of control and disease-associated chromosomes can be different for each pattern.

The HPM method can be easily augmented to produce significance-based point predictions, an extremely important utility in reducing the effects of uneven marker spacing and background linkage disequilibrium. For reasons of

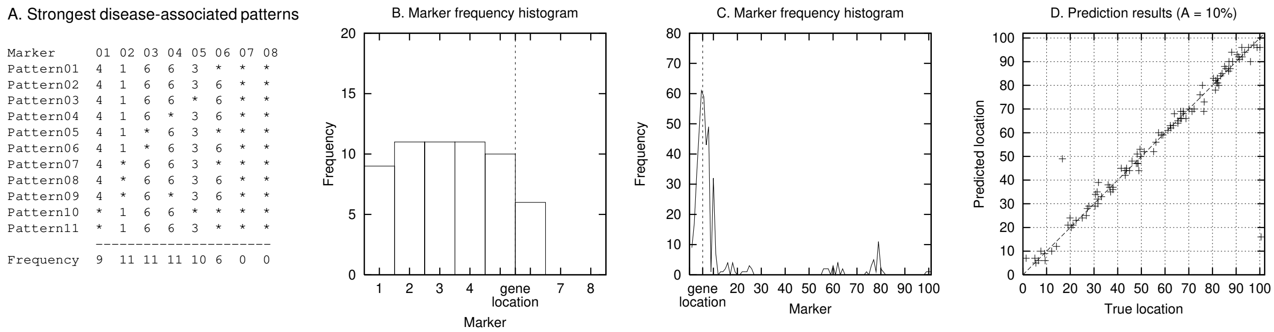

Figure 1. Panel A: Examples of strongly disease-associated haplotype patterns discovered in a data set with $A = 10\%$. Panel B: The corresponding marker frequencies; only the first 8 markers out of 101 are shown. ('Don't care' symbols * in gaps are included in marker frequencies.) Panel C: Marker frequencies in the same data set with all strongly disease-associated haplotype patterns. Panel D: A plot of the true vs. predicted locations of the DS gene for 100 data sets ($A = 10\%$).

brevity we only outline the extensions. The null hypothesis is that 'all the chromosomes are drawn from the same distribution'. The situation under the null hypothesis can be experimentally tested by randomly permuting the status fields of the chromosomes, keeping the proportions of affected and control chromosomes constant, and computing the marker frequencies in these permutations. An empirical p value is obtained for a marker from the fraction of permutations that gives at least at large frequencies for the marker than what was observed in the original data.

5. Experiments

We have evaluated the HPM method with realistic, simulated data sets that correspond to a recently founded, relatively isolated founder subpopulation. See Toivonen et al [13] for details of the simulation procedure and more experimental results, including a real data set of sib-pair families with type 1 diabetes from the UK.

In all the tests we simulated isolated founder populations which grow from 300 to approximately 100,000 individuals during 500 years. The genetic length of the chromosomes was 100 cM. Microsatellite markers were spaced at intervals of 1 cM, and their information content (PIC) was fixed at 0.7. We used a dominant disease model with a high phenocopy rate; the proportion of mutation-carrying chromosomes in the affected sample, denoted by A, was either 2.5%, 5%, 7.5% or 10%. The sample size was 200 individuals, out of which 100 were controls. High phenocopy rates and small sample sized were used, in order to explore the limits of the method and to study the performance in realistic situations.

As a basic setting within which to analyze the performance of the method, we used the following parameter values. The maximum length of haplotype patterns was restricted to seven consecutive markers, i.e., 6–8 cM. This is close to the average length of shared haplotypes in a population of approximately 500 years of age. To allow for reasonable flexibility, at most two gaps were allowed per haplotype, and their lengths were limited to one marker. These parameter values prune patterns which are not biologically conceivable (unreasonably long haplotype patterns, or those consisting mainly of gaps) and, from a practical point of view, they allow faster execution and experimenting with the method than more flexible parameters. With these parameter values, localization time for one simulated data set on a 400 MHz Pentium PC was around one minute. The association threshold for the signed $\pm\chi^2$ measure was set to $x = 9$, based on earlier work on similar data and methods (Ollikainen V, Mannila H, Kilpikari R, Koivisto M, Kärkkäinen H, Mäkelä M, Onkamo P, Smolander S, Kere J, unpublished data) and some experimenting. To ensure that the selection of these particular values is not critical for the method and to assess the robustness of HPM in this respect, we also experimented using patterns with unlimited length, with longer gaps, and without gaps.

5.1. Localization accuracy for one gene

In order to illustrate the HPM method, Figure 1A shows the list of 11 most strongly disease-associated haplotype patterns in a simulated data set with $A = 10\%$ (10% of disease-associated chromosomes carry the mutation; no missing or corrupted data). The chromosome has 101 markers, but the patterns with strongest association occur between markers 1 and 6. The bottom line gives the marker frequencies for these markers, and the frequencies are also plotted as a histogram in Figure 1B. Markers 2–4 have the

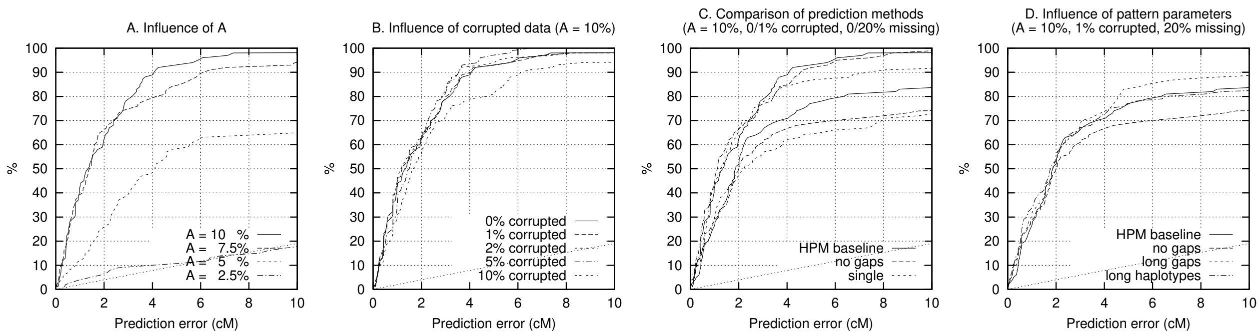

Figure 2. Effects of various parameters on the prediction accuracy. The y-axis shows which fraction of simulated data sets is within the error bound given on the x-axis, i.e., $y = P(\text{error} \leq x)$. The lowest, dotted curve is the prediction accuracy of random, uniform guesses. Panel A: The effect of the proportion of DS mutation carrying chromosomes "A". Panel B: The effect of corrupted data. Panel C: Comparison of prediction methods. The three topmost curves have been obtained with 0% corrupted and 0% missing data; the lower curves with 1% corrupted and 20% missing data. Panel D: The effect of pattern search parameters. 'HPM baseline': haplotype pattern searching as before, 'no gaps': haplotype pattern searching without gaps, 'single': the middle point of single most strongly associated haplotype without gaps is used for predicting the localization, 'long gaps': gaps of up to three markers allowed, 'long haplotypes': no length limit on the pattern lengths.

highest frequency, closely followed by markers 5 and 1. The true gene location is in this data set halfway between markers 5 and 6 (depicted by a dashed vertical line). Figure 1C shows a frequency histogram for the same data set, but this time with all haplotype patterns exceeding the association threshold of 9. Marker 5 has now the highest frequency and is therefore predicted as the gene location; a vertical line shows, again, the true location at position 5.5.

The true vs. predicted locations for 100 simulated data sets with $A = 10\%$ are shown in Figure 1D; the data set of Figure 1C is represented by a cross at $(5.5, 5)$. Overall, the predicted location shows good agreement with the true location. The localization accuracy and the effect of phenocopies was explored in more detail by plotting curves similar to power graphs: the height of the curve shows the fraction of data sets for which the localization was successful, as a function of the allowed localization error (Figure 2A). The solid line represents the results given by Figure 1D: for instance, in 90% of the simulations the error is 4 cM or less. For $A = 7.5\%$ the accuracy is near that for $A = 10\%$, but for $A = 5\%$ a clear drop can be observed and for $A = 2.5\%$ the localization method does not perform significantly better than random guessing. Our explicit aim was to test realistic (small sample sizes) but difficult $(2.5\% \leq A \leq 10\%)$ cases, in order to explore the limits of the method — which in this case and with respect to A seem to be somewhere around $A = 5\%$. For larger samples and lower phenocopy rates the results should obviously be at least as good as those presented here.

The effect of corrupted data, i.e., genotyping errors and sporadic marker mutations, was tested by randomly changing alleles in the data. Figure 2B shows the influence of having up to 10% of data corrupted (with $A = 10\%$). The influence of having up to 20% missing data was explored in a similar manner. There is hardly any difference in the accuracy with 0–5% of data corrupted or missing. Higher proportion (10% or more) results in a slight decrease in performance. The combined effect of corrupted and missing data contained no surprising interactions.

The HPM method was compared to two simpler alternatives (Figure 2C, $A = 10\%$). The first one was to take the single most strongly associated haplotype without gaps, and to predict the DS locus to be in the middle of that haplotype. The second was to localize with haplotype patterns without gaps. With correct data (three higher curves), there is not much difference between the performance of the methods for error bounds less than 4 cM. More differences appear as corrupted and missing data are introduced (lower three curves), and HPM method seems to outperform the other methods by finding the approximately correct region more consistently.

In order to assess the robustness of the method with respect to the selection of pattern search parameters, simulated data with $A = 10\%$, 1% corrupted and 20% missing, was reanalyzed (Figure 2D). The effect of gaps in the patterns was evaluated by either prohibiting gaps (as in Figure 2B) or by allowing the gaps to be up to three markers long instead of just one. In addition, a test was run where

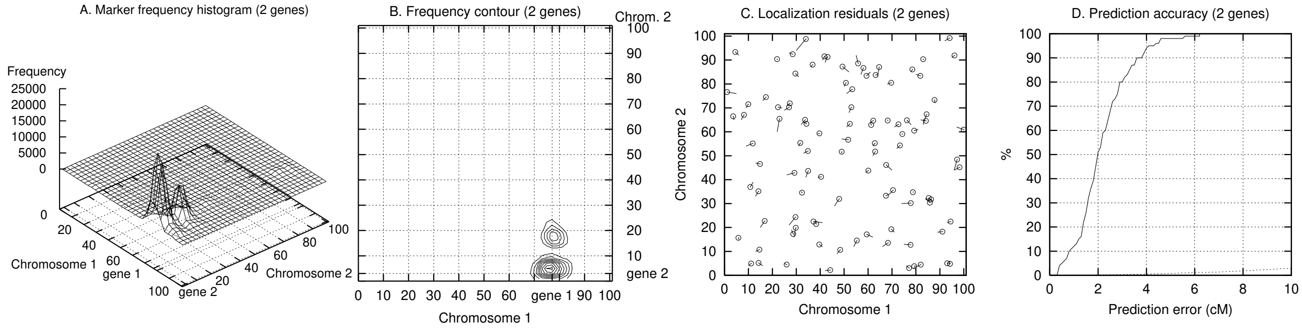

Figure 3. Panel A: Frequency histogram for pairs of markers from two chromosomes in three-dimensional form. Panel B: The same in contour representation (contour line interval is 2,500). The true locations of the genes are 77.1 on chromosome 1 and 3.0 on chromosome 2. Panel C: Prediction accuracy achieved in two-gene experiments; vectors represent the (Euclidean) residuals from the true locations (centers of the circles). Panel D: The cumulative proportion of analyses (in 100 data sets) with residuals below the bound given by x-axis. The dotted curve on the bottom is the prediction accuracy of random, uniform guesses for two gene locations.

the length of the haplotype patterns was not limited. Differences start to appear at error bounds of at least 2 to 4 cM: allowing longer gaps improves the performance somewhat, whereas prohibiting gaps altogether results in a decreased performance.

Permutation tests were used to obtain significance-based point predictions. The experimental results obtained with 1,000 random permutations show that the peaks observed in marker frequencies in the vicinity of DS locus typically clearly surpass those produced by background LD. However, the agreement of frequency-based and significance-based predictions is partly due to the "regularity" of the simulated data, in particular the even marker spacing. With real data the significance-based predictions are much more reliable; see Toivonen et al [13] for an example.

Our other experiments with biallelic SNP markers have shown that in the simulated data sets the localization accuracy of HPM with 3 SNPs per 1 cM is close to that of a map with one microsatellite per 1 cM [13]. Experiments with real data from UK, consisting of 25 microsatellite markers covering a 14 Mb region and containing the major Type 1 diabetes susceptibility locus mapped to the HLA region demonstrate that the method works with real data from a non-isolated population [13].

5.2. Localization accuracy for two-gene diseases with strong allelic heterogeneity

Mapping of oligogenic diseases is an interesting problem for which the HPM method can be generalized. We performed preliminary experiments on simulated data sets that contained two DS loci in different chromosomes. Both DS loci included several disease mutations, each originating from a different ancestral haplotype (max. 5 ancestral mutations per locus). The effect of all disease mutations to the disease susceptibility was, for simplicity, the same. The combined frequency of mutation alleles in current population was 0.05 for both loci. The overall frequency of carriers of at least one mutation allele in each disease locus (i.e. A–B–) was approximately 9.5/1,000 individuals. The penetrances were defined to be 0.9 for individuals with at least one DS mutation in both chromosomes (A–B–), and 0.001 for individuals with the remaining genotypes. The samples consisted of 100 affected and 100 randomly selected control individuals. Average genotype frequencies are summarized in Table 1. Table 2 shows the frequencies of the most frequent genotype in each genotype class, when different mutation alleles are not lumped together.

We carried out the two-gene experiments as follows. First, all the frequent haplotype patterns in the two chromosomes $C1$ and $C2$ were searched with the basic HPM algorithm, without considering the strength of their disease association. Then, new 'two-gene' patterns were formed by taking all pairwise combinations between patterns on $C1$ and on $C2$. We evaluated the association measure for these patterns, and predicted the DS locations for both chromosomes simultaneously based on the marker scoring method and the two-gene patterns.

It turned out that marker frequency surfaces generated by the strong two-gene haplotype patterns in the current data sets tended to have one strong peak, always in the immediate neighborhood of the DS gene (see Figure 3 for an example). Experimental results with 100 data sets indicate that the localization accuracy is consistently high: Figure 3C

Table 1. Absolute genotype frequencies in two-gene data sets of 100 affected and 100 control individuals (mean and estimated standard deviation). Because of etiological heterogeneity incorporated in the simulations, each mutation, denoted by a capital letter A or B, represents 1 to 5 (on average 3.2) founder mutations that are typically associated to different alleles. In each data set the most frequent mutation represents, on average, 43% of the mutations.

In affected	BB	Bb	bb	In controls	BB	Bb	bb
AA	0.1 (0.4)	2.3 (1.5)	0.0 (0.0)	AA	0.0 (0.0)	0.0 (0.0)	0.2 (0.5)
Aa	2.2 (1.4)	85.1 (3.4)	1.1 (1.0)	Aa	0.0 (0.0)	0.1 (0.3)	8.6 (3.0)
aa	0.0 (0.1)	0.9 (1.0)	8.3 (2.5)	aa	0.2 (0.5)	8.8 (2.9)	82.1 (4.0)

Table 2. Absolute frequencies for the most frequent combination of founder mutations in different genotype classes (mean and estimated standard deviation) computed from 100 data sets of 100 affected and 100 healthy individuals. Heterozygotes (Aa and Bb) carry the same copy of a founder mutation, whilst homozygotes (AA and BB) carry the same pair of founder mutations. For instance, the mean frequency of 0.7 for genotype class AaBB in the affected sample indicates that in average 0.7 affected individuals were heterozygous with respect to the same founder mutation in the first DS gene and, in addition, carry the same pair of founder mutations in the second DS gene.

In affected	BB	Bb	bb	In controls	BB	Bb	bb
AA	0.1 (0.3)	1.2 (0.8)	0.1 (0.2)	AA	0.0 (0.0)	0.0 (0.0)	0.2 (0.5)
Aa	1.1 (0.7)	29.7 (12.7)	0.8 (0.8)	Aa	0.0 (0.0)	0.1 (0.3)	5.7 (2.6)
aa	0.0 (0.2)	0.7 (0.7)	8.4 (2.7)	aa	0.2 (0.4)	5.4 (2.2)	82.1 (4.0)

shows the 2-dimensional residuals, Figure 3D gives the corresponding curve for prediction accuracy (in terms of the length of the residual vector).

6. Discussion

We have given an algorithm for Haplotype Pattern Mining (HPM), a new approach to LD mapping, and we presented experimental results with realistic, simulated data.

The HPM algorithm is based on finding all frequent patterns in the given data. The task is essentially the same as in the data mining problem of discovering all association rules [1, 2]. However, association rule algorithms have been designed to trade off low disk IO against high CPU costs, whereas here the data can be stored in main memory and all the computational costs are CPU-bound. In this respect, there is an apparent similarity to main memory methods used to discover patterns in protein sequences (e.g., [5]).

The method can be seen as an algorithm-based extension to traditional genetic association analysis. It works with a non-parametric statistical model and without any genetic models. Several statistical methods have been recently introduced to detect LD [12, 4, 7, 11, 10]. They put an emphasis on statistical models of LD around a disease susceptibility (DS) gene, rather than on patterns as we do. For instance, they try to take into account the recombination history from the common ancestors until present day, and some of the models are robust to high levels of etiologic heterogeneity. On the other hand, these methods make a number of assumptions about the inheritance model of the disease and the structure of the survey population, which may be misleading for the statistical inference. In addition, the methods can only consider association of one region at a time. Thus, they are currently best suited for fine mapping, rather than complex disease mapping or genome screening. The methods also tend to be computationally heavy.

Permutation tests similar to our approach have been used in mapping studies by Churchill and Doerge [3], Laitinen et al. [6], and Long et al. [9].

The HPM method seems to offer good possibilities for fine mapping, for example in proving DS loci when multiple candidate genes are screened with a dense map of markers after finding initial positive linkage. Our results indicate that the proposed method is robust to missing and erroneous data. It is suitable for application in complex disease mapping with high etiologic (allelic and locus) heterogeneity, as it can cope with a high degree of phenocopies. The power is good even with small data sets (100 affected and 100 control individuals).

Even though the simulation results are for a population isolate, the method itself does not require the study population to be an isolate, as experiments with the real data have shown [13]. At least moderate LD should exist between consecutive markers near the disease locus or loci, which means that the average marker distances in the region to be studied should be small with respect to the amount of LD in the study population. For a population isolate with expansive growth through approximately 20 generations, as in our simulations, the average microsatellite marker distance of 1 cM throughout the area of interest gave good results. For a typical linkage result, the area to fine map is 20–30 cM, which would in a similar setting require the genotyping of 20–30 markers. To find out which kind of average marker spacing is needed in different study populations, the power of the method should be explicitly studied in different situations. A crude estimate might be obtained by relating $4Nc$, the measure for average LD according to Long and Langley [8], to the power of the localization.

Improved techniques for high-throughput identification and genotyping of polymorphisms, such as SNPs, offer the possibility of using high numbers of markers in genome screening and candidate gene scanning in the near future. The proposed method scales well to large data sets of biallelic and multiallelic markers.

There are several ways to generalize the HPM method. First, we can allow approximate or inexact matching of patterns: the chromosomes that almost match a pattern could also be counted, perhaps with a reduced weight. Such an approach could provide a still more robust handling of errors and missing data than having gaps in fixed locations. Another possibility for relaxing pattern matching concerns uncertainties in haplotyping, as there is often some uncertainty concerning the alleles. The estimated probabilities of different alleles could be accounted for by weighting alternative true haplotypes according to their probabilities.

Improved flexibility can also be achieved by a more expressive pattern language. For instance, each marker could be associated with a set of alleles; one of them must appear in a chromosome. This kind of modification could be useful for the allele heterogeneity caused by marker mutations.

Interesting new possibilities are opened by incorporating environmental factors and phenotypic characteristics into the data. Conceptually the extensions to the pattern formalism and to the algorithms are simple. The input consists of vectors containing haplotype, environmental, and phenotypic data; the haplotype patterns can contain elements from all of these. Association of any such combination to the disease status, or to a set of phenotypic indicators, can be measured. For discrete factors this is almost straightforward, continuous measurements could be discretized as a pre-processing step.

Finally, methods for pattern discovery might be used as a pre-processing step for more detailed explicit statistical analysis. For example, the haplotype patterns might be used as a sample space for the reconstruction of ancestral haplotypes in DS chromosomes.

More information, including the simulated data sets, is publically available at
`http://www.genome.helsinki.fi/eng/research/projects/DM/index.html`.

Acknowledgements We are thankful to Mikko Koivisto for useful comments on a manuscript of this paper.

This research has been funded by the Academy of Finland, Graduate School in Computational Biology, Bioinformatics, and Biometry (ComBi), and Helsinki Graduate School in Computer Science and Engineering (HeCSE).

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *Proceedings of 1993 ACM SIGMOD Conference on Management of Data*, pages 207–216. ACM, Washington, DC, May 1993.

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo. Fast discovery of association rules. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI Press, Menlo Park, CA, 1996.

[3] G. Churchill and R. Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 138:963–971, 1994.

[4] B. Devlin, N. Risch, and K. Roeder. Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics*, 36:1–16, 1996.

[5] I. Jonassen, J. Collins, and D. Higgins. Finding flexible patterns in unaligned protein sequences. *Protein Science*, 4(8):1587–1595, 1995.

[6] T. Laitinen, P. Kauppi, J. Ignatius, T. Ruotsalainen, M. Daly, H. Kääriäinen, L. Kruglyak, and et al. Genetic control of serum IgE levels and asthma: linkage and linkage disequilibrium studies in an isolated population. *Hum Mol Genet*, 6:2069–2076, 1997.

[7] L. Lazzeroni. Linkage disequilibrium and gene mapping: an empirical least-squares approach. *Am J Hum Genet*, 62:159–170, 1998.

[8] A. Long and C. Langley. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res*, 9:720–731, 1999.

[9] A. Long, R. Lyman, C. Langley, and T. Mackay. Two sites in the Delta gene region contribute to naturally occurring variation in bristle number in Drosophila melanogaster. *Genetics*, 149:999–1017, 1998.

[10] M. McPeek and A. Strahs. Asssessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am J Hum Genet*, 65:858–875, 1999.

[11] S. Service, D. Temple Lang, N. Freimer, and L. Sandkuijl. Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *Am J Hum Genet*, 64:1728–1738, 1999.

[12] J. Terwilliger. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one ore more polymorphic marker loci. *Am J Hum Genet*, 56:777–787, 1995.

[13] H. Toivonen, P. Onkamo, K. Vasko, V. Ollikainen, P. Sevon, H. Mannila, M. Herr, and J. Kere. Data mining applied to linkage disequilibrium mapping. *Am J Hum Genet*, 67:133–145, 2000.