Haplotype Association Mapping by Density-Based Clustering in Case-Control Studies (Work-in-Progress)

Jing Li[1] and Tao Jiang[1,2]

[1] Department of Computer Science and Engineering, University of California
Riverside, CA 92521
{jili,jiang}@cs.ucr.edu
[2] Shanghai Center for Bioinformatics Technology

Abstract. Linkage disequilibrium (LD) mapping for complex diseases using haplotypes has been intensively studied recently due to increased availability of large-scale dense SNP (*single nucleotide polymorphism*) markers. Such an LD mapping has many applications, *e.g.* finding disease-associated haplotypes and predicting *disease susceptibility* (DS) gene loci from a whole genome scan. In this research, we develop a new algorithmic method for haplotype mapping based on a density-based clustering algorithm, and propose a new haplotype (dis)similarity measure. The mapping regards haplotype segments as data points in a high dimensional space. The DS gene embedded haplotypes, especially those mutants of recent origin, tend to be close to each other due to linkage disequilibrium, while other haplotypes can be regarded as random noise sampled from the haplotype space. Clusters are then identified using a density-based clustering algorithm. Pearson χ^2 statistic or Z-score based on the numbers of cases and controls in a cluster can be used as an indicator of the degree of association between the cluster and the disease under study. The method does not require any assumptions about the evolutionary model or the inheritance patterns of the disease. The proposed similarity measure is a generalization of several haplotype similarity measures currently used in the literature. It is robust against recent mutations/genotype errors and recombination events. Preliminary experimental results on an independent simulated data set, including both SNP markers and microsatellite markers, and on a real data set with the known DS gene location for type 1 diabetes show that our method could predict gene locations with high accuracy, even when the rate of phenocopies is high. This work is still in progress and more data are going to be tested.

Keywords: Haplotype, LD association mapping, SNP marker, clustering algorithm, case-control studies

1 Introduction

With the completion of the Human Genome Project [4, 12], an (almost) complete human genomic DNA sequence has become available, which is essential to the understanding of the functions and characteristics of human genetic material. An important next step in human genomics is to determine genetic variations among humans and the correlation between genetic variations and phenotypic variations (such as disease status, quantitative traits, *etc.*). To achieve this goal, an international collaboration, namely, the international HapMap project (`http://www.hapmap.org`), was launched in October, 2002. The main objective of the HapMap project is to determine the haplotype structure of humans, in the hope that we can identify associations of common haplotypes and common diseases by LD mapping using case-control data in the near future.

The rationale behind haplotype LD mapping is that when a disease mutated allele is in tight linkage disequilibrium with alleles surrounding them, haplotypes from (or transmitted to) affected individuals are expected to be more similar then those haplotypes from unaffected individuals (or untransmitted haplotypes). Various statistical methods have been proposed based on the degree of haplotype sharing in affected individuals (for example [8, 11] among others). While haplotype based methods have shown higher power and higher accuracy than traditional LD mapping using individual markers, most of the methods need explicit assumptions on disease inheritance patterns and/or the evolutionary models of the population under study, which are usually unknown in practice. The effects of violations of these assumptions are unpredictable in general. Recently, a nonparametric method called HPM (*haplotype pattern mining*) [10], inspired by data mining methods, has been proposed to identify disease associated haplotype patterns from case-control data. Toivonen *et al*. [10] showed that HPM does not require any assumptions on the inheritance patterns and has good localization power, even when the number of phenocopies is large. Later on, the method was extended to family data [13] and to QTL (*quantitative trait loci*) mapping [9]. However, methods based on HPM also have some limitations. First, by allowing “don’t care” symbols in a haplotype pattern, many haplotypes have been counted multiple times. The effect of this duplicate counting is unknown. Secondly, the frequency of identified haplotype patterns is used to predict gene locations. But the value is closely related to the sample size, and the statistical significance of the predicted gene location and obtained patterns cannot be assessed. Thirdly, the results in [10] showed that the effect of using a permutation test for HPM to predict gene locations is inconclusive. Finally, the authors of [10] showed in their experimental results that the prediction accuracy may be worse with dense (SNP) markers, which is undesirable and greatly limits the utility of the method.

In this study, we introduce a new haplotype mapping method based on a density-based clustering algorithm [1], which also does not require any assumptions on the evolutionary model or the inheritance patterns of the disease. The method works as follows. Haplotypes across a set of markers are first cast to a high dimensional discrete space. Clusters of haplotypes are then identified based on a new (dis)similarity measure via a density-based clustering algorithm. Our similarity measure, to be defined in section 2, generalizes several similarity measures in the literature and combines both Hamming similarity and the longest common substring. It is very flexible and robust against recent mutations/genotype errors and recombination events. Notice that the framework of our method is independent of the choice of haplotype similarity measure. The cluster with the most haplotypes from affected individuals is more likely to contain haplotypes with DS genes. By sliding a window of haplotype segments of certain length, a whole genome scan for association can be performed given because our method has a good scalability. The overall time complexity of our algorithm is $O(MN^2)$, where M is the total number of marker loci and N is the sample size which is around hundreds in most real data sets. The Pearson χ^2 statistic or Z-score (which are equivalent [2]) based on a contingency table derived from the numbers of case haplotypes and control haplotypes in a cluster can be used as an indicator of the degree of association of the cluster and the disease. Both measures can also be used as association/independence test statistics, properly adjusted (for example, using Bonferroni correction) for multiple tests. A statistical significance threshold can be chosen independent of the sample size and all findings that exceed the threshold will be reported. In the current algorithm, we only use the Z-score as an indicator of the degree of association. To truly reveal the performance of our method, we use the simulated data set from [10] and some real data sets. Preliminary experimental results on that simulated data and on the HLA real data set [3] with known disease gene location for type 1 diabetes show that our method could predict the gene location with high accuracy and its performance increases with denser markers. More benchmark tests will be performed in the future.

2 The Method and Similarity Measure

Our method consists of steps. We first define a general similarity measure. A corresponding distance measure can be defined based on the similarity measure in a straightforward way. Although the similarity is defined between two haplotypes, it is marker specific in the sense that we measure the similarity between two haplotypes relative to a specific marker. This is useful for whole genome association analysis because each marker position can have its own similarity score. The meaning will be more clear after we give the definition below. Clusters are then identified using a density-based clustering algorithm based on pairwise haplotype distance. There is no need to distinguish case and control haplotypes in these two steps. The labels of haplotypes are only used in the third step in assessing the degree of association of obtained clusters and the cluster with the largest Z-score is selected for each marker locus. For simplicity, we state the procedure by focusing on one locus and its surrounding markers. The procedure can be extended for a whole genome scan by sequentially examining each marker. At last, we draw a graph of the Z-score of each marker along a chromosome map. The true gene location(s) would be the one with the biggest score (or those with scores larger than a predefined threshold). A consensus haplotype pattern or a haplotype profile based on the cluster with the highest score will be reported as the disease associated pattern.

2.1 A general haplotype (dis)similarity measure

Suppose that we focus on a marker at locus 0, with loci $1, 2, \ldots, r$ on one side and $-1, -2, \ldots, -l$ on the other side. Assume that the genetic/physical distance from any locus to locus 0 is known and denoted as x_k, where $-l \leq k \leq r$. A haplotype h

spanning this region is just an $l+1+r$-dimensional vector and the k^{th} dimension of h, denoted as $h(k)$, is the allele at locus k. For a pair of haplotypes h_i, h_j , we define the similarity score of h_i, h_j (focusing on locus 0) as:

$$s_{i,j} = \sum_{k=-l}^{k=r} w_1(x_k) I(h_i(k), h_j(k)) + \sum_{k=1}^{k=r'} w_2(x_k) + \sum_{k=-1}^{k=-l'} w_2(x_k), \qquad (1)$$

where I is the identity function, $-l'$ and r' are two boundary loci such that the two haplotypes h_i, h_j are identical between these two loci and different at both locus $-l'-1$ and locus $r'+1$. The weights w_1 and w_2 are two decreasing functions so that the measure on each locus is weighted according to the distance from locus 0. The choices of the weights w_1 and w_2 will be discussed shortly.

The first summation in Equation 1 is a weighted measure of the number of alleles in common between haplotype h_i and h_j in the region, which can be thought of Hamming similarity. The remaining summations form a weighted measure of the longest continuous interval of matching alleles around locus 0, which has some resemblance to the longest common substring. This definition is quite flexible and generalize several similarity measures used in the literature [11]. For instance, by setting $w_1 = 1$ and $w_2 = 0$, the measure becomes the counting measure described in [11]. The length measure in the same article can be achieved by setting $w_1 = 0$ and $w_2 = 1$. This definition of haplotype similarity is more powerful than the above two specialized measures from [11] and can be used for different types of markers by choosing appropriate weighting functions. It has the strengths of both specialized measures: it is robust against recent marker mutations and genotyping/haplotyping errors, and it also apprehends partial sharing from a common ancestral haplotype due to historical recombination events. Notice that $s_{i,i} = s_{j,j}$, a distance metric between haplotypes h_i and h_j at marker locus 0 can be defined as:

$$d_{i,j} = \frac{s_{i,i} - s_{i,j}}{s_{i,i}} = \frac{s_{j,j} - s_{i,j}}{s_{j,j}}. \qquad (2)$$

The distance is normalized to the interval $[0,1]$ so it will not increase with the length of haplotypes.

The requirement for both weighting functions w_1 and w_2 is that they must be decreasing functions. It can be exponentially decreasing, quadratically decreasing or linearly decreasing. It can also be a discrete function and the values are only defined at marker positions. The user has the freedom of choosing the weighting function depending on the marker density of the input data. The selection of w_1 and w_2 in our simulation is depicted in Figure 1. More work needs to be done on how to choose these functions depending on genetic/physical distances of the input data.

2.2 A density-based clustering algorithm

As a general tool of mining useful information from massive data, clustering algorithms have been widely used in many fields including computational biology, especially in microarray data analysis. But as far as we know, no such methods have been used in gene mapping. In the haplotype association mapping setup, we are interested in identifying haplotype clusters that are strongly associated with the disease under study. This is unlike traditional clustering tasks that try to partition all the data points into certain clusters or try to build a hierarchical cluster tree. Under the assumption that the control haplotypes are randomly sampled, we do not expect them to form any clusters except by chance. On the other hand, haplotypes from affected individuals are expected to be more similar. But the difficulty lies in the fact that due to the phenomenon of phenocopy (*i.e.* people who were diagnosed with certain disease do not actually have any genetic material related to the disease), some haplotypes from affected individuals do not necessarily form a cluster. This is also a main reason why a gene mapping method using case-control data would likely fail in reality if it assumes, explicitly or implicitly, that all or at least most affected individuals do have disease-related genetic material. The key idea of our method is that we take the problem of finding strongly disease associated haplotype clusters as the problem of finding clusters from data with noise background. We use the concept of "density-based clusters" and adopt an algorithm called DBSCAN [1] with minor modifications. In order to keep the paper self-contained, we briefly introduce the DBSCAN algorithm in the context of haplotype mapping.

Before presenting the DBSCAN algorithm, we need some definitions. There are two input parameters for DBSCAN. One is a radius of the interested neighborhood ϵ and the other is a density threshold *MinPts*. A haplotype is called a *core* haplotype if there are more than *MinPts* haplotypes in its ϵ neighborhood. The haplotypes in the ϵ neighborhood are *directly reachable* from the core haplotype and a haplotype is *reachable* from a core haplotype if there is a chain of core haplotypes between these two haplotypes where each is directly reachable from the preceding one. Two haplotypes are *density-connected* if there is a core haplotype such that both haplotypes are reachable from it. A *density-based cluster* of haplotypes is a set of density-connected haplotypes with maximal density-reachability. All the above definitions are with respect to the two parameters ϵ and *MinPts*. DBSCAN examines every haplotype and starts to construct a cluster once a core haplotype is found. It then iteratively collects directly reachable haplotypes from a core haplotype, merging clusters when necessary. The process terminates when all haplotypes have been examined. The clusters are then output and the haplotypes that do not belong to any cluster are regarded as noise. More details about the algorithm can be found in [1].

2.3 Score of the degree of association

We measure the degree of association between a haplotype cluster and the disease of interest using the Z-score. Suppose that we are given m case haplotypes and n control haplotypes. Let m' and n' denote the number of case and control haplotypes in a cluster, respectively. A 2×2 contingency table like Table 1 can be constructed. The Z-score is defined as:

$$Z = \frac{m'/m - n'/n}{\sqrt{\frac{m'+n'}{m+n}(1 - \frac{m'+n'}{m+n})(1/m + 1/n)}}. \tag{3}$$

It is the weighted difference of relative frequencies of the case and control haplotypes in a cluster and follows approximately a normal distribution if we assume haplotypes randomly occur in the cluster. A large Z-score means strong association of the cluster (the haplotypes within the cluster) and the disease. The cluster with the highest score is

Table 1. A contingency table built for a cluster C.

	Case	Control
Cluster C	m'	n'
Remaining	$m - m'$	$n - n'$

taken as the prediction for each marker. The score is regarded as the point estimation of each marker locus and a consensus haplotype pattern or a haplotype profile based on the cluster can be used as diseased associated pattern centered at the current locus.

3 Preliminary experimental results

3.1 Simulation studies

To evaluate the performance of the proposed method, we perform simulation studies using the same data sets generated by Toivonen *et al.* [10] in their studies of the HPM method. We take the independent simulated data because our method could be applied to any population model or disease inheritance pattern, and the results obtained would truly reveal the performance of our method.

The data sets correspond to a recently founded, relatively isolated founder subpopulation that grows from the initial size of 300 to about 100,000 individuals in 500 years. A pair of homologous chromosomes are simulated for each individual with genetic length of 100 cM. Both microsatellite markers and SNP markers were simulated. Markers are evenly spaced along the chromosome with interval lengths of 1 cM and 1/3 cM for microsatellite marker and SNP marker respectively. The *polymorphism information content* (PIC) is set to 0.7 for microsatellite marker while allele frequency is set to 0.5 for SNP marker, and the PIC is thus fixed at 0.4375. A dominant disease is modeled. A small sample size with 200 control chromosomes and 200 case chromosomes are selected in order to study the performance of the method in a realistic situation. A high rate of phenocopy is used. The proportion of mutation-carrying chromosomes, denoted by A, is either 2.5%, 5.0%, 7.5%, or 10.0%, corresponding to overall relative risks of $\lambda = 1.2, 1.7, 2.7, 4.1$, respectively. Mutations are not modeled directly but compensated by introducing missing alleles randomly. A detailed description of simulation procedure can be found in the paper of Toivonen *et al.* [10].

The weighting functions w_1 and w_2 in the calculation of haplotype distance are depicted in Figure 1. The parameters (ϵ and *MinPts*) of DBSCAN clustering algorithm can be chosen based on the distribution of the pairwise haplotype distance. In our experiments, we set ϵ as 0.2. *MinPts* is dynamically determined by examining the number of neighbors of each haplotype given ϵ. We sort the numbers in an ascending order and select *MinPts* to be the 3/4 quartile. We take the lengths of haplotype segments as the same as in [10], which are 7 and 21 markers for microsatellite markers and SNP markers, respectively, corresponding to 6-7 cM.

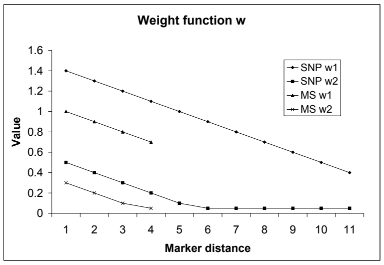

Fig. 1. Weighting functions used in our simulation studies. MS stands for microsatellite marker.

3.2 Results on microsatellite data

Figure 2 shows a typical Z-score distribution map for a data set. Two different haplotype segment lengths are presented, marker interval with length 5 on the left and marker interval with length 7 on the right. The true gene location is halfway between 5 and 6 (depicted by a vertical line in the figure) and the predicted gene location is at marker 5 for both length parameters, with Z-scores of 4.63 and 3.86 respectively. The two maps in general agree with each other. The parameters associated with identified clusters and haplotype patterns are summarized in Table 2. Figure 3 shows the predicted locations (y-axis) and true locations (x-axis) on 100 data sets. It illustrates that the localization accuracy of our method is very good with $A = 10\%$.

In order to compare our results to the results of the HPM algorithm, we take the same graph format to present the effect of phenocopies and sample sizes on the localization accuracy as illustrated in Figure 4. Overall the performances of the two algorithms are similar. For example, for a sample size of 200 (Figure 4 left versus Figure 2a in [10]), the prediction errors are small for $A = 10\%, 7.5\%$, but the error increases rapidly when $A = 5\%$. Neither methods can successfully predict gene locations when A drops to 2.5% (no significant difference from random guess). Our method is slightly better when only considering errors (distances from the predicted location to the true gene location) that are smaller than 4 cM. For both methods, doubling the sample size improves the prediction accuracy greatly. Our method gives better results when $A = 10\%, 7.5\%$ in this case. For example, for $A = 10\%$, all localization errors are within 4.5 cM, while only about 85% of HPM results achieve the same accuracy. HPM performs better when $A = 5\%$. None of the two methods could successfully handle data with $A = 2.5\%$, although the results are better than those with smaller sample sizes. The results show that our method performs consistently with the value of A. While the results of HPM on $A = 10\%$ are worse than its results on $A = 7.5\%$ (Figure 2b in [10]). This type of inconsistency also occurs in other results on HPM. For instance, the results (Figure 2c in [10]) with complete data are worse than the results with 5% corrupted data (by randomly changing 5% alleles), and the results with default parameters (Figure 2f in [10]) are worse than the results with long gaps.

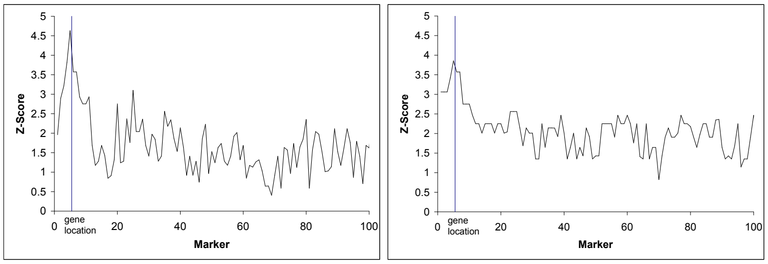

Fig. 2. The Z-score distribution for a data set with different haplotype segment lengths. Left: segment length = 5 markers. Right: segment length = 7 markers.

Table 2. Haplotype clusters identified with different haplotype segment lengths.

Length	Z-score	# haps	# case haps	# core haps	Consensus
5	4.63	24	23	23	66366
7	3.86	18	17	18	1663667

3.3 Results on SNP data

More and more SNP markers will be available for whole genome association studies of common diseases using case-control data. Thus, it is necessary to test the performance of our method on biallelic markers. Again we use the simulated SNP data from [10] with 3 SNPs per cM. We also use the same haplotype segment length of 21 markers corresponding to 7 cM. As expected, with denser SNP markers, our method performs better than on microsatellite markers. For instance, with $A = 10\%$, 98% of predicted errors are smaller than 5 cM and 81% of predicted errors are smaller than 2 cM for SNP markers and the results for microsatellite markers are 94% and 73% respectively. The comparison between SNP markers and microsatellite markers without missing alleles is not shown in [10]. With missing data (12.5% of alleles are randomly removed for both methods), our method first imputes missing alleles based on allele frequencies in order to calculate the pairwise haplotype distances. Figure 5 (right) shows the results on missing data. Comparing to Figure 4 in [10], we have similar prediction accuracy in the case of $A = 10\%$ and better performance in the cases of $A = 7.5\%$ and 5%.

3.4 Results on a real HLA data

We have also tested our method on a real dataset, consisting of affected sib-pair families with type 1 diabetes obtained from [3]. There are a total of 25 microsatellite markers spanning a 14Mb region on chromosome 6 including the entire HLA complex, with known type 1 diabetes-susceptibility locus. To test our algorithm, we first infer the haplotypes from the genotype data using the integer linear programming (ILP) algorithm [5] of our PedPhase program [6, 7]. We only take 89 families out of the original 385 families to run the ILP haplotyping algorithm. (The other families miss the genotypes of all members in at least one locus.) For each such family, a haplotype from the

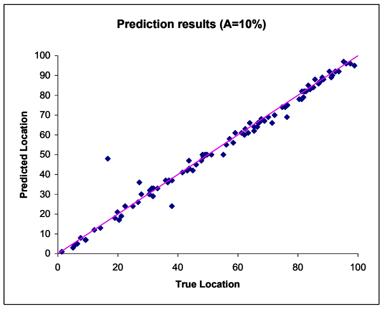

Fig. 3. Predicted location versus true DS gene location.

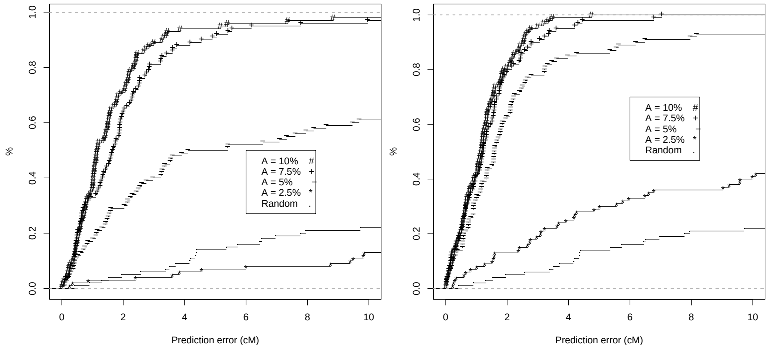

Fig. 4. The effects of A and sample size (left: 200 individuals, right: 400 individuals) on the prediction accuracy.

four parental haplotypes is assigned as a case haplotype if it appears in any of the two affected children. Otherwise it is selected as a control chromosome. There are totally 213 case haplotypes and 143 control haplotypes. Since we have only 25 markers, the length of haplotype segment is set to be 5. The results in Figure 6 show that our algorithm could find the true gene location at marker D6S2444 with a Z-score of 3.72. The associated cluster has 32 haplotypes and only 3 are from control haplotypes. The number of core haplotypes is 27 and the consensus haplotype pattern is 61429.

4 Discussion

W have proposed a new haplotype similarity measure and developed a new method for haplotype LD mapping using case-control data. The method is based on a density-based clustering algorithm and makes no assumptions about the population evolutionary model or disease inheritance patterns. Experimental results on simulated microsatellite/SNP data sets and a real data set of type 1 diabetes disease show that the method provides highly accurate predictions of the DS gene localization for realistic sample sizes, even when the degree of phenocopies is high. The method not only provides a

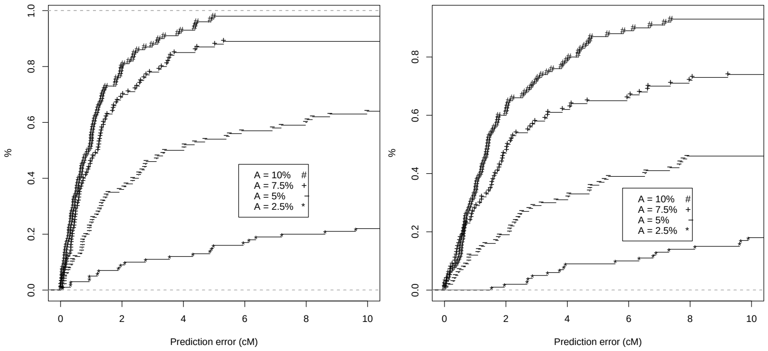

Fig. 5. Results on the SNP data set.

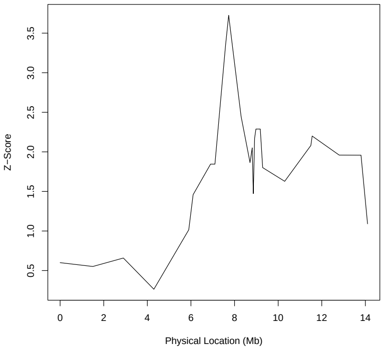

Fig. 6. Results on the real HLA data.

new strategy for gene fine mapping and disease associated haplotype pattern identification, it can also be used to perform whole genome association study with dense SNP markers, given its good scalability.

Preliminary experiment results show that the power of our method consistently increases with the sample size, higher proportion of mutated haplotypes, and denser markers. More experiments will be performed in the future using different parameter combinations. Permutation tests will be done by shuffling the disease status to assess the significant level of predicted results. Extending the method to original genotype data instead of using haplotype data can be achieved by defining a distance measure between genotype segments. For instance, we can define a similarity measure by counting the common alleles of each genotype, weighted by the genetic/physical distance to the locus of interest. Our method is not limited to oligogenic diseases since we can report all gene locations with Z-scores larger than a predefined threshold; but its ability to detect multiple DS associated genes requires more investigation. Quantitative traits that are important to human health can also be analyzed using our framework by first discretizing the involved continuous measurements.

5 Acknowledgement

JL's research is supported by NSF grant CCR-9988353. TJ's research is supported in part by NSF Grants CCR-9988353, ITR-0085910, and CCR-0309902 and National Key Project for Basic Research (973 #2002CB512801).

References

1. M. Easter, H.P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discoving clusters in large spatial databases with noise. *Proc KDD'96*, 226-231, 1996.
2. S. E. Fienberg. The analysis of cross-classified categorical data. MIT Press, 1977.
3. M. Herr *et al.* Evaluation of fine mapping strategies for a multifactorial disease locus: systematic linkage and association analysis of IDDM1 in the HLA region on chromosome 6p21. *Hum Mol Genet* 9(9):1291-1301, 2000.
4. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860-921, 2001.
5. J. Li and T. Jiang. An Exact Solution for Finding Minimum Recombinant Haplotype Configurations on Pedigrees with Missing Data by Integer Linear Programming. To appear *Proc. RECOMB'04*
6. J. Li and T. Jiang. Efficient rule-based haplotyping algorithms for pedigree data. *Proc. RECOMB'03*, pages 197-206, 2003.
7. J. Li and T. Jiang. Efficient inference of haplotypes from genotypes on a pedigree. *J Bioinfo Comp Biol* 1(1):41-69, 2003.
8. M.S. McPeek, and A. H. Strahs. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am J Hum genet*, 65(3):858-875, 1999.
9. P. Onkamo *et al.* Association analysis for quantitative traits by data mining: QHPM. *Ann Hum genet*, 66:419-429, 2002.
10. H. T. 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(1):133-145, 2000.
11. J. Y. Tzeng, B. Devlin, L. Wasserman, and K. Roeder. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am J Hum genet*, 72(4):891-902, 2003.
12. J. C. Venter*et al.* The sequence of the human genome. *Science*, 291(5507):1304-1351, 2001.
13. S. Zhang, K. Zhang, J. Li and H. Zhao. On a family-based haplotype pattern mining method for linkage disequilibrium mapping. *Proc. PSB'02*, 100-111, 2002.