Improvement of Association-based Gene Mapping Accuracy by Selecting High Rank Features

[1]Zahra Mahoor, [2]Mohammad Saraee, [3]Mohammad Davarpanah Jazi
[1,2,3]Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan 84156-83111, Iran
mahoor@ec.iut.ac.ir, saraee@cc.iut.ac.ir , mdjazi@cc.iut.ac.ir

Abstract

The association gene mapping methods based on the haplotype clustering analysis are vastly used to localize a mutation in a gene sequence. These methods employ the concept of "linkage disequilibrium" (LD). In many cases the locations that are found based on these methods have large errors. In this paper, we present a novel technique to decrease the mean error of the association gene mapping in the haplotype clustering analysis. In our technique, we utilize the information gain to select a set of important features (i.e., markers) that are used in the clustering process. In other words, each marker is assigned a rank and then the high ranked markers are fed into the HapMiner algorithm for localizing the disease. Therefore, we limit the HapMiner algorithm to search on a set of dominant features which are selected by the information gain. We have tested the performance of our technique on a set of simulated dataset. Our experiments show a significant reduction in the mean error of the gene mapping.

1. Introduction

A genetic disease is a disease caused by abnormal expression of one or more genes in a person, causing a clinical phenotype. The goal of the study which usually called disease susceptibility loci (DSL) is to locate these genes, so that one can detect, avoid or develop treatment for these diseases. This process is also called gene mapping. Many of the gene mapping techniques utilizes the haplotype analysis method. Most of the haplotype analysis methods employ the concept of "linkage disequilibrium" (LD), which refers to the tendency for alleles at closely linked loci to be associated with each other across unrelated individuals in a population. These gene mapping methods, that utilize the LD concept, are called association gene mapping. In this paper, we call the association gene mapping as gene mapping. By using LD, one can localize a disease-causing variant along a chromosome by detecting patterns of marker values that exist at a putative location at a higher frequency among diseased individuals than among healthy individuals. Haplotypes associated with disease are expected to look similar to one another around the location of the disease-causing mutation; if the functional mutation has occurred only once, they share a common ancestry at that point. Although haplotype-based methods utilize more information, they may lose power as a result of over parameterization, given a large number of haplotypes possible over even a few loci. Recently, new methods have been developed that cluster haplotypes with similar structure in the hope that this reflects shared genealogical ancestry. In the following paragraph, we review some of these clustering-based methods.

Molitor et al. [1] perform fine mapping by spatial clustering of haplotypes based on a similarity metric that measures the length of the shared region and by estimating the risk that each haplotype 'cluster' has for the trait. The authors in [2] propose an algorithm which its base is on [1] by modifying similarity measure and using Markov-chain Monte Carlo algorithm. There are also other methods based on clustering haplotypes including the work by Li et al. [3] that is based on density clustering. Igo et al. [4] have modified a generalized linear model approach for association analysis by incorporating algorithm [3] to reduce the number of coefficients in the model. Durrant et al. [5] use hierarchical clustering to produce approximations of genealogical trees and map genes based on these trees.

There are also other gene mapping methods which are not based on clustering technique. For example, [6] haplotype pattern mining (HPM) is such method. The algorithm finds all haplotype fragments (patterns) of arbitrary length that show statistical association with the disease. In [7], tree pattern mining for gene mapping (TreeDT) has been presented. At each locus, trees that approximate the genealogy of the haplotypes at that locus are constructed. After that a disequilibrium test is performed on each of trees to test if there is a small set of subtrees with relatively high proportions of disease-associated chromosomes, suggesting shared genetic history for those and a likely disease-gene location. In [8] disease-susceptibility genes can be localized directly by measuring the statistical significance of haplotype similarity in the cases without explicit clustering or goodness of fit tests, such as x^2.

The gene mapping methods search through all the markers to find the gene that causes a disease. They do not utilize the general knowledge that which markers have the most effect on the condition of a disease. In this study, we find the markers that have the most effects on the haplotype status, i.e., case-control, by utilizing a feature selection technique where features are the markers. Therefore, the process of finding the location of a gene is only limited to analyzing the markers that have been selected by the feature selector.

Feature selection (also known as subset selection) is a process commonly used in machine learning, wherein a subset of the features available from the data is selected for application of a learning algorithm. The best subset contains the least number of dimensions that most contribute to accuracy; we discard the remaining, unimportant dimensions. This is an important stage of preprocessing and is one of two ways of avoiding the curse of dimensionality. We combine the feature selection approach of [9] with gene mapping density-based clustering algorithm of Li and Jiang [3] for association mapping.

The rest of this paper is organized as follows. In Section 2, our approach for computing the Information Gain and gene mapping with haplotype clustering is presented. Section 3 describes the dataset and the results of the proposed approach on this dataset. Conclusions and future research work are given in Section 4.

2. Methodology

Figure 1 shows the general methodology of our approach. As shown in the figure, we first use information gain as the feature selection method to weight markers in haplotypes. We then select the markers with the highest weights for input to gene mapping that utilizes the HapMiner technique. In other words, in the HapMiner technique rather than sliding a window on all haplotype, we insert window around markers that information gain has assigned them a high weight.

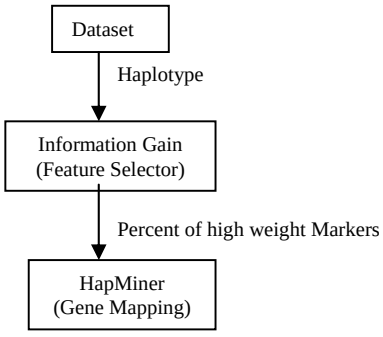

Figure 1. General framework.

2.1. Computation of Information Gain

[9] proposed a classification algorithm called *ID3*, which introduces the concept of information gain. In [10], they use information theory that underpins the criterion to construct the best decision tree for classifying as follows:
"The information conveyed by a message depends on its probability and can measure bits as minus the logarithm to base 2 of that probability."
Let S be the set of n instances and let C be the set of k classes. Let $P(C_i,S)$ be the fraction of the example in S that have class C_i . Then, the expected information from this class membership is as follows:

$$Inf(S) = -\sum_{i=1}^{k} P(Ci,S) \times \log(P(Ci,S)) \quad (1)$$

If particular, attribute A has v distinct values, the expected information required for the decision tree with A as the root is the weighted sum of expected information of the subsets of assorting to distinct valued. Let S_i be the set of instances whose value of attribute A is A_i.

$$Info_A(S) = -\sum_{i=1} \frac{|S_i|}{|S|} \times Info(S_i) \quad (2)$$

Then, the difference between *info(S)* and *infoA(s)* gives the information gained by partitioning S according to testing A.

$$Gain(A) = Info(S) - Info_A(S) \quad (3)$$

2.2 Gene Mapping with Haplotype Clustering

For fine mapping, we use HapMiner algorithm which designed in [3]. In order to keep the paper self-contained, we briefly introduce the HapMiner algorithm. This method directly explores the sharing of haplotype segments in affected individuals that are rarely present in normal individuals. The measure of sharing between two haplotypes is defined by a similarity metric that combines the length of the shared segments and the number of common alleles around any marker position of the haplotypes, which is robust against recent mutations/genotype errors and recombination events. For a pair of haplotypes h_i ,h_j, [3] define the similarity measure as:

$$s_{i,j} = \sum_{k=-l}^{r} w_1(x_k) I(h_i(k), h_j(k)) + \sum_{k=-l'}^{r'} w_2(x_k) \quad (4)$$

You can find more explanation about this similarity measure in [3]. This algorithm scans each marker one by one. For each marker position, a haplotype segment with certain length centered at the position will be considered. Clusters are formed in regions of high density. A haplotype is designated a "core" haplotype if enough density, determined by the density threshold *MinPts*, is located within a given distance ε from it. Haplotypes within this ε

neighborhood are clustered together. A score for each marker will be calculated as follows. We measure the degree of association between a haplotype cluster and the disease of interest using the *Z*-score. In other words, the marker with the highest *Z*-score will be predicted as disease susceptibility loci. Let *m'* and *n'* denote the numbers of case and control haplotypes in a cluster, respectively. A *2× 2* contingency table can be constructed and the *Z*-score of the cluster is defined as:

$$Z = \frac{m'/m - n'/n}{\sqrt{\frac{m'+n'}{m+n}(1-\frac{m'+n'}{m+n})(1/m+1/n)}} \quad (5)$$

3. Experiments and Results

3.1 Data

We use a simulated dataset which was generated by Toivonen [6] to test our technique for gene mapping. The simulated dataset corresponds to a recently founded, relatively isolated founder subpopulation that grew from the initial size of 300 to about 100,000 individuals in 500 years. The considered region is at the chromosome level with genetic length of 100 cM. Both microsatellite markers and SNP markers are simulated. Markers are evenly spaced along the chromosome with interval lengths of 1 cM and 1/3 cM for microsatellite markers and SNP markers, respectively. A dominant disease is modeled, with a large number of phenocopies. The proportion of mutation carrying chromosomes from all the case chromosomes, denoted by *A*, is either 2.5%, 5.0%, 7.5%, or 10.0%, corresponding to overall relative risks = 1.2, 1.7, 2.7, 4.1, respectively. SNP data has 301 markers and microsatellite data has 101 markers.

3.2. Results

We use Information Gain feature selection provided in Weka package [11] to rank the markers of haplotypes. This selection process is based on cross-validation technique by folding 10. In the HapMiner algorithm, we select a window of 9 markers for the microsattelite and a window of 21 markers for the SNP. The weights for calculating the similarity between the haplotypes are $w1=w2=1-0.05\times d$ and $w1=w2=1-0.1\times d$ for the SNP and the microsattellite, respectively. The *MinPts* is set to *25%* and the ε is set to *0.2*. We executed various experiments by choosing 30, 50, 100, 150, 200, 250 or 300 high ranked markers for the HapMiner algorithm.

The results of gene mapping for various experiments are presented in term of the root mean square (rms) error between the actual location of the genes and the predicted location on simulated SNP data by our technique in Table 1. The first row of the table shows the rms of the errors by utilizing the HapMiner technique without ranking the markers and the other rows show the rms errors for various numbers of high ranked markers fed into the HapMiner algorithm. The columns show the effect of different values of *As* on the error. As the table shows, ranking the markers provides a lower rms error rate than the using the markers without ranking. Even in the case that all the ranked markers are used by the HapMiner, the rms error is less than the time that the markers are not ranked. This clearly illustrates the advantage of feature selection on the process of gene mapping in the haplotype clustering approach.

Figures 2(a-c) illustrate the results of gene mapping for SNP and microsatellite data in terms of cumulated prediction error. In these figures, the *x* coordinate represents the distance from the true gene position and the *y* coordinate represents the average fraction (power) of the predictions that were within the distance on 100 genes in the dataset. As the figure shows, our approach outperforms the HapMiner technique alone without ranking the markers.

In addition, we have found that for the microsatellite data, the mean square error rates for different number of selected high ranked markers and the without rank are the same. The results of simulated microsatellite data are presented in Table 2. However, for the SNP data, feature selection has a great impact on reducing the error rate, because information gain is more useful for features that have binary value. With the advance of genotyping technology, more SNP markers will be available for whole-genome association studies of common diseases using case-control data in the near future. Therefore, for any gene mapping method, it is desirable to see that the performance of the method improve with denser markers. Indeed, our approach has performed better on SNP markers than on microsatellite markers.

Table 1. Comparisons of our approach with HapMiner in terms of the root mean square error rate for different *A*s on SNP simulated dataset.

HapMiner	A=10%	A=7.5%	A=5%	A=2.5%
All markers without ranking	28.35	51.89	66.60	126.07
30 high rank markers	**23.58**	**38.79**	**66.42**	**121.78**
50 high rank markers	**14.61**	**42.28**	**59.95**	**117.78**
100 high rank markers	**17.71**	**44.87**	**58.10**	**122.42**
150 high rank markers	**17.74**	**44.87**	**60.93**	**122.45**
200 high rank markers	**17.67**	**46.41**	**63.41**	**123.75**
250 high rank markers	**17.63**	**46.22**	**63.40**	**121.78**
All markers with ranking	**17.70**	**46.25**	**63.21**	**122.61**

Table 2. Comparisons of our approach with HapMiner in terms of the root mean square error rate for different *A*s on Microsaltellite simulated dataset.

	A=10%	A=7.5 %	A=5%	A=2.5 %
All markers without ranking	8.28	16.83	30.5	44.79
30 high rank markers	**8.24**	17.82	31.50	**42.80**
50 high rank markers	8.75	17.52	31.50	**42.80**
All markers with ranking	8.77	**16.66**	**27.22**	**43.68**

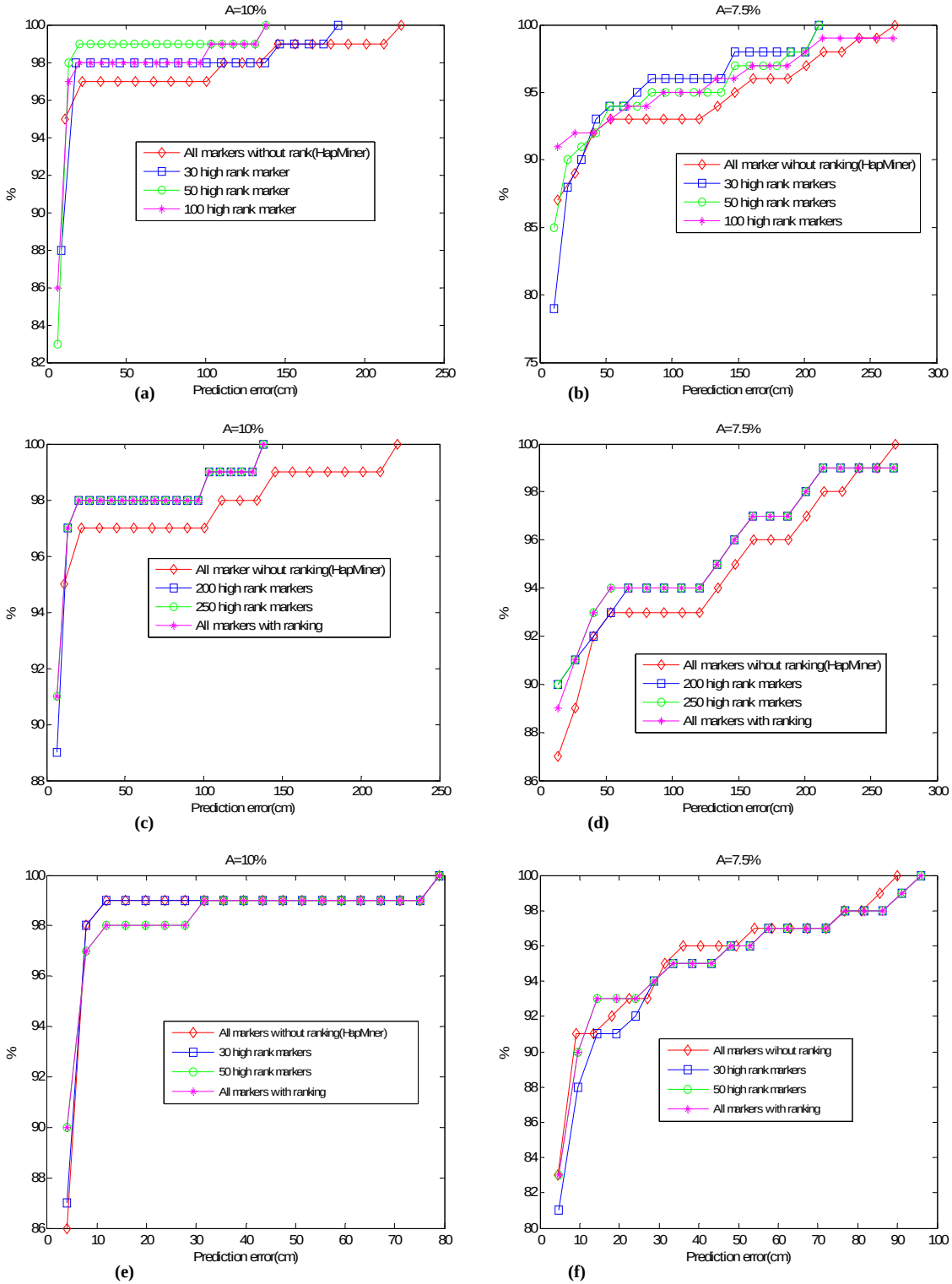

Figure 2. The results on simulated SNP and microsattelite datasets. Comparison between the HapMiner with and without selecting a portion of high rank marker for a sample size of 200 case and 200 controls. SNP data: (a) and (b) A=10%; (c) and (d) A=7.5%; Microsatteliate data: (e) A=10% and (f) A=7.5%.

4. Conclusion

In this paper, we have investigated the effect of information gain of the attributes as the feature weights on the accuracy of gene mapping. Information gain was calculated by standard formula in [9] for the nominal attributes. In the gene mapping with haplotype clustering, by assigning weights to the features, we have significantly improved the accuracy of gene mapping on SNP simulated dataset by mean square root error rate criterion. However, for the microsatellite simulated dataset, root mean square error rate with and without ranking of the markers had the same performance. In other words, our approach performed better on SNP markers than on the microsatellite markers. This is due to the fact the Information Gain functions performs better for the binary attributes. In the future, we will also test the performance of our technique on a real dataset. We will also apply our approach on datasets with different markers to evaluate the effect of the number of the markers in gain information efficiency.

5. Acknowledgments

We are grateful to H.T. Toivonen for sharing the simulated dataset. We thank Jing Li for providing her program HapMiner.

10. References

[1] J. Molitor, P. Marjoram, and D. Thomas, , 'Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques'*, Am. J. Hum. Genet*, 2003, Vol. 73, pp. 1368 – 138.

[2] E.R.B. Waldron, J.C. Whittaker, and D.J. Balding, "Fine Mapping of Disease Genes Via Haplotype Clustering", *Genetic Epidemiology*, 30: 170–179,2006.

[3] J. Li, T. Jiang, "Haplotype-based linkage disequilibrium mapping via direct data mining', *Bioinformatics*, 2005, Vol. 21, pp. 4384 – 4393.

[4] R. P Igo, D. Londono, K. Miller, A. R. Parrado et al., "Density-based clustering in haplotype analysis for association mapping", *BMC Proceedings*, 2007.

[5] C. Durrant, K.T. Zondervan,L.R. Cardon et al. , "Linkage disequilibrium mapping via cladistic analysis of singlenucleotide polymorphism haplotypes', *Am. J. Hum. Genet*. Vol. 75, 2004, pp. 35 – 43.

[6] 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.* 2000, 67:133-145.

[7] P. Sevon, H. Toivonen and V. Ollikainen, "TreeDT: Tree pattern mining for gene mapping', *in IEEE/ACM Transactions on Computational Biology and Bioinformatics* , 2006.

[8] J.Y. Tzeng, B. Devlin, L. Wasserman, et al., "On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit", *Am. J. Hum. Genet.* ,2003, Vol. 72, pp. 891 – 90.

[9] J.R. Quinlan., *Induction of decision trees.* Machine learining,1, 1986.

[10] J.R. Quinlan. , *C4.5: Programs for Machine Learning.* Morgan Kaufmann, California, 1993.

[11] I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition. Morgan Kaufmann, San Francisco, 2005.