Keywords: haplotypes and SNP, tagging, haplotype prediction, linear algebraic methods, phasing
Linear Algebraic Tag SNP Selection and Haplotype Reconstruction

Abstract

Constructing a complete human haplotype map is helpful when associating complex diseases with their related SNPs. Unfortunately, the number of SNPs is very large and it is costly to sequence many individuals. Therefore, it is desirable to reduce the number of SNPs that should be sequenced to a small number of informative representatives called tag SNPs. Also, the tag SNP selection may reduce the noise introducing by irrelevant SNPs for disease association. In this paper, we propose a new linear algebraic method for tag SNP selection and haplotype reconstruction. Our new haplotype reconstruction method is purely combinatorial and can be applied to any set of tag SNPs. We compare the quality of our new linear algebraic methods with several previously known methods. We use the data sets, evaluation methodology, and sometimes tag SNPs suggested by the respective authors. In our comparisons, the proposed linear algebraic algorithm considerably improves the quality of haplotype reconstruction. For example, for the LPL [5] and Chromosome 21 data [15] when 10% of SNPs are used as tags, the new linear algebraic algorithm reaches 80% accuracy, while the methods of Halldorsson et al. [9] and Zhang et al. [19] only reach 20% accuracy.

1 Introduction

In the disease association and susceptibility studies, the DNA of individuals from two population (healthy and sick individuals) is sampled. Then, discrepancies in the haplotype structure of the two population are served as an evidence for the correlation between the haplotypic structure and the disease. Clearly, the statistical significance of the study is affected by the size of the sample population. On the other hand, the total cost of the study is also affected by the number of SNPs typed. Therefore, to reduce cost, one wishes to select the small subset of SNPs, the so called tag SNPs, which predicts the rest of SNPs. Moreover, the tag SNP selection may reduce the noise introducing by irrelevant SNPs for disease association.

The Predictive Tag SNP Selection Problem can be formulated as follows:

Given a sample $S$ of a population $P$ of haplotypes (or genotypes) on $m$ SNPs, find positions of $k$ ($k < m$) tag SNPs such that one can reconstruct the entire haplotype (or genotype) from its restriction on the $k$ tag SNPs.

The solution of this problem consists of two steps. The first step is to identify the tag SNPs in a population of known haplotypes (or genotypes). The second step is to reconstruct a haplotype (or genotype) from its tag SNPs. We can formulate the problem as the following:

We assume that the sample $S$ consists of haplotype data (otherwise, one can phase $S$ using, e.g., PHASE [17]). Note, The haplotype (or genotype) restricted to the tag SNPs may also miss some data.

The first step, tag SNP selection, has received considerable attention in recent years. One established way of selecting tag SNPs is based on linkage disequilibrium (LD) in which the entire SNP sequence is partitioned into blocks, i.e., contiguous SNP segments within which the
number of different haplotypes is relatively small (see [2, 4, 13, 15, 19, 18]). Due to the low diversity within a block, the SNPs are highly correlated and a very small number of tag SNPs can predict values of all other SNPs. In the paper of Clark et al. [6], one can find a valuable discussion of the tasks and limitations of LD approaches. An alternative way is to ignore block structure and select tag SNPs across the whole region under study [1, 14, 16, 9].

The second step, reconstructing a haplotype (or genotype) from its typed tag SNPs, has received much less attention. The paper of Zhang et al. [19] presents a method for selecting tag SNPs based on haplotype (or genotype) data, then reconstructing haplotypes from typed tag SNPs with the partition-ligation-expectation-maximization algorithm. The method is tested on the data set described in the paper of Daly et al. [7] and is capable of recovering 90% of the original haplotypes using only 35% of the SNPs as tags. In contrast to the block-based methods of Zhang et al. [19], Halldorsson et al. [9] describes a block-free approach to tag SNP selection. Their method considers a graph whose vertices are SNPs; two vertices are connected with an edge if one SNP can be used to reliably predict the other. The vertices (SNPs) with high degree are chosen as tags. To reconstruct the value of an unknown SNP in a given haplotype, that SNP’s neighbor’s values are inspected and a majority vote is taken. The method is tested on three different data sets with leave-one-out cross-validation and can recover 90% of the haplotype data using only 20% of SNPs as tags. The paper of Forton et al. [8] is a study of the accuracy of haplotype reconstruction using tag SNPs. They compared the block-based approach of Zhang et al. [19] with the block-free, entropy-based approach of Ackerman et al. [1] on both the 5q31 cytokine gene cluster and the IL-8 region on chromosome 4 across a European and a West African population. The study concludes that the block-free method generally chooses fewer SNPs as tags, while block-based method has higher accuracy when reconstructing haplotypes from tag SNPs.

He et al. [10, 12] presents a linear algebraic method for selecting tag SNPs and reconstructing the haplotype. Their method is based upon linear dependency between SNPs. Gauss-Jordan elimination is used to identify a set of SNPs which are linearly independent across the population of haplotypes. By multiplying the restricted haplotype matrix (i.e. the haplotype matrix with all non-tag SNPs removed) by the row-reduced echelon form of the unrestricted haplotype matrix, the original haplotypes are reconstructed.

Our model assume that the tag values for the genotypes are known 1. We obtain the tag values of the haplotypes by phasing (PHASE et.). In the traditional approach (see Forton et al. [8], the complete haplotype is extrapolated from the tags by finding the best matched haplotype in the sample with the minimum Hamming distance between its restricted haplotype and the tags. The alternative way is to reconstruct the complete haplotype from its tags; the final haplotype is identified as the bests matched haplotype in the sample to the reconstructed haplotype. In this paper, we apply the new linear algebraic algorithm using the second model which obtains 2% higher reconstruction accuracy.

In this paper, we propose a new linear algebraic method for tag SNP selection and haplotype reconstruction which exploits linear dependency between SNPs. We apply our method to real and simulated data sets. In all of our comparisons, the proposed linear algebraic approach is a considerable improvement on the quality of full haplotype reconstruction from typed SNP tags as well as tag SNP selection.

The rest of the paper is organized as follows. The next section introduces the basic linear algebraic method for tag SNP selection and haplotype reconstruction. Then, we present the
The Basic Linear Algebraic Method for Tag Selection and Haplotype Reconstruction

This section first briefly describes linear dependency of SNPs and haplotypes following He et al. [10]. Then we describe our basic linear algebraic method which uses linear dependency of tags and matrix multiplication for haplotype/genotype reconstruction.

Typically, in genetic sequences derived from human haplotypes, the number of sites is much larger than the number of individuals [15]. Because of such disproportion, many columns corresponding to SNP sites are similar. Indeed, as noted in [15], the number of equivalent sites in real data is considerably large. The 0-1-column-site $s_i$ is equivalent to the site $s_j$ if either $s_i = s_j$, or $s_i$ is complimentary to $s_j$ (i.e., $s_i$ becomes $s_j$ after each 0 is replaced with 1 and each 1 is replaced with 0). It is common to keep only one site out of several equivalent sites since they do not carry any additional information.

In general, if one column-site can be restored from several other columns, then it can be dropped without loss of information. In this paper we consider restoration of one column-site using a linear combination of other column-sites.

As noted in [2], one cannot straightforwardly apply linear combinations of column-sites since equivalent columns are linearly independent. But one can easily overcome this obstacle by replacing 0’s with −1’s [11]. From now on we will change SNP notations: −1 corresponds to the wild type and 1 corresponds to the mutation. The advantage of (−1, 1)-notations is that two sites are equivalent if and only if they are collinear (i.e., linearly dependent). Similarly the genotypes attain the notation of (0, 1, 2)-notation where 2 corresponds heterogenous site
are replace with \((-1,1,0)\)-notation. As the result, genotype vector is linear combination of two haplotypes, i.e., \( g = (h + h')/2 \). Our tagging method is based on keeping only linearly independent SNPs as tags.

The basic linear algebraic method for tagging assumes that if there is a linear dependency between certain SNPs in the given sample \( H \), then the same dependency is likely to hold for these SNPs in the entire population \( P \). The experimental study shows that this assumption is true for simulated and real data.

The basic linear algebraic method for tagging consists of the following steps:

- From the sample haplotype matrix \( H \), extract the maximum number \( r = \text{rank}(H) \) of linearly independent columns-SNPs \( T(H) = \{H_{t_1}, \ldots, H_{t_r}\} \) forming a basis of columns-SNPs of \( H \). The columns-SNPs in \( T(H) \) form the set of tag SNPs.
- For each column-SNP \( H_{j}, j = 1, \ldots, m \) in \( H \), find a unique representation of \( H_{j} \) as a linear combination of tag SNPs
  \[
  H_{j} = \sum_{i=1}^{r} \alpha_{i,j} H_{t_i}
  \]
  For example, if \( H_{j} \) is a tag, i.e., \( H_{j} = H_{t_i} \), then \( \alpha_{i,j} = 1 \) and \( \alpha_{i',j} = 0 \), \( i \neq i' \).
- Output the positions \( \{t_1, \ldots, t_r\} \) of tag SNPs of \( T(H) \) and the matrix \( F = (\alpha_{i,j}) \) of coefficients of linear combinations.

The linear algebraic method can be efficiently implemented by applying Gauss-Jordan elimination to obtain reduced row echelon (RREF) \( R \) from \( H \). The RREF will have exactly \( r \) tag SNPs formed by linearly independent column-sites corresponding to nonzero rows can be easily found from \( R \). Let \( F \) be the matrix \( R \) in which zero rows are dropped, so \( F \) is an \( r \times m \) matrix. Then for any haplotype \( h \) with the tag SNP values \( h_r \), the predicted reconstruction \( \tilde{h} = f(h_r) \) equals
\[
\tilde{h} = h_r \times F
\]
One cannot guarantee all the values of \( \tilde{h} \) to be either 1 or -1. Therefore we postprocess \( \tilde{h} \) as follows: if the value of an SNP in \( \tilde{h} \) is negative, we set it to \(-1\), otherwise we set it to 1.

In \((-1,1,0)\)-notations, a genotype vector \( g \) is obtained from haplotype vectors \( h \) and \( h' \) if and only if \( g = (h + h')/2 \). Therefore, we directly apply the above method for genotype reconstruction \( \tilde{g} = g_r \times RREF(G) \), where \( G \) is the matrix of a genotype sample, \( g_r \) is the tag SNPs of a genotype \( g \).

3 Linear Algebraic Method for Tag Selection

He et al. [10, 12] uses linear reduction method for selecting tag SNPs and reconstructing haplotypes. The LRP algorithm selects linearly independent SNPs from the left side of the haplotype matrix as tags. The RLRP randomly permutes the columns of the haplotype matrix before performing the linear algebraic method. In He et al. [10, 12], it is shown that the RLRP method has a lower error rate than the LRP method under identical conditions. The ‘question’ which basis will be the better set’ is still open. In this section, we attempt to predict the quality of basis from the corresponding RREF.
Figure 2 plots the number of zeros in each column of the row-reduced echelon form of the haplotype matrix and reconstruction error rate for each column in the sample using the RLRP method. It is easy to see that the number of zeros in RREF is highly correlated to the reconstruction rate. There is an intuitive explanation for this correlation – if in the RREF corresponding to a non-tag SNP contains many zeros, then it can be represented as a linear combination of relatively few tag SNPs. This in turn indicates that this linear combination is not result of noise and unlikely to be artefact.

It is computationally infeasible to exhaustively search for the optimal set of linearly independent column-sites which produces the largest number of zeros in the row-reduced echelon form of the haplotype matrix. In stead of exhaustively search, we apply the greed method. We suggest the following Separable Linear Algebraic Tagging (SLT) method. First, the row-reduced echelon form of the haplotype matrix \( R = \text{RREF}(H) \) is computed using Gauss-Jordan Elimination and the number of zeros in \( R \) is calculated. Then, for each column \( r_i \) in \( R \) containing a pivot, the corresponding column \( s_i \) in \( H \) is swapped with each column \( s_j \) in \( H \) that does not contain a pivot and \( R = \text{RREF}(H) \) is recomputed. If the number of zeros has increased, then we keep the swap, otherwise we swap the columns back to their original positions. The algorithm stops when no swap gives increased number of zeros. the rearranged columns of \( H \) gives the largest number of zeros in \( R \).

The SLT method is not an exhaustive search since it doesn’t test every linearly independent subset of columns of size \( r = \text{rank}(H) \), and it is possible that when the algorithm stops the optimal subset has not been achieved. In practice, however, the SLT algorithm find the close to optimal solutions.

In case when the required number of tags \( k \) is less than the linear rank of \( H \), following He et al. [12] we reduce the sample to \( k \) linear independent haplotypes. it is noticed that it is better to choose the most representative haplotypes, i.e., haplotypes that can predict all others with the least number of errors.

In case when the required number of tags \( k \) is more than the linear rank \( r \) of \( H \), following
He et al. [12] we add $k - r$ column-SNPs which contains less zeros in RREF(H) obtained by applying the SLT method on the $H$.

4 Linear Algebraic Method for Haplotype Reconstruction

He et al. [12] suggests the haplotype reconstruction algorithm separately for the case (i) when the required number of tags $k$ is greater or equal to the rank of the sample $r$ and (2) when the sample rank $r$ is small. We found out that while finding additional tags in case (i) works well, their method does not work as good for the second case. In this section, we first describe their voting algorithm for the case (i) and then show how to reduce the case (i) to the case (ii) by linear algebraic means. Finally, based on the reconstruction algorithm for the case (ii), we suggest a new way how to handle missing tags.

Case (i): $k \geq r = \text{rank}(S)$. The method in [12] suggests to repeatedly choose $k - r + 1$ different subsets of $r$ linearly independent tags and reconstruct all other SNP using the basic linear algebraic method from Section 2. This can be done by placing $k$ tag columns to the beginning of the sample matrix $S$, $k - r + 1$ times randomly permuting these $k$ tags and finding RREF of the resulted matrices.

Since there are $k - r + 1$ different RREF’s, there that many different reconstructions for each haplotype. The aggregation of all these reconstructions is suggested to be done by “voting”: each SNP attained the value of $-1$ (respectively, 1) if the majority of $k - r + 1$ reconstructions suggests $-1$ (respectively, 1).

Case (ii): $k < r = \text{rank}(S)$. The method in [12] does not suggests any difference in haplotype reconstruction from the case when $k = \text{rank}(S)$. Just put the $k$ tag SNP columns in the beginning of the sample matrix $S$ and find RREF of the resulted matrix keeping only its first $k$ rows. According to the basic reconstruction algorithm, the tag vector of the given haplotype/genotype should multiplied by this RREF to obtain reconstruction of the full unknown haplotype/genotype vector. As a result the reconstruction is poor since it does not exploit all information from the sample $S$. Indeed, a single RREF obtained from $S$ and uses only to the first $k$ rows.

In this paper, we suggest just to reuse the algorithm from the case (i) after transposing the matrix $S$. Indeed, for the transposed matrix $S^T$, we get into the same problem as in case (i) – how to choose $k$ columns out of $r = \text{rank}(S^T)$ columns. First we need to correctly choose $r - k$ additional columns corresponding to the different sample haplotypes. Here we repeat the procedure used for selecting additional tags – find the columns that have the worst reconstruction by first $k$ haplotypes. These haplotypes will have the maximum number of non-zero values in RREF of $S^T$.

Reconstruction algorithm is also very similar to the case (i). We first find $r$ linear independent rows of $S$ and use RREF of $S$ to predict the given haplotype/genotype. After that the rows of $S$ are permuted $r - k$ times and more $r - k$ different reconstruction of the given haplotype/genotype are obtained. Finally, the aggregation of all these $r - k + 1$ reconstructions is done by the same “voting” as in case (i).
Handling missing tag SNPs. If genotyping of tag SNPs of unknown haplotype/genotype contains '?'s, i.e., lacks certain values, then it can be handled in the same way as above. We just simply treat such event as if we would never expected to know these tags. If as a result the number of tags becomes less than the rank, we use the reconstruction algorithm from case (ii), otherwise we use algorithm from case (i).

5 Experimental Results

We compare the SLT method with the methods described in [10, 12, 9, 19, 8] on the following four data sets:

Chromosome 5q31. The Daly et al. [7] data set consists of 516 haplotypes containing 103 SNPs that are derived using trio phasing described in Brinza et al. [3] from 129 family trios on a 616 kilobase region of human chromosome 5q31 that may contain a genetic variant responsible for Crohn’s disease. For each trial, sample data sets of size 10-100 are extracted from the data, and missing data is introduced into the remaining haplotypes at 0%, 5%, and 10% of the total number of SNPs. The SLT method as described in Section 3 is used to reconstruct the remaining haplotypes. The results are averaged over 10 trials and are reported as the average hamming distance between the reconstructed and real haplotypes as a percentage of the haplotype length (see Figure 3). We compare the SLT method with the methods of He et al. [10, 12] on this data set. The linear algebraic method reliably (error rate below 2%) recovers all SNPs based upon a very small portion of tag SNPs (e.g., 32 tags out of total 103 SNPs) while sampling below 8% of the population.

21 data set consists the first 1,000 of 24,047 SNPs typed over 20 haploid copies of human Chromosome 21 [15]. This subset was found to be highly representative to the entire data set. Both the LPL and Chromosome 21 data sets were used in Halldorsson et al. [9]. We compare the SLT method with the methods of Halldorsson et al. [9] and Zhang et al. [19] on these two data sets by using leave-one-out cross-validation to evaluate the quality of the solution.

The haplotype left out is reconstructed by the SLT method. The Hamming distance between the reconstructed haplotype and the leave-one-out haplotype is recorded; the error rate is the number of errors in the reconstruction as a percentage of the haplotype length. The average error rate in reconstructing all of the haplotypes is used as a measure of the overall accuracy of the tagging method on the data set. The methods of Zhang et al. [19] and Halldorsson et al. [9] impute a SNP based on the tag SNPs in the same block or neighborhood. Therefore, if there is no tag SNP in the block or neighborhood, then these methods do not predict a value for the SNP. The SLT method reconstructs each SNP based on the values of all tag SNPs. Figure 4 compares the SLT method with the methods of Zhang et al. [19] and Halldorsson et al. [9]. The figure shows that, for the LPL and Chromosome 21 data when 10% of SNPs are used as tags, the SLT method reaches 80% accuracy, while the methods of Halldorsson et al. [9] and Zhang et al. [19] only reach 20% accuracy.

Figure 4: The x-axis shows the number of tag SNPs, and the y-axis shows the fraction of SNPs correctly imputed in a leave-one-out experiment. (A) Results from SLT, Halldorsson et al. and Zhang et al. for the LPL data set. (B) Results from the SLT method, Halldorsson et al. and Zhang et al. for the Chromosome 21 data set.

Large simulated data sets. The Forton et al. [8] data set consists of 32 family trios of European descent genotyped over 122 SNPs across 654 kilobase of the 5q31 cytokine gene cluster. SNPs that violate the Hardy-Weinberg law and SNPs with frequency less than 5% are cleansed from the population. After cleansing, the data set contains 99 SNPs. 95 haplotypes and their frequencies are computed using Phamily and PHASE as described in [17]. To create model populations with which to conduct haplotype reconstruction experiments, first an initial population of 100,000 individuals is created by selecting random pairs of haplotypes, while ensuring that the haplotype frequencies remained unchanged. From this initial population, 5 smaller populations consisting of 760 unrelated individuals are created by randomly selecting individuals. Five levels of missing data (1%, 2%, 5%, 10% and 20%) are introduced to the
The method of Ackerman et al. [1] identifies 22 tag SNPs for the 5q31 cytokine gene cluster. To reconstruct the haplotypes of each population, we assume that the tag values for the genotypes are known. We obtain the tag values of the haplotypes by SNPHAP. In Forton et al. [8], the complete haplotype is extrapolated from the tags by finding the haplotype in the sample with the minimum Hamming distance between its restricted haplotype and the tags. In contrast to Forton et al. [8], we reconstruct the complete haplotype using the linear algebraic method on the same set of tags; the final haplotype is identified as the haplotype in the sample with the minimum Hamming distance to the reconstructed haplotype. The error is recorded as the percentage of incorrect haplotype reconstructions (see Figure 5A.). The SLT method is computational more expensive but obtains higher reconstruction accuracy. For example, at the 20% missing level, we obtain 2% higher reconstruction accuracy. Figure 5B shows the percentage of error at each haplotype locus over all simulated populations with different levels of missing data.

![Graphs showing percentage of error at each haplotype locus over all simulated populations with different levels of missing data.](image)

Figure 5: (A) The x-axis shows the percentage of missing data, and the y-axis shows the percentage of incorrect haplotype reconstructions. Results are from the simulated data sets. (B) The percentage of errors at each haplotype locus over all simulated populations with different levels of missing data of the simulated data sets.

### 6 Conclusions & Future Work

We have suggested a new linear algebraic method for tag SNP selection and haplotype reconstruction. Our new haplotype reconstruction method is purely combinatorial and can be applied to any set of tag SNPs. The experimental results show the proposed new linear algebraic algorithm is a considerable improvement on the quality of haplotype reconstruction from typed SNP tags. For example, for the LPL [5] and Chromosome 21 data [15] when 10% of SNPs are used as tags, the new linear algebraic algorithm reaches 80% accuracy, while the methods of Halldorsson et al. [9] and Zhang et al. [19] only reach 20% accuracy. In our future work, we will apply tag selection on genotype data and will apply tag SNP selection for predicting disease association and susceptibility.
References


