



Haplotype Reconstruction from Genotype Data using Imperfect Phylogeny

Eran Halperin¹ and Eleazar Eskin²

¹CS Division, University of California Berkeley, Berkeley, 92093-0114, CA and

²School of Computer Science and Engineering, Hebrew University, Jerusalem, Israel, 91904,

ABSTRACT

Critical to the understanding of the genetic basis for complex diseases is the modeling of human variation. Most of this variation can be characterized by single nucleotide polymorphisms (SNPs) which are mutations at a single nucleotide position. To characterize the genetic variation between different people, we must determine an individual's *haplotype* or which nucleotide base occurs at each position of these common SNPs for each chromosome. In this paper, we present results for a highly accurate method for haplotype resolution from genotype data. Our method leverages a new insight into the underlying structure of haplotypes which shows that SNPs are organized in highly correlated "blocks". In a few recent studies (see Daly *et al.* (2001); Patil *et al.* (2001)), considerable parts of the human genome were partitioned into blocks, such that the majority of the sequenced genotypes have one of about four common haplotypes in each block. Our method partitions the SNPs into blocks and for each block, we predict the common haplotypes and each individual's haplotype. We evaluate our method over biological data. Our method predicts the common haplotypes perfectly and has a very low error rate (less than 2% over the data from Daly *et al.* (2001).) when taking into account the predictions for the uncommon haplotypes. Our method is extremely efficient compared to previous methods, such as PHASE and HAPLOTYPYER. Its efficiency allows us to find the block partition of the haplotypes, to cope with missing data and to work with large data sets.

Availability: The algorithm is available via webserver at <http://www.calit2.net/compbio/hap/>

Contact: eran@eecs.berkeley.edu,eeskin@cs.huji.ac.il

INTRODUCTION

Most of the variation within a population can be characterized by single nucleotide polymorphisms (SNPs) which are mutations at a single nucleotide position that occurred once in human history and were passed on through heredity. Approximately 10 million common SNPs (NIH (2002); Group (2001)), each with a frequency of 10% to

50% account for the majority of the variation between DNA sequences of different people (Patil *et al.* (2001)). To characterize an individual's variation, we must determine an individual's *haplotype* or which nucleotide base occurs at each position of these common SNPs for each chromosome. By correlating an individual's haplotypes with the presence of a disease, researchers can better understand complex diseases. The effort to characterize human variation, currently a major focus for the NIH, will be a tremendous undertaking requiring obtaining the haplotype information from a large collection of individuals from diverse populations (NIH (2002)).

Although the two chromosomes of an individual can be separated and analyzed independently as in Patil *et al.* (2001), current technology suitable for large scale polymorphism screening obtains *genotype* information at each SNP. The genotype gives the bases at each SNP for both copies of the chromosome, but loses the information as to the chromosome on which each base appears. Consider a SNP where there are two common bases, *A* or *G*. There are four possible cases for the haplotype. Two of the cases are *homozygous* genotypes, where either both chromosomes contains *A* or both chromosomes contain *G*. The other two cases are *heterozygous* genotypes, where the first chromosome contain *A* and the second contain *G* and vice versa. For this SNP, there are three possible cases for the genotype information. In the homozygous cases, the genotype will be either *A* or *G* respectively and we can infer that the base appears in both chromosomes. In the heterozygous cases, the genotype will be *H* (for heterozygous) and we can infer that in one chromosome, we have an *A* and in the other we have a *G*, but we can not infer on which chromosome each appears which causes ambiguities in reconstructing the haplotypes. Consider the example where an individual at four successive SNPs, with possible values *A* or *G*, has a genotype *AHGG*. In this case, the individual's haplotypes have two possibilities: either *AAAG* on one chromosome and *AGGG* on the other chromosome or *AAGG* and *AGAG*. Without any other information, such as the genotypes from related individuals, it is impossible to determine the individual's

| Haplotype | 0,1 Representation | Frequency |
|-----------|--------------------|-----------|
| CCGAT | 00000 | 66 |
| CTGAC | 01001 | 24 |
| ATACT | 11110 | 10 |
| CTGAT | 01000 | 6 |
| ATGAT | 11000 | 1 |
| ATGCC | 11011 | 1 |
| CCGAC | 00001 | 1 |

Table 1. Block 6 from Daly *et al.* (2001). The block contains 5 SNPs over 11 kilobases. The horizontal line separates the common haplotypes from rare haplotypes. The first column shows the haplotypes from the transmitted chromosomes. The second column shows the same haplotypes but mapped to 0,1 representation. The 0 represents the common nucleotide at the position, while the 1 represents the rare nucleotide at the position. The third column is the frequency of the haplotype block in the transmitted chromosomes. Note that any chromosome that contained any ambiguity in the block due either to missing data or heterozygous genotypes for all members of the trio was omitted.

actual haplotypes. This problem of haplotype resolution is often referred to as the phase problem.

Recent studies in linkage disequilibrium (Goldstein & Weale (2001); Reich *et al.* (2001)) characterizing human haplotype structure observe that for the individuals and the specific regions they consider, the SNPs are grouped into “blocks” of limited diversity perhaps due to recent bottlenecks in human history. In each block containing n SNPs, typically around four haplotypes account for the majority of the haplotypes in the population. Consider the haplotype block shown in Table 1 consisting of 5 SNPs over 11 kilobases from a recent paper, Daly *et al.*, 2001. We can map each of these haplotypes to the 0,1 representation where 0 represents the common nucleotide and 1 represents the rare nucleotide. The 0,1 representation for block 6 is also shown in Table 1. Note that 90% of the haplotypes in the population are one of four common haplotypes.

Existing methods to resolve haplotypes include the original approach of (Clark (1990)), parsimony approaches (Gusfield (2000, 2001); Lancia *et al.* (2001)), maximum likelihood methods (Excoffier & Slatkin (1995); Fallin & Schork (2000); Hawley & Kidd (1995); Long *et al.* (1995)), statistical methods such as PHASE (Stephens *et al.* (2001)) and HAPLOTYPYER Niu *et al.* (2001), and perfect phylogeny-based approaches (Gusfield (2003)). Some of those previous methods implicitly assume limited diversity of the haplotypes in the population. This assumption does not hold over long regions. Consider the entire 616 kilobase region examined in Daly *et al.* (2001). Out of the 258 haplotypes in the population, the most common haplotype appears 45 times and 169 haplotypes are unique. Understandably, the error rate of predictions of one of the most accurate methods, PHASE, over this long region is 25.25%. Similarly, another accurate

method HAPLOTYPYER over the first 55 SNPs[†] in the region has an error rate of 15.67%.

In this paper, we present results for a highly accurate method for haplotype resolution from genotype data, the HAP algorithm. HAP takes as input a population of genotypes and partitions the SNPs into blocks of limited diversity. For each block HAP predicts the common haplotypes as well as the haplotypes of each individual in the population. We also show that the common haplotypes roughly fit a perfect phylogeny model. Essentially, HAP can effectively predict the haplotypes for *unrelated* individuals. This ability significantly reduces the costs and difficulties of characterizing human variation since it eliminates the need for collecting genotype data from complete trios.

Haplotypes can be resolved from genotype data by making the assumption that most of the haplotypes within a block of limited diversity will loosely fit the perfect phylogeny model. This method for resolving haplotypes was first proposed in Gusfield (2003). The perfect phylogeny model assumes an infinite site mutation model and allows no recombinations (Hudson (1990)). The infinite site mutation model makes the assumption that at each SNP site, a mutation only happened once in human history. This model forbids recurrent mutations or back mutations. The assumptions of the model imply that a chromosome with a mutation at a SNP is a direct descendant from the chromosome of the ancestor in which the mutation occurred. Likewise, any chromosome without the mutation can not be a descendant of a chromosome that has the mutation. Clearly, these assumptions are not realistic although it is reasonable to assume that recombinations and recurrent mutations are relatively rare events within a block. Thus, we consider a relaxed model which allows for a certain number of recurrent mutations and recombinations within a block. A similar approach to ours is the strict perfect phylogeny model approach of Gusfield (2003); Bafna *et al.* (2002a). However, in this paper, we show that only if we relax the assumptions of the perfect phylogeny model do these type of approaches work over real data.

We compared HAP to two of the most popular methods for phase resolution (HAPLOTYPYER and PHASE), and found that HAP is considerably faster than both in all cases, and just as or more accurate than both methods in most cases. In fact, in some of the experiments we ran, HAPLOTYPYER was still running without giving results after hours of CPU while HAP made predictions in seconds.

RESULTS

[†]This was the largest subset of the SNPs that HAPLOTYPYER was able to handle using its default settings.

Predicting Haplotypes from Genotypes We performed our experiments over two data sets presented in Daly *et al.* (2001) and Gabriel *et al.* (2002). The data from Daly *et al.* (2001) contains a set of genotypes for 103 SNPs collected from 129 mother father and child trios of European decent. We use the portion of the data from Gabriel *et al.* (2002) which has pedigree information. This portion consists of genotypes of SNPs from 62 regions from two populations. The first portion consists of genotypes from 3277 SNPs from 93 individuals from 12 multi-generational pedigrees of European ancestry and the second consists of genotypes from 3061 SNPs from 90 individuals from 30 trios from Yoruba.

The pedigree information allows us to infer the haplotypes from the genotypes for most of the SNPs, assuming Mendelian genetics, that is, assuming no recombination. For the Daly *et al.* (2001) data, we only use the children’s genotypes as an input to the algorithms, ensuring that the genotypes are independent. We use the pedigree information in order to evaluate the accuracy of the algorithms.

The Daly *et al.* (2001) data is partitioned into 11 blocks of limited diversity. Our first set of experiments over this data assumes that we are given the block partition for the 11 blocks in the data. Our second set of experiments assumes that we have no prior information about the block partition of the 103 SNPs and are only given their genotypes. We apply our algorithm to determine the block partition. We then compare our predictions to the predictions of PHASE and HAPLOTYPYER.

Over the Gabriel *et al.* (2002) data, since no partition of the SNPs is defined in the data, we partition the regions into blocks and predict the haplotypes within each block. We also make predictions using PHASE and HAPLOTYPYER over our predicted blocks.

Predicting Haplotypes within a Block In the first set of experiments, for each of the 11 blocks as defined in Daly *et al.* (2001) we predicted the common haplotypes from the genotypes of the children in the trios as well as each child’s haplotype. In all cases the predictions for the common haplotypes are correct. A significant portion of the data is missing, 10.03% of the total genotype data. This missing data comes from various sources of experimental error. In addition to resolving the heterozygous genotypes, we resolve the missing data. Throughout the paper, the error rate is the number of bases predicted incorrectly divided by the number of heterozygous and missing genotypes. The number of bases predicted incorrectly is the Hamming distance between the predicted haplotypes and the correct haplotypes in the best of the two possible assignments of the pairs of predicted to correct haplotypes. In our predictions, over the 129 individuals and 103 SNPs, the error rate is only 0.91% and the error rate in the presence of missing data

is only 1.27%, significantly lower the amount of missing genotype data. Over individuals that contain the common haplotypes, our predictions are perfect. The errors only occur for individuals who have an uncommon haplotype. A comparison of the error rates with those of PHASE and HAPLOTYPYER are shown in Table 2 and 3.

The program takes only a few seconds to make each of these haplotype predictions, while PHASE and HAPLOTYPYER took significantly longer. Running times for each of these programs is shown in Table 4.

| SNPs | Predicted Common Haplotypes | Freq. | HAP Error Rate | PHASE Error Rate | HAPLOTYPYER Error Rate |
|--------|---|-----------------------------|----------------|------------------|------------------------|
| 1-8 | GGACAACC AATTCGTG | 215 38 | 0.0000 | 0.0000 | 0.0298 |
| 10-14 | TTACG CCCAA | 217 35 | 0.0000 | 0.0000 | 0.0106 |
| 16-24 | CGGAGACGA GACTGGTCG CGCAGACGA CGGATACGA | 139 52 34 15 | 0.0188 | 0.0209 | 0.0230 |
| 25-35 | CGCGCCCGGAT CTGCTATAACC CTGCCCCGGCT TTGCCCAACC | 142 39 35 25 | 0.0048 | 0.0016 | dnf |
| 36-40 | CCAGC CCACC GCGCT CAACC | 146 51 30 12 | 0.0193 | 0.0193 | 0.1159 |
| 41-45 | CCGAT CTGAC ATACT | 152 63 31 | 0.0326 | 0.0181 | 0.0688 |
| 78-84 | CGTTTAG TGTTGA TGATTAG CGTCTAG TGTTGGA | 142 53 20 12 10 | 0.0111 | 0.0084 | 0.0250 |
| 86-91 | ACAACA GCGGTG ACGGTG GTGACG | 145 71 14 13 | 0.0223 | 0.0198 | 0.0273 |
| 92-98 | GTTCTGA TGTGTAA TGTGCGG TGCGTAA | 142 49 32 15 | 0.0131 | 0.0183 | 0.0471 |
| 99-103 | CGGCG TATAG TATCA | 112 105 35 | 0.0000 | 0.0000 | 0.0436 |

Table 2. Predictions over blocks defined by Daly *et al.* (2001). The second column shows the common haplotypes as presented in Daly *et al.* 2001. The third column shows the predicted common haplotypes and the fourth gives their frequencies. The fifth through seventh columns give the error rate after resolving all missing data. The error rate is the total number of errors in the predictions divided by the total number of heterozygous and missing genotypes in the block. “dnf” in the error rate corresponds to when the HAPLOTYPYER program failed on the input. The total error rate over all blocks for HAP is 0.0127, for PHASE is 0.0165 and for HAPLOTYPYER is 0.04 over the regions that HAPLOTYPYER returned a prediction.

| SNPs | Predicted Common Haplotypes | Freq. | HAP Error Rate | PHASE Error Rate | HAPLOTYPYER Error Rate |
|-------|---|----------------------------|----------------|------------------|------------------------|
| 46-76 | CCCTGCTTACGGTGCAGTGGCACGTATTGCA TCCCATCCATCATGGTCGAATGCGTACATTA CCCCGCTTACGGTGCAGTGGCACGTATATCA CATCACTCCCCAGACTGTGATGTTAGTATCT CCCTGCTTACGGTGCAGTGGCACGTATTGCA | 137 59 19 10 9 | 0.0143 | 0.0307 | dnf |

Table 3. Predictions over data from Daly *et al.* (2001) (continued).

| SNPs | HAP CPU Time | PHASE CPU Time | HAPLOTYPYER CPU Time |
|--------|--------------|----------------|----------------------|
| 1-8 | 0.18 | 46.89 | 4.29 |
| 10-14 | 0.05 | 44.58 | 54.56 |
| 16-24 | 0.06 | 52.74 | 7.93 |
| 25-35 | 0.08 | 60.99 | dnf |
| 36-40 | 0.04 | 34.64 | 20.36 |
| 41-45 | 0.18 | 35.57 | 90.97 |
| 46-76 | 0.06 | 207.14 | dnf |
| 78-84 | 0.06 | 41.21 | 299.12 |
| 86-91 | 0.06 | 47.48 | 346.52 |
| 92-98 | 0.06 | 46.83 | 247.61 |
| 99-103 | 0.06 | 40.31 | 45.61 |

Table 4. CPU time measured in seconds for prediction of haplotypes within blocks in data from Daly *et al.* (2001). “dnf” stands for “did not finish” after several hours of computation.

Predicting Blocks from Genotypes Typically, we must determine the block partition directly from the genotype data. We first make haplotype predictions for all possible blocks using the local haplotype prediction algorithm and discard any blocks with more than five common haplotypes since we are looking for low diversity regions. We chose the number five, since in Daly *et al.* (2001), Patil *et al.* (2001) and in Gabriel *et al.* (2002) the number of common haplotypes is smaller than five in the vast majority of cases. However, we can use the same dynamic programming using any criterion for determining candidate blocks.

Out of those candidate blocks, we choose the optimal block partition. There are a few possible criteria to determine which is a good block partition such as linkage disequilibrium based criteria Daly *et al.* (2001) or information based criteria Bafna *et al.* (2002b). One possible criterion for determining a good block partition is minimizing the sum of the number of representative SNPs over all blocks. This criterion has been used to partition blocks on a larger scale (Patil *et al.* (2001); Zhang *et al.* (2002)).

For each block, we can define a set of representative SNPs that are sufficient to determine an individual’s haplotypes. In Table 1, the second, third and fifth SNPs are sufficient to determine the haplotype. For example, if we observe T , A , and T in these SNPs, we can infer that the individual has the third haplotype (assuming the

| Population | Number of SNPs | HAP Error Rate | PHASE Error Rate | HAPLOTYPYER Error Rate |
|------------|----------------|----------------|------------------|------------------------|
| European | 3277 | 0.0352 | 0.0375 | 0.0478 |
| Yoruba | 3061 | 0.0380 | 0.0441 | 0.0489 |

Table 5. Results from predictions over the Gabriel *et al.* (2002).

individual has one of the common haplotypes). On the other hand, if we observe T , G , and C , we can infer the individual has the second haplotype. The minimum number of representative SNPs is three. That is, no two SNPs can distinguish the four common haplotypes.

Using dynamic programming, we choose the best block partition for the data from Daly *et al.* (2001) where the objective is to minimize the number of representative SNPs over the entire block partition. The total error rate for predictions in these blocks for HAP is 0.0109, for PHASE is 0.0190 and for HAPLOTYPYER is 0.0363 over the regions that HAPLOTYPYER returned a prediction.

The exact predicted block partition is described in supplementary materials on the website. The block partition varies from the partition described in Daly *et al.* (2001) (shown in Table 2) since the criteria for defining block partitions are different. Our criterion, to minimize the number of representative SNPs, is consistent with the criterion in Patil *et al.* (2001) while the criterion in Daly *et al.* (2001) determines blocks by estimating the recombination frequencies between blocks.

For the Gabriel *et al.* (2002) data, we partition each of the 62 regions for both populations into blocks and make predictions over the haplotypes within each block. For each block, we make predictions for each individual. We measure the error rate by inferring the actual haplotypes from the pedigrees. In some cases for the multi-generational pedigrees, there were conflicts in the inference perhaps due to recombination. For the evaluation, we omitted any pedigrees that had conflicts. Table 5 summarizes the results over the data. Details of the predictions are provided in the supplementary materials on the website.

We wish to emphasize that the issue of efficiency and speed is crucial, since for instance, for the Daly *et al.* (2001) data, in order to optimally partition the data, we

need to make predictions for over 1000 candidate blocks since the number of candidate blocks increases linearly with the size of the data. Therefore, the fact that our algorithm is considerably faster than both HAPLOTYPER and PHASE is an important advantage of our algorithm.

DISCUSSION

Recent studies in haplotype structure have shown that haplotypes are structured into blocks with limited diversity. Over these blocks, many methods can effectively resolve haplotypes from genotypes of SNPs for populations. For longer regions, the SNPs must be partitioned into regions of limited diversity. We have presented a method, HAP, for partitioning genotypes of SNPs into blocks and making predictions for haplotypes for each block.

A recent paper by Gusfield (2003) suggested the use of perfect phylogeny to reconstruct the haplotypes. However, the actual haplotypes do not fit the perfect phylogeny model. For a given set of haplotypes, we can measure the percentage of conflicts to the perfect phylogeny model. If we consider all haplotypes, even the uncommon ones, the data does not fit the perfect phylogeny model as shown in Figure 1A. If we consider instead a relaxed perfect phylogeny model using an error threshold which allows a small number of haplotypes to be excluded when determining conflicts, we notice that the haplotypes fit the model much better. The results for the Chromosome 5q31 data are in Figures 1A-C. Clearly, as the error threshold increases, the number of conflicts significantly decreases. This is due to the fact that the infrequent haplotypes cause the majority of the conflicts with the perfect phylogeny model.

We have demonstrated our method over actual haplotype data and verified the accuracy of our predictions to the haplotypes inferred by pedigrees. Our method is highly accurate and efficient. The predictions differ from the haplotypes inferred by the trios by less than 2% over the data from Daly *et al.* (2001) even after resolving approximately 10% of the missing genotype data. We also present a method for determining the block partition from genotype data and a method for extending haplotype predictions beyond single blocks.

The program for predicting haplotype structure is publicly available via a webserver at <http://www.calit2.net/compbio/hap/>.

We note that one disadvantage of the perfect phylogeny method with respect to the statistical methods is that it is not clear how to give estimates of uncertainty in the predictions. We note that one could possibly extend our methods in order to provide such estimates.

In many cases, the data can be split into blocks or regions of low diversity. In cases where there is no underlying "block" structure, HAP would partition the

data into very small regions. For this type of data, it is not clear if splitting into regions of low diversity is the best approach to the problem.

METHODS

Dataset Description We use two data sets for our experiments. The first data set is a 500 kilobase region of chromosome 5q31 containing 103 SNPs from the studies of Daly *et al.* (2001) and Rioux *et al.* (2001). In this study, genotypes for the 103 SNPs are collected from 129 mother, father, child trios from a European-derived population in an attempt to identify a genetic risk factor for Crohn's disease. A significant portion of the genotype data (10.03%) is missing with an average of 10 SNPs per individual's genotype missing. The 103 SNPs were split into 11 blocks containing from 5 to 31 SNPs and ranging from 3 to 92 kilobases. For each of these blocks, four haplotypes correspond to 90% of the individual chromosomes. Since this set consists of trios, we can infer each individual's haplotypes in all positions except for the positions where all three individuals are heterozygous.

We use populations *A* and *D* from the Gabriel *et al.* (2002) data which has pedigree information. The data consists of genotypes of SNPs from 62 regions. Population *A* consists of 93 individuals from 12 multi-generational pedigrees of European ancestry and population *D* consists of 90 individuals from 30 trios from Yoruba.

To evaluate our predictions of haplotypes, we make predictions over the genotype data of the individuals and then compare our predictions to the haplotypes inferred from the pedigrees.

Inferring Haplotypes from Trios We use data collected in trios to measure the accuracy of our method. Given the genotypes for a mother, father, child trio, in most cases, we can infer the haplotypes for each of the individuals. We infer the haplotypes at each SNP independently assuming Mendelian genetics. We define each parent to have a transmitted chromosome and an untransmitted chromosome. The child has both transmitted chromosomes from the parents. Each SNP for each chromosome can be represented by either 0 or 1 for the common base or mutation base respectively. For these four chromosomes, there are a total of 16 possibilities. Each SNP in the genotype can be denoted either 0, 1 or 2 which represents homozygous with the common base, homozygous with the mutation base, or heterozygous respectively. Although there are 27 possible genotypes for each trio at a given SNP, many of them are invalid such as the case where the father and child are homozygous for the common base and the mother is homozygous for the mutation base. In any valid case where at least one of the genotypes in the trio is homozygous, we can uniquely determine the haplotypes for that SNP. Only

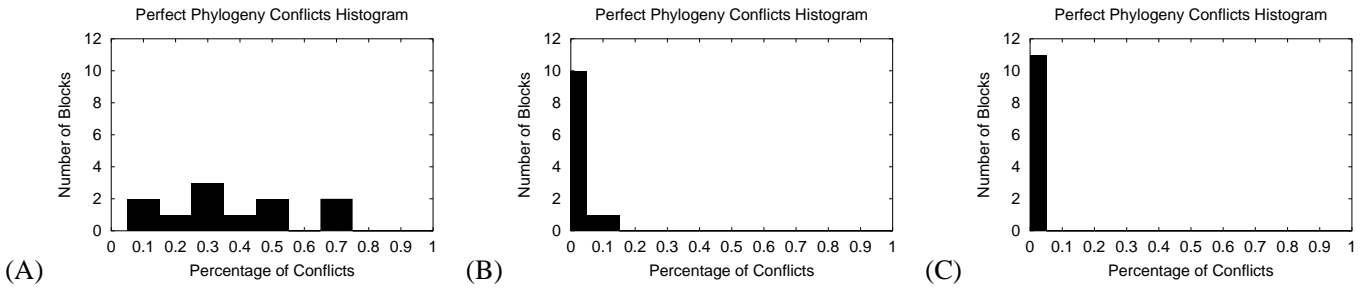


Fig. 1. Histograms of percentage of conflicts under different error thresholds for the blocks defined in Daly *et al.* (2001). Thresholds are (A) 0% (B) 5% and (C) 10%.

in the case where all three of the genotypes are heterozygous is there more than one possible resolution.

Measuring Perfect Phylogeny from Haplotypes The perfect phylogeny model implicitly defines a phylogenetic tree for the haplotype data such as the one for the four common haplotypes from Table 1 shown in Figure 2. At each edge of the tree, we have a mutation labeled with the position of the mutation. Under the perfect phylogeny model assumptions, there can only be one edge for each site in the data. Once a mutation occurs at an edge, the mutation must be present in each individual in the subtree rooted at that edge and only in the subtree. Each haplotype at a node contains all of the mutations along the path from the root node to the current node.

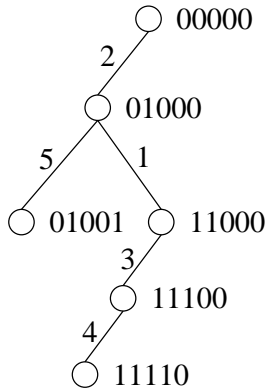


Fig. 2. The Perfect Phylogeny Tree for the data from Table 1

We can measure how well a set of haplotypes fits the perfect phylogeny model by constructing a phylogenetic tree for the haplotypes. These trees can be determined by inferring the relations between sites from the individual’s haplotypes. For instance, if there exists an individual which has a mutation at both sites i and j , we can infer that sites i and j must be along the same path in the tree, and thus one of them is an ancestor of the other.

In general, for each pair of sites, each row determines some constraints on the relation of these sites in the tree. Since there are only four possible configurations, (that is $(0, 0)$, $(0, 1)$, $(1, 0)$ and $(1, 1)$), there are only three possible constraints ($(0, 0)$ does not imply a constraint under the assumption that the root is the all zeros vector).

Not all haplotypes fit the perfect phylogeny model. In many cases the above constraints contradict each other. It is a well known fact that whenever all four possible constraints occur, this implies the nonexistence of such a tree which is a conflict to the model.

We can measure how well a block fits perfect phylogeny by counting the number of conflicts between pairs of sites within the block. For a block containing n SNPs, we can normalize the count by $\binom{n}{2}$ to compare blocks that contain a different number of SNPs. In general, the infrequent haplotypes cause many conflicts with the perfect phylogeny model. We adapt this measure to evaluate how well the majority of the data fits the perfect phylogeny model, by introducing an error threshold. We consider a pair of sites to have a conflict if the number of individuals that contain $(1, 1)$, $(1, 0)$ and $(0, 1)$ are all above the error threshold. For example consider sites 4 and 5 in Table 1 considering all of the haplotypes. For these two sites we have 25 individuals that have $(0, 1)$, ten individuals that have $(1, 0)$ and only a single individual that has $(1, 1)$. For an error threshold of 1 or higher, we would not consider this as a conflict. In Figure 1 we illustrate the effect of different thresholds on the model.

Haplotype Resolution Via the Perfect Phylogeny Model The problem of haplotype resolution as perfect phylogeny was proposed in Gusfield (2003), where he also proposed a polynomial time algorithm. His algorithm is based on heavy machinery from matroid theory, and it is not clear how to extend it to data which does not fit exactly to the perfect phylogeny model. We therefore use as a base to our algorithm the algorithm of Eskin *et al.* (2003). We further extend this algorithm to the case where the data is noisy or simply does not fit exactly the perfect phylogeny model.

| Genotype | 0,1 Representation | Frequency |
|----------|--------------------|-----------|
| CHGAH | 02002 | 23 |
| CCGAT | 00000 | 20 |
| HHHHT | 22220 | 11 |
| HTHHH | 21222 | 4 |
| CTGAH | 01002 | 3 |
| CTGAC | 01001 | 2 |
| CCGAH | 00002 | 1 |
| AHHHT | 12220 | 1 |
| HTHHT | 21220 | 1 |
| ATHCH | 11212 | 1 |
| CHGAT | 02000 | 1 |

Table 6. Genotypes from block 6 from Daly *et al.* (2001). The block contains 5 SNPs over 11 kilobases. The block represents SNPs 41-45 of the 103 SNPs. The first column shows the genotypes from the 129 children with H representing the heterozygous genotype. The second column shows the same genotypes but mapped to 0,1,2 representation. The 0 represents the homozygous genotype of the common nucleotide at the position, while the 1 represents the homozygous genotype of the rare nucleotide at the position. A 2 represents the heterozygous genotype. The third column is the frequency of the genotype among the 129 children. Note that any genotypes that contained any missing data were omitted.

In order to describe this extension, one has to be familiar with the algorithms in Eskin *et al.* (2003). We give here a very simplified sketch of the algorithm.

The algorithm is basically inferring the different relations between the pairs of sites. As mentioned above, these relations are determined by the possible configurations in each of the rows. For each row we add a constraint, depending whether it is a (0, 1) configuration, (1, 0) or (1, 1). Since only the genotypes are given, in many cases one cannot predict the actual constraints. The algorithm uses the constraints that can be predicted in order to construct all possible trees.

Since the data does not perfectly fit the model, we change that algorithm by deciding that a constraint is valid only if there is enough evidence for the constraint, that is, there must be enough rows that imply that constraint. In our experiments, if at least 2% of the rows imply the constraint then we consider the constraint as valid. This results in a more relaxed model since we remove some of the constraints.

Maximum Likelihood Model for Local Haplotype Reconstruction We choose the “best” solution from the set of candidate solutions that roughly fit the perfect phylogeny model using a maximum likelihood model. The maximum likelihood model estimates the likelihood of observing the population of genotypes given the predicted haplotype frequencies. Our likelihood model assumes a Hardy-Weinberg equilibrium, that is, random and independent mating.

Given a population of n individuals, we denote the two haplotypes of the i th individual as i_1 and i_2 . We use the notation $f(i_1)$ to denote the frequency of the haplotype i_1 in the population. The likelihood of a

haplotype i_1 is $\frac{f(i_1)}{2^n}$. The likelihood for each genotype of an individual is simply the product of the likelihoods of their two haplotypes $\frac{f(i_1)f(i_2)}{(2^n)^2}$. The likelihood of a candidate solution for a population of genotypes is

$$L = \prod_{i=1}^n \frac{f(i_1)f(i_2)}{(2^n)^2} \quad (1)$$

This model is consistent with previous maximum likelihood models for choosing haplotypes from genotypes (Excoffier & Slatkin (1995); Hawley & Kidd (1995); Long *et al.* (1995); Fallin & Schork (2000); Stephens *et al.* (2001)). These algorithms are using various methods (such as MCMC or the EM algorithm) to find a local maxima of the likelihood function. The idea in our algorithm is to save time by not trying to reach a local maxima, but rather use the solutions given by the perfect phylogeny procedure as candidates solutions, and picking the best one out of them. In Gusfield (2003) it is shown that the number of possible solutions that fit the perfect phylogeny model is bounded by 2^m where m is the number of SNPs. Since m is quite small in practice, and since in practice much less than 2^m solutions are found, we are able to do that by enumerating over all possible solutions.

We believe that this explains the running time and the accuracy of our algorithm. Our method uses the perfect phylogeny to speed up the algorithm, but uses the maximum likelihood approach in order to increase the sensitivity. We note that it is possible to take the result of algorithm and use it as a seed to PHASE Stephens *et al.* (2001), which is a method based on Markov Chain Monte Carlo. Assuming that our algorithm gives a solution which is close to a local maxima, we expect it to speed up the running time of PHASE, and get more accurate results.

Resolving Missing Data Missing data is resolved after the algorithm resolves heterozygous genotypes. For this, we use a simple extension of Eskin *et al.* (2003) in which the decision for the relationships between SNPs ignores missing data. We then apply the maximum likelihood model over the possible solution given by the perfect phylogeny procedure. To compute this solution, we only use the genotypes which do not have any missing data. We then resolve the missing data by choosing the most likely SNP based on the maximum likelihood model. Effectively, we resolve the missing data by choosing the SNP to match the common haplotypes.

Computing Block Partitions from Genotypes Our method predicts block partitions directly from the genotype data. We first define a set of candidate blocks. Given a maximum block length, we slide a window across the data for each block length to define our candidate blocks. For each candidate block, we apply the local haplotype prediction

algorithm to predict the haplotypes. Our algorithm accurately predicts haplotypes only if there is limited diversity within a block. To ensure accuracy of our predictions, we discard all candidate blocks that have more than five common haplotypes. For each remaining candidate block, we determine the number of representative SNPs. This is done by enumerating over all subsets of the SNPs in the block and checking to see if they distinguish between the common haplotypes.

To compute the block boundaries for the haplotypes, we use a straightforward dynamic programming technique similar to the one presented in Zhang *et al.* (2002). The main difference is that in our setting, there is no missing data since it is resolved by the local prediction algorithm. Note that the block partition in Daly *et al.* (2001) does not assign several SNPs to blocks. We can easily modify the dynamic programming algorithm to optimize a block partition where several SNPs are allowed to be left out.

REFERENCES

- Bafna, V., Gusfield, D., Lancia, G. & Yoosheph, S. (2002a). Haplotyping as perfect phylogeny: A direct approach. *Technical Report UC Davis CSE-2002-21*.
- Bafna, V., Halldorsson, B., Schwartz, R., Clark, A. & Istrail, S. (2002b). Haplotypes and informative snp selection algorithms: Don't block out information. In *Proceedings of the 7th International Conference on Computational Molecular Biology (RECOMB 2003)*, pp. 19–27.
- Clark, A. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. *Journal of Molecular Biology and Evolution*, **7**, 111–122.
- Daly, M., Rioux, J., Schaffner, S., Hudson, T. & Lander, E. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics*, **29**, 229–32.
- Eskin, E., Halperin, E. & Karp, R. M. (2003). Efficient reconstruction of haplotype structure via perfect phylogeny. *Journal of Bioinformatics and Computational Biology*, **1**, 1–20.
- Excoffier, L. & Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, **12**, 921–7.
- Fallin, D. & Schork, N. (2000). Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *American Journal of Human Genetics*, **67**, 947–59.
- Gabriel, G., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E., Daly, M. & Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.
- Goldstein, D. & Weale, M. (2001). Population genomics: linkage disequilibrium holds the key. *Current Biology*, **11**, R576–R579.
- Group, T. I. S. M. W. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–33.
- Gusfield, D. (2000). A practical algorithm for optimal inference of haplotypes from diploid populations. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*.
- Gusfield, D. (2001). Inference of haplotypes from samples of diploid populations: complexity and algorithms. *Journal of Computational Biology*, **8**, 305–23.
- Gusfield, D. (2003). Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions (extended abstract). In *Proceedings of the 6th International Conference on Computational Molecular Biology (RECOMB 2002)*.
- Hawley, M. & Kidd, K. (1995). Haplo: a program using the em algorithm to estimate the frequencies of multi-site haplotypes. *Journal of Heredity*, **86**, 409–11.
- Hudson, R. (1990). Gene genealogies and the coalescent process. *Oxford Survey of Evolutionary Biology*, **7**, 1–44.
- Lancia, G., Bafna, V., Istrail, S., Lippert, R. & Schwartz, R. (2001). Snps problems, algorithms and complexity, european symposium on algorithms. In *Springer-Verlag, (ed.) Proceedings of the European Symposium on Algorithms (ESA-2001), Lecture Notes in Computer Science*, volume 2161, pp. 182–193.
- Long, J., Williams, R. & Urbanek, M. (1995). An e-m algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human Genetics*, **56**, 799–810.
- NIH (2002). Large-scale genotyping for the haplotype map of the human genome. RFA: HG-02-005.
- Niu, Qin, Xu & Liu (2001). In silico haplotype determination of a vast set of single nucleotide polymorphisms. Technical report, Department of Statistics, Harvard University.
- Patil, N., Berno, A., Hinds, D., Barrett, W., Doshi, J., Hacker, C., Kautzer, C., Lee, D., Marjoribanks, C., McDonough, D., Nguyen, B., Norris, M., Sheehan, J., Shen, N., Stern, D., Stokowski, R., Thomas, D., Trulson, M., Vyas, K., Frazer, K., Fodor, S. & Cox, D. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–23.
- Reich, D., Cargill, M., Bolk, S., Ireland, J., Sabeti, P., Richter, D., Lavery, T., Kouyoumjian, R., Farhadian, S., Ward, R. & Lander, E. (2001). Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.
- Rioux, J., Daly, M., Silverberg, M., Lindblad, K., Steinhart, H., Cohen, Z., Delmonte, T., Kocher, K., Miller, K., Guschwan, S., Kulbokas, E., O'Leary, S., Winchester, E., Dewar, K., Green, T., Stone, V., Chow, C., Cohen, A., Langelier, D., Lapointe, G., D, G., Faith, J., Branco, N., Bull, S., McLeod, R., Griffiths, A., Bitton, A., Greenberg, G., Lander, E., Siminovitich, K. & Hudson, T. (2001). Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to crohn disease. *Nature Genetics*, **29**, 223–8.
- Stephens, M., Smith, N. & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.
- Zhang, K., Deng, M., Chen, T., Waterman, M. & Sun, F. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Science*, **99**, 7335–9.