Haplotypes versus Genotypes on Pedigrees

B. Kirkpatrick*

Jan 7, 2010

Abstract

Genome sequencing will soon produce haplotype data for individuals. For pedigrees of related individuals, sequencing appears to be an attractive alternative to genotyping. However, methods for pedigree analysis with haplotype data have not yet been developed. In addtion, the computational complexity of such methods has been an open question. Furthermore, it is not clear in which scenarios haplotype data would provide better estimates than genotype data for quantities such as recombination rates. This paper addresses each of these open questions.

We give a reduction from genotype problem instances to haplotype problem instances and show that solving the haplotype problem yields the solution to the genotype problem, up to constant factors or coefficients. The pedigree analysis problems we consider are the likelihood, maximum probability haplotype, and minimum recombination haplotype problems.

We also introduce a hidden Markov model for haplotype data and compare its recombination estimates to estimates produced by a genotype HMM. While having haplotype data on all individuals usually produces better estimates, having untyped founders can produce estimates from the haplotype data that have similar accuracy as the estimates from genotype data.

1 Appendix

Recall that x is a vector of haplotypes, with two haplotypes per individual in the pedigree, and that s is a vector of inheritance indicators that describe which grand-parental haplotype each of the haplotypes originated from. We describe the probability of a haplotype vector as $\mathbb{P}[x] = \sum_s \mathbb{P}[x|s]\mathbb{P}[s]$. We are interested in the following three pedigree problems:

*Electrical Engineering and Computer Sciences, University of California Berkeley, e-mail: bbkirk@eecs.berkeley.edu

Likelihood. Find the probability of the observed data by summing over all the possible unobserved haplotypes, i.e.

$$\sum_x \sum_s \mathbb{P}[x|s]\mathbb{P}[s].$$

Maximum Probability. Find the values of $x^m_{i,j}$ and $x^p_{i,j}$ that maximize the probability of the data, i.e.

$$\max_x \sum_s \mathbb{P}[x|s]\mathbb{P}[s].$$

Minimum Recombination. Find the values of $x^m_{i,j}$ and $x^p_{i,j}$ that minimize the number of required recombinations, i.e.

$$\min_{x,s} \sum_i \sum_{j=2}^{l} \mathbb{I}[s^p_{i,j-1} \neq s^p_{i,j}] + \mathbb{I}[s^m_{i,j-1} \neq s^m_{i,j}]. \tag{1}$$

To prove the complexity of these three problems, we use a reduction from the genotype problem to the haplotype problem, and we show that the solutions to the haplotype problems yield a corresponding solution to the original genotype problem. Specifically, the solution to the haplotype version of the problem is the solution to the genotype version with the values of the functions being related by constant factors or coefficients, depending on whether the function is a recombination count or a probability.

Mapping. Given a pedigree with genotype data, for any of the three pedigree problems, we define a polynomial mapping to a corresponding haplotype problem with exactly $5|G|$ individuals haplotyped. First we create the pedigree graph for the new haplotype instance, and later we construct the required haplotype observations from the genotype data.

Let $G \subset I$ represent the set of genotyped individuals in a pedigree having individuals I and edges E. We will create a haplotype instance of the problem, with individuals $H \cup I$ and edges $R \cup E$. To obtain the set H, we add five individuals, i_0, i_1, i_2, i_3, i_4, to H for every individual $i \in G$. The set of new relationship edges, R, will connect individuals in sets H and G. Specifically, the edges stipulate that i and i_0 are the parents of full-siblings i_1, i_2, i_3, and i_4 by including the edges: $i_0 \rightarrow i_1$, $i_0 \rightarrow i_2$, $i_0 \rightarrow i_3$, $i_0 \rightarrow i_4$, $i \rightarrow i_1$, $i \rightarrow i_2$, $i \rightarrow i_3$, and $i \rightarrow i_4$. We will refer to these five individuals, i_0, i_1, i_2, i_3, and i_4, and their relationships with i as the *proxy family* for individual i. This produces a pedigree graph with exactly $5|G| + |I|$ individuals and $8|G| + |E|$ edges.

To obtain the new haplotype data from the genotype data, we type only individuals in $|H|$ such that the corresponding genotyped individual in G is required, by the rules of inheritance, to have the observed genotypes. Without loss of generality, assume that the genotype alleles are sorted such that $g^0_{i,j} < g^1_{i,j}$. Now we can easily constrain the parental genotype for individual $i \in G$ by giving the spouse, i_0, homozygous haplotypes of all ones while giving child i_1 the haplotypes $\{\vec{1}, g^0_i\}$, child i_2 haplotypes $\{\vec{1}, g^1_i\}$. These two children guarantee the

correct genotype. we will use the remaining two children to represent possible re-assortments of the genotyped parent's T_i heterozygous loci, indexed by t_j where $1 \leq j \leq T_i$. In addition to the haplotype $\vec{1}$, child i_3, will have haplotype consisting of $h_{i_3,t_j} := g_{i,t_j}^{1-j \mod 2}$ while child i_4 has the genotyped parent's complementary alleles $h_{i_4,t_j} := g_{i,t_j}^{j \mod 2}$. This results in child i_3 and i_4 alternating in having the smaller allele at every other heterozygous locus.

To prove the result for the minimum number of recombinations, we exploit the alternating pattern of alleles assigned to the four children. We find that this pattern forces there to be two recombinations, among the four children, between consecutive heterozygous loci.

Lemma 1. *Let r_g be the minimum number of recombinations in the genotype problem instance. The mapping yields a haplotype problem instance having*

$$r_h = r_g + \sum_{i \in G} 2(T_i - 1)$$

for the minimum number of recombinations, where T_i is the number of heterozygous sites in genotype i.

Proof of Constant Dependence for Minimum Recombination. Consider the haplotype instance of the problem. Recall that set G is defined as the individuals who are genotyped in the genotype problem instance, and, by construction, they are not haplotyped in the haplotype problem instance. For each $i \in G$ the rules of inheritance applied to i's proxy family dictate that the set of alleles at each position are given by $g_{i,j}^0$ and $g_{i,j}^1$. Therefore, the proxy family dictates the genotype of i.

Since the haplotypes for all the typed individuals are completely given, we only need to consider the assortment of the alleles from g_i^0 and g_i^1 into the maternal and paternal alleles of individual i. Clearly this assortment determines the number of recombinations that the proxy family contributes to Eqn. (1). However, we will use induction along the genome to show that every possible phasing of the parental genotype induces the same minimum number of recombinations among the four children, namely $2(T_i - 1)$.

Now we define an arbitrary assortment of the genotype alleles into two haplotypes for person i. We can think of this parental genotype for l loci as a string $s \in \{H, T\}^l$, where H represents a homozygous site and T a heterozygous site. Recall that T_i is the number of heterozygous sites in the genotype string, and those sites appear at indices t_j where $1 \leq j \leq T_i$. For this genotype there are $2^{T_i - 1}$ pairs of haplotypes that phase the given genotype. Represent each pair by setting $T_i - 1$ binary variables

$$P_{t_j} = \begin{cases} 0, & \text{if } x_{i,t_j}^p < x_{i,t_j}^m, \\ 1, & \text{otherwise.} \end{cases}$$

Note, that we are only interested in the origin of the children's haplotypes, rather than in the origin of i's haplotypes, so the p and m can arbitrarily label either haplotype.

Since $\{i_1, i_2\}$ between them have the parent genotype at every locus, one of them has origin p while the other has origin m, and similarly for $\{i_3, i_4\}$. For each locus, indicate the paternal origin of the allele for individuals i_1 and i_3, respectively with Q_j and S_j. Formally, $Q_j = 1$ if both $h_{i_1,j} = x^p_{i,j}$ and $h_{i_2,j} = x^m_{i,j}$ while $Q_j = 0$ otherwise. Similarly, $S_j = 1$ if both $h_{i_3,j} = x^p_{i,j}$ and $h_{i_4,j} = x^m_{i,j}$ while $Q_j = 0$ otherwise.

Define R_j as the minimum recombination count before locus j. Notice that P_{t_1} sets the origin of all the child haplotypes, therefore $R_{t_1} = 0$, since all preceding homozygous loci can have the same origin as locus t_1.

From t_j to t_{j+1} we have two cases:

1. If $P_{t_j} = P_{t_{j+1}}$, then $Q_{t_j} = Q_{t_{j+1}}$ and $S_{t_j} \neq S_{t_{j+1}}$, by the alternating construction of children i_3 and i_4 as compared with i_1 and i_2.
2. Similarly, if $P_{t_j} \neq P_{t_{j+1}}$, then $Q_{t_j} \neq Q_{t_{j+1}}$ and $S_{t_j} = S_{t_{j+1}}$.

Furthermore, regardless of the number of homozygous loci separating t_j and t_{j+1}, the number of recombinations can only be increased. Therefore, we have the recursion

$$R_{t_{j+1}} = 2 + R_{t_j},$$

proving the lemma. □

To prove that the reduction works for the probability distributions, we need to sum over the ways for the children to inherit their haplotypes from the proxy parent. This means we are summing over all the possible numbers of recombinations between consecutive heterozygous loci. We can easily do this by constructing a 16-state Markov chain with a transition step at each locus. Due to the construction of the proxy family, this calculation depends only on the proxy parent's genotype and is independent of the haplotypes assigned to that parent. After we prove this, it is straight-forward to apply the result to the likelihood and maximum probability haplotype problems.

Lemma 2. *The mapping yields a haplotype problem instance having haplotype probabilities proportional to the haplotype probabilities of the genotype instance. Specifically, for all x,*

$$\begin{aligned}\sum_s \mathbb{P}_h[x|s]\mathbb{P}[s] \;\; = \;\; & \left(\sum_{\{s_i|i\in I\}} \mathbb{P}_g\big[\{x_i|i \in I\}|\{s_i|i \in I\}\big] \; \mathbb{P}\big[\{s_i \in I\}\big] \right) \\ & \cdot \prod_{i\in G} p_t(i) \prod_j \mathbb{P}[x^p_{i_0,j} = 1]\mathbb{P}[x^m_{i_0,j} = 1]\end{aligned}$$

where the proxy family transmission probability is a function of genotype g_i, the recombination rate $\theta \leq 0.5$, and of the transition matrices P, Q_{0110}, and Q_{1001},

$$p_t(i) = \left(\frac{1}{16}\right) \vec{1} \cdot P^{h_0} \prod_{j=0}^{T_i} \big(O_j Q_{0110} + (1 - O_j)Q_{1001}\big) \cdot P^{h_j} \cdot \vec{1}^T$$

and O_j indicates whether index j is odd, h_0 is the number of homozygous loci that begin proxy parent's genotype, and h_j is the number of consecutive homozygous loci after the j'th heterozygous locus where there are T_i heterozygous loci for proxy parent i.

The transition probabilities are given by $P_{ij} = \theta^{H(i,j)}(1-\theta)^{4-H(i,j)}$ where $H(i,j)$ is the Hamming distance between inheritance states i and j. Let Q_{0110} be a transition matrix having non-zero recombination probabilities only in column 0110 *(i.e. $Q_{0110,i,j} = P_{ij}$ when $j =$* 0110*). Similarly, let Q_{1001} be a transition matrix with non-zero recombination probabilities only in column* 1001*.*

Proof of Proportional Haplotype Probability. Without loss of generality, assume that individuals $i \in G$ are all fathers in their proxy family. This is simply for convenience of notation.

Let x be any fixed assignment of haplotypes to all the individuals in the pedigree. When conditioning on the assigned haplotypes for individual i, the probability of the proxy family of i is independent of the probability for the rest of the pedigree. Since we can say this for all the proxy families, the terms in the probability for the pedigree individuals in set I (i.e. those also in the genotype pedigree) are equal to the probability on the genotype data in the genotype pedigree. Therefore, we write that

$$\begin{aligned}\sum_s \mathbb{P}_h[x|s]\mathbb{P}[s] &= \sum_s \mathbb{P}_g\big[\{x_i|i\in I\}|\{s_i|i\in I\}\big]\ \mathbb{P}\big[\{s_i\in I\}\big]\ \prod_{i\in G}\left(\prod_j \mathbb{P}[x^p_{i_0,j}=1]\mathbb{P}[x^m_{i_0,j}=1]\right)\\ &\quad\cdot\left(\prod_k \mathbb{P}[x^p_{i_k}|x^p_{f(i_k)},x^m_{f(i_k)},s^p_{i_k}]\mathbb{P}[x^m_{i_k}|x^p_{m(i_k)},x^m_{m(i_k)},s^m_{i_k}]\mathbb{P}[s^p_{i_k}]\mathbb{P}[s^m_{i_k}]\right).\end{aligned}$$

The sum over vector s can be split into sums over the component pieces. The sums involving the s_{i_k} can be distributed into the product over k, since that is the only place they are used. Let $s_{i_k} = (s^p_{i_k}, s^m_{i_k})$. We easily see that $\mathbb{P}[x^m_{i_k}|x^p_{m(i_k)},x^m_{m(i_k)},s^m_{i_k}]\mathbb{P}[s^m_{i_k}] = 1$, since there are two ways to inherit the 1-allele from the mother, and all of them are compatible.

$$\begin{aligned}\sum_s \mathbb{P}_h[x|s]\mathbb{P}[s] &= \sum_{\{s_i|i\in I\}} \mathbb{P}_g\big[\{x_i|i\in I\}|\{s_i|i\in I\}\big]\ \mathbb{P}\big[\{s_i\in I\}\big]\ \prod_{i\in G}\left(\prod_j \mathbb{P}[x^p_{i_0,j}=1]\mathbb{P}[x^m_{i_0,j}=1]\right)\\ &\quad\cdot\left(\prod_k \sum_{s_{i_k}} \mathbb{P}[x^p_{i_k}|x^p_{f(i_k)},x^m_{f(i_k)},s^p_{i_k}]\mathbb{P}[s^p_{i_k}]\right).\end{aligned}$$

Let $p_t(i)$ be the transmission probability for the proxy family, defined as

$$p_t(i) = \prod_k \sum_{s_{i_k}} \mathbb{P}[x^p_{i_k}|x^p_{f(i_k)},x^m_{f(i_k)},s^p_{i_k}]\mathbb{P}[s^p_{i_k}].$$

View this probability as a Markov chain along the genome with a state space of size 2^4 where each state indicates the inheritance of $(s_{i_1}, s_{i_2}, s_{i_3}, s_{i_4})$. The transition probabilities are given

by $P_{ij} = \theta^{H(i,j)}(1-\theta)^{4-H(i,j)}$ where $H(i,j)$ is the Hamming distance between inheritance states i and j. By design, the transitions allowed by the data have an unusual structure dictated by the heterozygous loci of the proxy parent. Specifically, at a heterozygous locus, there is exactly one inheritance state that satisfies the children's haplotypes. At homozygous loci, all the inheritance states are allowed. So, we compute this probability using the l-state transition probabilities to determine the contribution of a particular stretch of l homozygous loci that are followed by a heterozygous locus. Notice that the heterozygous locus has, as inheritance indicators, either $(0,1,1,0)$ or $(1,0,0,1)$, and these alternate between consecutive heterozygous loci.

Let Q_{0110} be a transition matrix having non-zero recombination probabilities only in column 0110 (i.e. $Q_{0110,i,j} = P_{ij}$ when $j = 0110$). Similarly, let Q_{1001} be a transition matrix with non-zero recombination probabilities only in column 1001. Let h_0 be the number of homozygous loci that begin proxy parent's genotype and let h_j be the number of consecutive homozygous loci after the j'th heterozygous locus where $1 \leq j \leq T_i$ and T_i is the number of heterozygous loci for proxy parent i. Now, we can write the transmission probability in terms of matrix operations

$$p_t(i) = \left(\frac{1}{16}\right) \vec{1} \cdot P^{h_0} \prod_{j=0}^{T_i} \big(Z_j Q_{0110} + (1 - Z_j) Q_{1001}\big) \cdot P^{h_j} \cdot \vec{1}^T$$

where Z_j indicates whether the j'th heterozygous locus has inheritance indicators $(0,1,1,0)$. The column vector of ones at the end simply sums all final state probabilities to obtain the total probability.

Finally, notice that the two heterozygous inheritance states $(0,1,1,0)$ and $(1,0,0,1)$ are arbitrarily labeled. The main feature is that these states alternate at heterozygous loci, and it does not matter which one occurs first. So, we can write

$$p_t(i) = \left(\frac{1}{16}\right) \vec{1} \cdot P^{h_0} \prod_{j=0}^{T_i} \big(O_j Q_{0110} + (1 - O_j) Q_{1001}\big) \cdot P^{h_j} \cdot \vec{1}^T$$

where O_j indicates the event that j is odd. Now we have a quantity that is a function of the genotype data and not dependent on the haplotypes under consideration.

□

Corollary 3. *The mapping yields a haplotype problem instance having a likelihood and maximum probability proportional, respectively, to the likelihood and maximum probability of the genotype instance. Specifically,*

$$\begin{aligned} \sum_x \sum_s \mathbb{P}_h[x|s]\mathbb{P}[s] \;\; = \;\; & \left(\sum_{\{x_i | i \in I\}} \sum_{\{s_i | i \in I\}} \mathbb{P}_g\big[\{x_i | i \in I\} | \{s_i | i \in I\}\big] \; \mathbb{P}\big[\{s_i \in I\}\big] \right) \\ & \cdot \prod_{i \in G} p_t(i) \prod_j \mathbb{P}[x^p_{i_0,j} = 1] \mathbb{P}[x^m_{i_0,j} = 1] \end{aligned}$$

and

$$\max_x \sum_x \mathbb{P}_h[x] \quad = \quad \left(\max_{\{x_i | i \in I\}} \mathbb{P}_g[\{x_i | i \in I\}] \right) \cdot \quad \prod_{i \in G} p_t(i) \prod_j \mathbb{P}[x^p_{i_0,j} = 1] \mathbb{P}[x^m_{i_0,j} = 1]$$

where $p_t(i)$ is proxy family i's transmission probability as defined in Lemma 2.

Proof of Proportional Likelihood and Maximum Probability. Lemma 2 shows that X is independent of the coefficient of proportionality between the haplotype probability and the genotype probability. Therefore, this coefficient factors out of both the likelihood and the maximum probability equations. □

References