Conditional Probability Methods for Haplotyping in Pedigrees

Guimin Gao,* Ina Hoeschele,*1 Peter Sorensen† and Fengxing Du‡

*Virginia Bioinformatics Institute and Department of Statistics, Virginia Tech, Blacksburg, Virginia 24061, †Department of Animal Breeding and Genetics, Danish Institute of Agricultural Sciences, 8830 Tjele, Denmark and ‡The Monsanto Company, Chesterfield, Missouri 63198

ABSTRACT

Efficient haplotyping in pedigrees is important for the fine mapping of quantitative trait locus (QTL) or complex disease genes. To reconstruct haplotypes efficiently for a large pedigree with a large number of linked loci, two algorithms based on conditional probabilities and likelihood computations are presented. The first algorithm (the conditional probability method) produces a single, approximately optimal haplotype configuration, with computing time increasing linearly in the number of linked loci and the pedigree size. The other algorithm (the conditional enumeration method) identifies a set of haplotype configurations with high probabilities conditional on the observed genotype data for a pedigree. Its computing time increases less than exponentially with the size of a subset of the set of person-loci with unordered genotypes and linearly with its complement. The size of the subset is controlled by a threshold parameter. The set of identified haplotype configurations can be used to estimate the identity-by-descent (IBD) matrix at a map position for a pedigree. The algorithms have been tested on published and simulated data sets. The new haplotyping methods are much faster and provide more information than several existing stochastic and rule-based methods. The accuracies of the new methods are equivalent to or better than those of these existing methods.

HAPLOTYPING in a pedigree refers to the reconstruction of haplotypes from observed genotype data within the pedigree. It is an important computational step in the fine mapping of quantitative trait locus or complex disease genes in animal and human complex pedigrees. For a simple, small pedigree with a small number of linked loci, it is not difficult to enumerate all possible consistent haplotype configurations, calculate their likelihood or conditional probabilities, and identify the most probable configurations (Sobel et al. 1996). A consistent haplotype configuration is an assignment of haplotypes to all individuals in the pedigree, which is consistent with all observed genotype data and the pedigree structure. However, for larger and more complex pedigrees and with a larger number of linked loci, the number of consistent haplotype configurations is usually too large for an exhaustive search to be feasible. Methods that are capable of handling a larger number of loci are not guaranteed to find the most probable configuration (Lin and Speed 1997).

Various computer algorithms and programs for haplotyping in pedigrees have been developed. Some of these algorithms are purely logical and rule based (Wijmman 1987; O’Connell 2000; Tapadar et al. 2000; Qian and Beckmann 2002), while others are based on likelihood or conditional probability computations (Lander and Green 1987; Sobel and Lange 1996; Sobel et al. 1996; Lin and Speed 1997; Thomas et al. 2000; Abecasis et al. 2002). The rule-based procedures are deterministic and fast and hence can be used for large pedigrees with large numbers of linked loci, but they do not use the distance between markers and are not suitable for situations with substantial amounts of missing data or uninformative markers. The likelihood- or conditional probability-based algorithms are typically stochastic (Sobel and Lange 1996; Sobel et al. 1996; Lin and Speed 1997; Thomas et al. 2000), and although they can be applied to complex pedigrees, their computing time requirements may become unacceptable.

In the space of all consistent haplotype configurations on a pedigree (SACHC), typically most configurations have very small probabilities conditional on the observed genotype data, so that only a relatively small subset of configurations is relevant. In this contribution, we provide two new methods, a conditional probability approximation and a conditional enumeration method. The first method produces a single, approximately optimal haplotype configuration. The second method eliminates haplotype configurations with low conditional probabilities from SACHC, identifies a subset of haplotype configurations with high conditional probabilities, ranks the configurations in the subset by their likelihood, and calculates their numbers of recombinants. These methods use the closest informative flanking markers, which may not be the immediate flanking markers of a locus under consideration, and information from close relatives, par-
ents, and offspring. For a large pedigree with a large number of linked loci, both new methods are not guaranteed to find the most probable configuration and the true configuration. Below we first describe our new methods and subsequently compare them with SimWalk2 (Sobel and Lange 1996; Sobel et al. 1996) and the methods of Lin and Speed (1997) and Qian and Beckmann (2002).

METHODS

Notation and definitions: In this article, we assume that all individuals in a pedigree have been genotyped for all markers. The combination of a specific individual and a specific marker locus is termed a person-marker, and each person-marker either is homozygous or has an observed heterozygous genotype with two possible ordered genotypes. A multilocus genotype of an individual with an ordered genotype at each locus is termed a haplotype pair. For each founder in the pedigree, we look for the first heterozygous locus in the linkage group and assign arbitrarily an ordered genotype for this locus. For some descendants in the pedigree, the genotypes at some heterozygous loci can be ordered with certainty conditional on parental genotypes (Pong-Wong et al. 2001; Qian and Beckmann 2002), so that haplotypes can be partially reconstructed with certainty. Our haplotype reconstruction methods are based on this prior, partial reconstruction (PPR). The observed data in a pedigree after PPR are denoted by D.

A person-marker with an observed, unordered, heterozygous genotype is referred to as an unordered person-marker. For a pedigree with N members and a set of L linked marker loci, there are k unordered person-markers (U₁, . . . , U₁^k) in individual i after PPR, where U₁^j denotes the jth unordered person-marker in individual i. The set of all unordered person-markers in a pedigree is denoted by U = {U₁, . . . , U₁^j, . . . , U₁^k}. To reconstruct the haplotype configuration for the entire pedigree, one needs to reconstruct only the haplotype configuration for the person-markers in U.

Let \( t = \sum_{i=1}^{n} k \) denote the total number of unordered person-markers in the pedigree, let \{M₁, M₂, . . . , M₁\} be a specific ordering of the person-markers in U, and let \( m = a_j/ \alpha_2 \) denote an ordered genotype for person-marker \( M_j \), where allele \( a_j \) is of paternal origin. The joint probability of a haplotype configuration for the person-markers in U, \( \{m, m_2, . . . , m_1\} \), conditional on the pedigree data after PPR (D), can be factored using the method of decomposition for a given order, or

\[
\Pr(m_1, m_2, . . . , m_n|D) = \Pr(m_1|m_2, m_3, . . . , m_n)|D) \Pr(m_2|m_1, m_3, . . . , m_n), \quad . . . \Pr(m_n|m_1, m_2, . . . , m_{n-1}). \quad (1)
\]

Conditional probability approximation: Sobel et al. (1996) proposed a conditional probability method for haplotyping in a pedigree to identify an approximately optimal haplotype configuration from SACHC, by assigning a haplotype pair to each individual in the pedigree sequentially in a given order, on the basis of the conditional probabilities. This method is designed for pedigrees of small or moderate size and a small number of linked loci. We use a similar, sequential approach to reconstruct a haplotype configuration for the person-markers in U, and we extend it to larger pedigrees and larger numbers of loci by an approximation method.

We assign an ordered genotype to each unordered person-marker in U sequentially in a given order \{M₁, M₂, . . . , M₁\}. We note that this is different from assigning an optimal haplotype pair to each individual. The order in which the assignment occurs is termed a reconstruction order. After the first \( i = 1 \) person-markers have been assigned the set of ordered genotypes \( m_1, m_2, . . . , m_{i-1} \), the two ordered genotypes at person-marker \( M_i \) are compared by means of their conditional probabilities \( \Pr(m_i|m_1, m_2, . . . , m_{i-1}, D) \). The \( m_i \) with the higher conditional probability is assigned to person-marker \( M_i \). Concatenating these conditionally optimal ordered genotypes provides an approximately optimal haplotype configuration for the \( t \) person-markers in U and hence for the entire pedigree.

Consequently, at each step of the algorithm, the conditional probabilities of the ordered genotypes at a person-marker, \( \Pr(m_i|m_1, m_2, . . . , m_{i-1}, D) \), must be calculated. Exact calculation will be time consuming and is therefore approximated. The conditional probability method with the approximation is termed the conditional probability approximation. The approximation is described next.

Assume a given reconstruction order \{M₁, M₂, . . . , M₁\}, let \( M_i \) represent the person-marker for locus \( j \) and individual \( k \), and assume that the first \( i = 1 \) person-markers have been assigned the ordered genotypes \( m_1, m_2, . . . , m_{i-1} \). Also, let \( V_k \) denote the subset of loci, which contains locus \( j \) and all loci with known ordered genotypes in individual \( k \) after the assignment of ordered genotypes to the first \( i = 1 \) person-markers. Then conditional probabilities at person-marker \( M_k \), \( \Pr(m_i|m_1, m_2, . . . , m_{i-1}, D) \), can be calculated approximately by conditioning only on marker information at loci in \( V_k \) from individual \( k \) and those of its “close” relatives, parents, and offspring, which have a known, ordered genotype at locus \( j \), or

\[
\Pr(m_i|m_1, m_2, . . . , m_{i-1}, D) \approx \Pr(m_i|G_{k_0}, G_{k_1}, H_{al}), \quad (2)
\]

where \( G_{k_0}, G_{k_1}, \) and \( H_{al} \) are the partially ordered multilocus genotypes at the loci in \( V_k \) of individual \( k \), its father and its mother, respectively, and \( H_{al} \) is the set of partially known haplotypes, inherited from individual \( k \), by all of its offspring, all after PPR and assignment of \( \{m_1, m_2, . . . , m_{i-1} \} \) to person-markers \{M₁, M₂, . . . , M₁\}. If any of the relatives of individual \( k \) has an unordered genotype at locus \( j \) after the assignment of ordered genotypes to the first \( i = 1 \) person-markers, then this relative is dropped.
from Equation 2. If all close relatives of individual \( k \) have unordered genotypes at locus \( j \), then \( \Pr(m_1 | m_2, \ldots, m_{j-1}, D) = 0.5 \). A partially ordered multilocus genotype of an individual has a known (fixed) ordered genotype at some loci (the genotype list of this individual contains only one element for these loci), while it has two possible ordered genotypes at some other loci (the genotype list contains two elements for these loci). A partially known haplotype has known parental origin at some loci (the allele list of this haplotype contains only one allele for these loci), while it has unknown parental origin at some other loci (the allele list contains two alleles for these loci).

At the loci in \( V_i \), \( G_k \) has only one unordered person-marker \( M_i \) (at locus \( j \)), and the two ordered genotypes at person-marker \( M_i \) are denoted by \( m_i^1 \) and \( m_i^2 \). Let \( G_{i}^1 \) be the corresponding multilocus genotypes obtained from \( G_k \) with ordered genotype \( m_i^1 \) (\( m_i^2 \)) at locus \( j \). Then

\[
\Pr(m_i^1 | G_k, G_i, G_m, H_{all}) = \frac{\Pr(G_i | G_k, G_i, G_m, H_{all}) \Pr(m_i^1 | G_i)}{\sum_{m_i^1} \Pr(G_i | G_k, G_i, G_m, H_{all}) \Pr(m_i^1 | G_i)}
\]

(3)

A marker is informative for an individual and one of its parents, if the ordered genotype is known for both relatives, and if the parent is heterozygous. After the assignment of \( [m_1, m_2, \ldots, m_{j-1}] \) to person-markers \( [M_1, M_2, \ldots, M_{j-1}] \), for the person-marker \( M_1 \) in individual \( k \), let \( M_{i1}^o \) and \( M_{i1}^p \) denote the closest informative flanking markers for individual \( k \) and its father; let \( M_{i2}^o \) and \( M_{i2}^p \) be defined analogously for individual \( k \) and its mother; and let \( M_{i1}^o \) and \( M_{i1}^p \) be defined similarly for individual \( k \) and its offspring \( o \). Let \( G_{i1}^o \) be the subset of ordered genotypes at loci \( M_{i1}^o, M_{i2}^o, \) and \( M_{i2}^p \) in the multilocus genotype \( G_k^i \) of individual \( k \) \((j = 1, 2)\); let \( G_{i1}^p \) and \( G_{i2}^p \) be the paternal and maternal haplotypes at loci \( M_{i1}^o, M_{i2}^o, M_{i2}^p, M_{i1}^p \) at \( M_{i1}^o, M_{i2}^o, M_{i2}^p, M_{i1}^p \), respectively, in the multilocus genotype \( G_k^i \) of individual \( k \); let \( H_{all}^{o}(M_{i1}^o, M_{i2}^o, M_{i2}^p) \) be the haplotype containing only loci \( M_{i1}^o, M_{i2}^o, M_{i2}^p \) inherited by offspring \( o \) from individual \( k \); let \( G_{i1}^o(M_{i1}^o, M_{i2}^o, M_{i2}^p) \) and \( G_{i2}^p(M_{i1}^o, M_{i2}^o, M_{i2}^p) \) denote the ordered genotypes at loci \( M_{i1}^o, M_{i2}^o, M_{i2}^p \) and at \( M_{i1}^o, M_{i2}^o, M_{i2}^p \) in the multilocus genotypes \( G_i \) of the father and \( G_m \) of the mother, respectively. From Equation 3 we have

\[
\Pr(m_i^1 | G_k, G_i, G_m, H_{all}) = \frac{\Pr^*(H_i^o | G_i) \Pr^*(H_i^p | G_m) [\Pr^*(H_o | G_i) \Pr^*(H_i^o | G_i)]}{\sum_{m_i^1} \Pr^*(H_i^o | G_i) \Pr^*(H_i^p | G_m) [\Pr^*(H_o | G_i) \Pr^*(H_i^o | G_i)]},
\]

(4)

where

\[
\Pr^*(H_i^o | G_i) = \Pr(H_i^o(M_{i1}^o, M_{i2}^o, M_{i2}^p) | G_i(M_{i1}^o, M_{i2}^o, M_{i2}^p)),
\]

\[
\Pr^*(H_i^p | G_m) = \Pr(H_i^p(M_{i1}^o, M_{i2}^o, M_{i2}^p) | G_m(M_{i1}^o, M_{i2}^o, M_{i2}^p))
\]

\[
\Pr^*(H_o | G_i) = \Pr(H_o(M_{i1}^o, M_{i2}^o, M_{i2}^p) | G_i(M_{i1}^o, M_{i2}^o, M_{i2}^p)), \quad n = 1, 2.
\]

For example, \( \Pr(H_i^o(M_{i1}^o, M_{i2}^o, M_{i2}^p) | G_i(M_{i1}^o, M_{i2}^o, M_{i2}^p)) \) is the transmission probability of haplotype \( H_i^o(M_{i1}^o, M_{i2}^o, M_{i2}^p) \) by genotype \( G_i(M_{i1}^o, M_{i2}^o, M_{i2}^p) \).

If either of the two closest informative flanking markers for one of the close relatives and individual \( k \) does not exist, then this flanking marker is dropped from the corresponding transmission probability in Equation 4; if there are no informative flanking markers for a close relative and individual \( k \), then this relative is dropped from Equation 4; and if there are no informative flanking markers for any of the close relatives and individual \( k \), then

\[
\Pr(m_i^1 | G_k, G_i, G_m, H_{all}) = 0.5.
\]

To illustrate the approximation method, we use a three-generation pedigree with partially ordered genotypes at seven markers in eight individuals as depicted and explained in Figure 1.

We consider calculating the conditional probabilities of the ordered genotypes at locus 3 in individual 4. For this calculation, only the marker information of the parents (1, 2) and offspring (6, 7) at the loci in \( V_i \) is
used, where \( V_1 = [1, 2, 3, 5, 7] \). The genotype data in boxes are discarded temporarily for calculating the conditional genotype probabilities at this person-marker (locus 3, individual 4), but may of course be used for other person-markers (e.g., locus 5 in individual 6). Marker information at loci 4 and 6 are not used, because individual 4 is not ordered at these loci, and the marker data of individual 8 are not used because this offspring has an unordered genotype at locus 3. In individuals 6 and 7, only the haplotypes inherited from individual 4 are used. At locus 5 of individual 6, for example, the genotype is not ordered, and hence the paternal haplotypes are used.

Markers 2 and 4, the closest informative flanking markers are markers 1 and 5 for individual 4 and its father, markers 2 and 5 for individual 4 and its mother, and markers 2 and 7 for individual 4 and offspring 6.

A problem with the conditional probability approximation method is that the reconstruction order can greatly influence the conditional probability of the reconstructed haplotype configuration. Therefore, it is important to determine an approximately optimal reconstruction order and to execute the conditional probability method with this order. Because we identify an ordered genotype for each person-marker in \( U \) on the basis of the information from the individual under consideration and its close relatives, we assign ordered genotypes earlier to those person-markers that have more information in the individual and its close relatives than to others. Thus, the conditional probability method is implemented with the following steps:

1. Calculate the probabilities of the ordered genotypes for each person-marker in \( U \) conditional on the data \( D \), retain the larger probability for each person-marker, and determine the person-marker with the highest probability. This person-marker is set to \( M_1 \). Then the ordered genotype with higher probability, \( m_1 \), is assigned to \( M_1 \).

2. Calculate the probabilities of the ordered genotypes for all remaining person-markers in the subset of \( U \) not used in step 1, \( U - \{M_1\} \), conditional on \( m_1 \) and \( D \), retain the larger probability for each person-marker, and determine the person-marker with the ordered genotype \( m_2 \) with the highest probability. This person-marker becomes \( M_2 \), and the ordered genotype \( m_2 \), having the larger probability of the two choices, is assigned to \( M_2 \).

3. Similarly, order and assignment for the remaining person-markers are determined, resulting in an approximately optimal order \( M_1, \ldots, M_t \) and a corresponding assignment \( m_1, m_2, \ldots, m_t \).

To illustrate how to reconstruct haplotypes for a pedigree with the conditional probability approximation in an approximately optimal order and to show the influence of reconstruction order on the result, we simulated a haplotype configuration of a two-generation pedigree with eight markers in six individuals, which is depicted in Figure 2. The person-markers requiring reconstruction; (..) an unordered genotype. For the ordered genotypes, the left allele is of paternal origin.

![Simulated pedigree data for illustration of influence of reconstruction order on the haplotype reconstruction result.](image)

**Figure 2.**—Simulated pedigree data for illustration of influence of reconstruction order on the haplotype reconstruction result. The distance between adjacent markers is 1 cM. (*) The person-markers requiring reconstruction; (..) an unordered genotype (the other genotypes are ordered). For the ordered genotypes, the left allele is of paternal origin.
Conditional Probability Haplotyping

TABLE 1

Haplotype reconstruction in an approximately optimal order \( (M_1, M_2, M_3, M_4) \) with the conditional probability approximation

<table>
<thead>
<tr>
<th>Larger conditional probability for each person-marker</th>
<th>Optimal reconstruction order and assigned ordered genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Pr(U^i_2 = 1/2</td>
<td>D) = 0.961 ), ( \Pr(U^i_1 = 1/2</td>
</tr>
<tr>
<td>( \Pr(U^i_2 = 2/1</td>
<td>D) = 0.980 ), ( \Pr(U^i_1 = 1/2</td>
</tr>
<tr>
<td>( \Pr(U^i_2 = 1/2</td>
<td>U^i_1 = 1/2, D) = 0.5 ), ( \Pr(U^i_1 = 1/2</td>
</tr>
<tr>
<td>( \Pr(U^i_2 = 2/1</td>
<td>U^i_1 = 1/2, D) = 0.981 ), ( \Pr(U^i_1 = 1/2, U^i_2 = 2/1, D) = 0.980 )</td>
</tr>
<tr>
<td>( \Pr(U^i_2 = 2/1</td>
<td>U^i_1 = 1/2, U^i_2 = 2/1, D) = 0.998 )</td>
</tr>
</tbody>
</table>

TABLE 2

Haplotype reconstruction in an arbitrary order, different from that in Table 1, with the conditional probability approximation

<table>
<thead>
<tr>
<th>Larger conditional probability for each person-marker</th>
<th>Reconstruction order and assigned ordered genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Pr(U^i_2 = 1/2</td>
<td>D) = 0.961 )</td>
</tr>
<tr>
<td>( \Pr(U^i_1 = 1/2</td>
<td>U^i_2 = 1/2, D) = 0.667 )</td>
</tr>
<tr>
<td>( \Pr(U^i_2 = 2/1</td>
<td>U^i_1 = 1/2, U^i_2 = 1/2, D) = 0.667 )</td>
</tr>
<tr>
<td>( \Pr(U^i_2 = 1/2</td>
<td>U^i_1 = 1/2, U^i_2 = 2/1, D) = 0.667 )</td>
</tr>
</tbody>
</table>

...
creasing $\lambda$ value. When $\lambda = 1$, this method becomes the exhaustive enumeration method (Sobel et al. 1996), and SACHC = SACHC*. The exhaustive enumeration method is exact, but it becomes computationally expensive or infeasible for large pedigrees with large numbers of loci. When $\lambda = 0.5$, the conditional enumeration method becomes the conditional probability approximation method, which is always fast computationally, but it provides only a single approximately optimal haplotype configuration. The SACHC* identified by the conditional enumeration method ($0.5 < \lambda < 1$) always contains the single, approximately optimal haplotype configuration identified by the conditional probability method ($\lambda = 0.5$). The SACHC may contain a subset of haplotype configurations all having the same and highest conditional probability. The conditional probability method cannot identify this subset, but the conditional enumeration method can identify all or most of the configurations in this subset. The number of haplotype configurations retained in SACHC*, the accuracy, and the computing time for the conditional enumeration method can be controlled with the value of $\lambda$ (see below).

RESULTS FOR A PUBLISHED DATA SET

To compare our methods with existing methods, we analyzed a published data set, which was previously analyzed by Sobel et al. (1996) and Lin and Speed (1997). The data came from a medical genetic study of the Krabbe disease by Oehlmann et al. (1993) in a pedigree of nine individuals with unordered genotypes at eight polymorphic markers on chromosome 14, with marker order D14S47, D14S52, D14S43, D14S53, D14S55, D14S48, D14S45, and D14S51, and with distances between adjacent markers of 12.5, 22.3, 3.3, 8.4, 1.9, 18.8, and 1.5 CM.

The approximately optimal haplotype configuration identified with the conditional probability approximation ($\lambda = 0.5$) is shown in Figure 3 and is identical to the most likely configuration obtained by Sobel et al. (1996) and Lin and Speed (1997). The computing time of the conditional probability approximation was $<1$ sec on 2.00 GHz Intel Xeon CPU (1.047,546 kB RAM; Microsoft Windows 2000).

After PPR, SACHC had a total of 32,768 configurations. Using the exhaustive enumeration method ($\lambda = 1$, exact method), configurations were ranked by their likelihood from highest (rank 1) to lowest (rank 32,768) in ~4 min of CPU time. The sums of the conditional probabilities of the 10, 50, 100, and 500 top configurations are 0.95172, 0.99839, 0.99988, and $\approx 1.0$, respectively. The conditional probability of each configuration from rank 51 to rank 32,768 was $<0.00011$, and the sum of the conditional probabilities of these configurations was only 0.00161. The 10 top configurations in SACHC are listed in Table 3.

FIGURE 3.—The most likely haplotype configuration for the Krabbe disease pedigree (nine individuals, eight marker loci). The left allele at any locus in any individual is of paternal origin.

Using the conditional enumeration method with $\lambda = 0.995$, SACHC* contained 128 haplotype configurations, which were selected from the 32,768 configurations in SACHC. The computation time was ~2 sec, much less than that of the exhaustive enumeration method ($\lambda = 1$). The total conditional probability of the 128 configurations in SACHC* accounted for 99.89% of the total probability of all the configurations in SACHC, and the SACHC* contained all of the 50 top configurations in SACHC, except for the two configurations ranked 36 and 39 (with conditional probabilities of 0.00034 and 0.00029, respectively). The 10 configurations with the highest likelihoods and conditional probabilities in SACHC* were the same as the 10 top configurations in SACHC listed in Table 3.

Configuration 1 in Table 3 is the most likely configuration as confirmed by the exhaustive search. Column 2 in Table 3 shows the differences between configurations identified by the conditional enumeration method and the most likely configuration. For example, [2, 7] denotes that the corresponding configuration has a different ordered genotype at marker 7 of individual 2. Column 5 presents the conditional probabilities estimated by the ratio of the likelihood of the corresponding configuration to the sum of the likelihoods of all configurations.
TABLE 3
Haplotype configurations with highest likelihoods identified by the conditional enumeration method and numbered from 1 to 10 by decreasing likelihood or conditional probability

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Difference from the most likely configuration</th>
<th>Likelihood</th>
<th>No. of recombinants</th>
<th>Probability (%)^a</th>
<th>λ = 0.995</th>
<th>λ = 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>None</td>
<td>6.27 × 10^{-12}</td>
<td>6</td>
<td>57.27</td>
<td>57.21</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>[2, 7], [2, 8]</td>
<td>1.28 × 10^{-12}</td>
<td>7</td>
<td>11.69</td>
<td>11.68</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>[8, 7]</td>
<td>1.16 × 10^{-12}</td>
<td>7</td>
<td>10.64</td>
<td>10.63</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>[8, 6], [8, 7]</td>
<td>6.43 × 10^{-12}</td>
<td>7</td>
<td>5.88</td>
<td>5.87</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>[4, 1]</td>
<td>2.61 × 10^{-15}</td>
<td>8</td>
<td>2.38</td>
<td>2.38</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>[2, 7], [2, 8], [8, 7]</td>
<td>2.38 × 10^{-15}</td>
<td>8</td>
<td>2.17</td>
<td>2.17</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>[2, 3], [2, 4], [2, 5], [2, 7], [2, 8]</td>
<td>2.35 × 10^{-15}</td>
<td>9</td>
<td>2.14</td>
<td>2.14</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>[2, 7], [2, 8], [8, 6], [8, 7]</td>
<td>1.31 × 10^{-15}</td>
<td>8</td>
<td>1.20</td>
<td>1.20</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>[8, 6]</td>
<td>1.20 × 10^{-15}</td>
<td>8</td>
<td>1.09</td>
<td>1.09</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>[2, 3], [2, 4], [2, 5], [2, 7], [2, 8]</td>
<td>8.73 × 10^{-14}</td>
<td>9</td>
<td>0.80</td>
<td>0.80</td>
<td></td>
</tr>
</tbody>
</table>

^a For example, [2, 7] denotes that the corresponding configuration has an ordered genotype at marker 7 of individual 2, which differs from that in the most likely configuration.

^b The conditional probability estimated by the ratio of the likelihood to the sum of the likelihoods of all configurations in SACHC*. When λ = 1.0, the conditional probabilities are calculated exactly.

Simulations in SACHC*. When λ = 1, the conditional probabilities are calculated exactly.

Using their Gibbs-Jump method, LIN and SPEED (1997) identified their most and second-most likely configurations with estimated conditional probabilities of 0.69 and 0.15, respectively. These two configurations are identical to configuration 1 (with highest likelihood) and configuration 3 (with third-highest likelihood) as described below. Using their rule-based minimum-recombinant haplotyping method, QIAN and BECKMANN (2002) would identify only configuration 1, which has the minimum number of recombinants. Moreover, several configurations in Table 3 have eight recombinants, but different likelihoods, and recombinant counting cannot discriminate among these configurations. The optimality criteria based on likelihood and recombinant count are equivalent, or produce the same ranking of the configurations, only if the distances among adjacent markers are all equal and if all markers are informative.

SIMULATION STUDIES

To further evaluate the performance of our new methods and to compare their results and computing times with those of SIMWALK 2 (SOBEL and LANGE 1996; SOBEL et al. 1996), we simulated several pedigrees as described below.

Data simulation: Founder haplotypes were generated assuming Hardy-Weinberg equilibrium within and linkage equilibrium between loci. Haplotypes for nonfounders were then simulated conditional on their parental haplotypes, assuming Haldane’s no interference mapping function.

The first pedigree had 88 members (20 founders) over five generations. The linkage group consisted of 20 biallelic markers with allele frequency of 0.5 and with a distance between adjacent markers of 1 cM. Each parent had one spouse, and each full-sib family had two children. The likelihood of the simulated haplotype configuration was 1.3329 × 10^{-81} and the number of recombinants was 18.

Several additional pedigrees with SNP and microsatellite markers were simulated similarly. The first four pedigrees had 330 members (30 founders) with SNP markers over six generations. The linkage group consisted of 10 biallelic markers. Each parent had two spouses, and each full-sib family had three children. The four pedigrees differed in the allele frequency (0.5 or 0.9) and in the distance between adjacent markers (0.5, 1, and 3 cM). The fifth (sixth) pedigree had 546 members and 42 founders over seven generations with SNP markers (microsatellite markers having 10 alleles each) and the same structure as the 330-member pedigrees, but intermarker distance was 1 cM, frequency for each allele was 0.5 (0.1), and the linkage group consisted of 40 markers.

Results for simulated pedigrees: Table 4 presents the (approximately) optimal haplotype configurations identified by our new methods and SIMWALK2 from the analysis of the 88-member pedigree. Each of these con-
TABLE 4
Comparison of the haplotype reconstruction results of the new methods and SIMWALK2 from the analysis of the 88-member pedigree

<table>
<thead>
<tr>
<th>Method</th>
<th>No. of optimal configurations identified*</th>
<th>Size of SACHC*</th>
<th>Time (hr:min:sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIMWALK2</td>
<td>1 (No)</td>
<td>41:39:35</td>
<td></td>
</tr>
<tr>
<td>Conditional probability</td>
<td>0.50</td>
<td>1 (No)</td>
<td>41:39:35</td>
</tr>
<tr>
<td>Conditional enumeration</td>
<td>0.965</td>
<td>64 (No)</td>
<td>00:00:01</td>
</tr>
<tr>
<td></td>
<td>0.990</td>
<td>64 (No)</td>
<td>00:00:06</td>
</tr>
<tr>
<td></td>
<td>0.995</td>
<td>128 (Yes)</td>
<td>00:00:07</td>
</tr>
<tr>
<td></td>
<td>0.997</td>
<td>128 (Yes)</td>
<td>02:27:45</td>
</tr>
</tbody>
</table>

*Each of the (approximately) optimal configuration(s) identified by any of the methods had the same likelihood as the true configuration. This likelihood is the estimated (approximately) highest likelihood over all configurations in SACHC. The true highest likelihood is unknown (exhaustive search not feasible).

Figurations had the same likelihood as the true (simulated) configuration. This likelihood is the estimated (approximately) highest likelihood over all configurations in SACHC. The true highest likelihood is unknown (exhaustive search not feasible).

SIMWALK2 had to be restarted several (three) times before it identified a (single) approximately optimal configuration, whose likelihood had the same value as the likelihood of the true configuration. The likelihoods of the configurations obtained prior to the final run were much lower, and the corresponding recombinant counts were higher. The two new methods identified configurations with the same likelihood as the true configuration in just one run and were much faster than SIMWALK2. For 0.965 \( \leq \lambda \leq 0.997\), the conditional enumeration method identified a subset of 64–128 haplotype configurations with likelihood and recombinant counts equal to those of the true configuration. To identify such a subset with SIMWALK2, the required multiple starts and restarts would probably take several weeks, instead of seconds and minutes with the conditional enumeration method. For \( \lambda > 0.995\), the number of configurations with the estimated highest likelihood equal to that of the true configuration did not increase, but the number of configurations in SACHC* and the computing time increased rapidly.

Results of the conditional enumeration method for the four 330-member pedigrees and the 546-member pedigrees are presented in Table 5. For each of the six pedigrees, the conditional probability approximation identified an approximately optimal haplotype configuration with approximately highest likelihood in <1 sec. When the distance between adjacent markers was 1 or 0.5 cM, the haplotype configuration identified by the conditional probability approximation had likelihood equal to or higher than that of the true configuration and the same recombinant count as the true configuration. When the distance was 3 cM, the identified configuration had a higher likelihood and a lower number of recombinants than the true configuration. As noted earlier, any SACHC* identified with the conditional enumeration method contains the single, approximately optimal haplotype configuration identified with the conditional probability method.

For each pedigree, Table 5 presents results for several (ascending) \( \lambda \) values. In particular, column 7 contains the ratio of the sums of the likelihoods of all configurations in two different SACHC* corresponding to two different \( \lambda \) values (current row over previous row). For example, for the first 330-member pedigree (intermarker distance 0.5 cM, 1.045 is the ratio of the total likelihood for the SACHC* corresponding to \( \lambda = 0.995\) to the total likelihood for the SACHC* corresponding to \( \lambda = 0.965\). The results in column 7 indicate that when the \( \lambda \) value and the size of SACHC* become sufficiently large, an additional increase in \( \lambda \) often results in a very large increase in the size of SACHC* but only a very small increase in the total likelihood. For example, for the second 330-member pedigree (intermarker distance 1 cM, allele frequency 0.5), when \( \lambda \) was raised from 0.990 to 0.995, the size of SACHC* increased 319 times, but the total likelihood increased only by 4.2%. For the fourth 330-member pedigree (intermarker distance 3 cM, allele frequency 0.5), the increase in the total likelihood, due to increasing \( \lambda \) and the size of SACHC*, was higher when compared with the increases for the pedigrees with shorter intermarker distances (1 and 0.5 cM).

The results in Table 5, column 8, show that a few hundred up to 1000 of the top configurations (those with the highest likelihoods) always contained most of the information in SACHC*. The exception was the pedigree with the largest intermarker distance of 3 cM. In this case, determining SACHC* with a larger \( \lambda \) value and/or retaining more than the 1000 top configurations may be beneficial. However, for small intermarker dis-
Table 5 presents the reconstruction results of the conditional enumeration method for the 330-member pedigrees and the 546-member pedigrees with different marker allele frequencies and distances between adjacent markers.

<table>
<thead>
<tr>
<th>N (type)</th>
<th>I</th>
<th>cM</th>
<th>Freq</th>
<th>λ</th>
<th>Size of SACHC*</th>
<th>Ratio of likelihoods of 50/100/500/1000 top configurations (%)</th>
<th>Total probability of 50/100/500/1000 top configurations (%)</th>
<th>Time (hr:min:sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>330 (SNP)</td>
<td>10</td>
<td>0.5</td>
<td>0.5</td>
<td>0.965</td>
<td>64</td>
<td>93.7/—/—/—</td>
<td>89.7/98.3/—/—</td>
<td>0:00:02</td>
</tr>
<tr>
<td></td>
<td>0.995</td>
<td>256</td>
<td>1.045</td>
<td>88.3/96.8/99.9/99.99</td>
<td>0:23:25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.996</td>
<td>40,960</td>
<td>1.015</td>
<td>92.6/97.6/99.9/99.99</td>
<td>0:00:36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>330 (SNP)</td>
<td>10</td>
<td>1.0</td>
<td>0.5</td>
<td>0.965</td>
<td>64</td>
<td>99.4/—/—/—</td>
<td>99.4/94.1/99.4/99.8</td>
<td>3:03:19</td>
</tr>
<tr>
<td></td>
<td>0.950</td>
<td>1,024</td>
<td>1.088</td>
<td>88.9/94.1/99.4/99.9</td>
<td>0:00:51</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.950</td>
<td>32,768</td>
<td>1.042</td>
<td>7.8/15.6/78.1/99.1</td>
<td>0:01:27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>330 (SNP)</td>
<td>10</td>
<td>1.0</td>
<td>0.9</td>
<td>0.850</td>
<td>1,024</td>
<td>7.2/14.5/72.3/91.7</td>
<td>6.1/12.2/60.8/77.1</td>
<td>0:10:00</td>
</tr>
<tr>
<td></td>
<td>0.950</td>
<td>2,048</td>
<td>1.190</td>
<td>48.5/63.6/96.0/99.4</td>
<td>0:01:46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.950</td>
<td>16,384</td>
<td>1.190</td>
<td>30.2/41.2/71.3/82.7</td>
<td>0:22:32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>330 (SNP)</td>
<td>10</td>
<td>3.0</td>
<td>0.5</td>
<td>0.950</td>
<td>32,768</td>
<td>19.1/26.7/50.3/61.5</td>
<td>19.1/26.7/50.3/61.5</td>
<td>6:01:52</td>
</tr>
<tr>
<td></td>
<td>0.950</td>
<td>589,824</td>
<td>1.599</td>
<td>19.1/26.7/50.3/61.5</td>
<td>6:01:52</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>546 (SNP)</td>
<td>40</td>
<td>1.0</td>
<td>0.5</td>
<td>0.940</td>
<td>768</td>
<td>53.3/70.2/98.5/—</td>
<td>45.6/60.1/87.8/94.7</td>
<td>0:01:36</td>
</tr>
<tr>
<td></td>
<td>0.965</td>
<td>12,288</td>
<td>1.170</td>
<td>42.3/55.6/82.1/89.8</td>
<td>0:24:33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.975</td>
<td>147,456</td>
<td>1.079</td>
<td>99.9/1.0/—/—</td>
<td>4:50:10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>546 (Mic)</td>
<td>40</td>
<td>1.0</td>
<td>0.1</td>
<td>0.975</td>
<td>1</td>
<td>99.9/1.0/—/—</td>
<td>99.9/98.9/99.9/99.99</td>
<td>24:35:49</td>
</tr>
<tr>
<td></td>
<td>0.985</td>
<td>128</td>
<td>1.149</td>
<td>98.1/98.9/99.9/99.99</td>
<td>0:00:23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.991</td>
<td>524,288</td>
<td>1.126</td>
<td>98.1/98.9/99.9/99.99</td>
<td>24:35:49</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **N** denotes the number of individuals in the pedigree, and **type** denotes the marker type, where SNP (Mic) denotes SNP (microsatellite) markers.
- **I** denotes the number of linked marker loci.
- **cM** denotes the distance between adjacent markers.
- **Freq** denotes the allele frequency. For the microsatellite markers, each allele has the frequency 0.1.
- **λ** denotes the ratio of the sums of the likelihoods of all configurations in two different SACHC* corresponding to two different λ values.
- **Time** denotes the sum of likelihoods of top configurations over sum of likelihoods of all configurations in SACHC*.

Finally, Table 5 also shows that when marker allele frequency increased from 0.5 to 0.9 (330-member pedigrees), the increase in the number of homozygous genotypes resulted in a decrease of the information content of the markers and an increase in the computing time of the conditional enumeration method (for the same λ values).

**DISCUSSION**

We have presented a conditional probability approximation and a conditional enumeration method for haplotyping in pedigrees. The conditional probability approximation method identifies a single, approximately optimal (in terms of likelihood or conditional probability) haplotype configuration in a very short run time. For any given threshold λ > 0.5, the conditional enumeration method finds a set of configurations (SACHC*), which always contains a subset of configurations with approximately highest likelihood. In the set SACHC*, some configurations may have very low likelihood, and some configurations with relatively high likelihood may not have been included. For the first problem, a subset of
configurations from SACHC* can be obtained by eliminating the configurations with low likelihood as desired. The second problem can be controlled through the choice of the value for $\lambda$, subject to computational feasibility, as for $\lambda = 1$ no configuration will be missed. In our experience with simulated data (as presented in this article and beyond), when the distance between adjacent markers is $<1.5$ cM, then it is possible to set $\lambda$ to high values (e.g., $\lambda \approx 0.96$) while maintaining computational efficiency. However, when the distance exceeds $2$ cM, setting $\lambda$ to high values may cause a substantial increase in computing time.

Both methods sequentially assign ordered genotypes to the unordered person-markers in a pedigree in approximately optimal orders, by using only the marker information of the individual under consideration and its close relatives (parents and offspring), and conditional on previously assigned ordered genotypes of other person-markers preceding the current one in the order chosen.

In the rule-based method of Qian and Beckmann (2002), the haplotype assignment is based on minimizing the number of recombinants by considering two generations at a time, a nuclear family or a parent-offspring trio. In contrast, our methods are based on maximizing the number of recombinants by considering two person-markers conditional on the marker in the spring trio. In contrast, our methods are based on maximizing the number of recombinants by considering two unordered person-markers preceding the current one in the order for the pedigree at a specific genome location (for definition of IBD matrix, see Pong-Wong et al. 2001). The IBD matrix conditional on the observed data ($\mathbf{D}$) is a weighted average of all IBD matrices, each conditional on a haplotype configuration in SACHC, where the weight of each configuration is the conditional probability of the configuration in SACHC. The IBD matrix conditional on the observed data ($\mathbf{D}$) can be calculated by the expression

$$Q_{D} = \sum_{\omega} Q_{\omega} \cdot \Pr(\omega | \mathbf{D})$$  \hspace{1cm} (5)

(Wang et al. 1995; Hoeschele 2001), where $\omega_{i}$ is a specific haplotype configuration of the pedigree in SACHC, and $Q_{D} (Q_{\omega})$ is the IBD matrix of the pedigree given $\mathbf{D}$ ($\omega$). $\Pr(\omega | \mathbf{D})$ is the probability of $\omega$, conditional on the observed data $\mathbf{D}$. The summation in Equation 5 is over all configurations in SACHC. For a large pedigree with a large number of loci, the IBD matrix $Q_{D}$ can be estimated approximately by Equation 5 with the summation over the subset of configurations identified by the conditional enumeration method, and the probability $\Pr(\omega | \mathbf{D})$ can be estimated approximately by the ratio of the likelihood of $\omega_{i}$ to the sum of the likelihoods of all configurations in the identified subset.

In this article we demonstrated applications of the new methods to pedigrees with up to 546 members and 40 linked loci. We anticipate that pedigrees of several thousand individuals and up to 100 linked loci can be analyzed efficiently by choice of suitable $\lambda$ values without significant loss in accuracy of evaluation of the IBD matrix and QTL localization (work in progress).

In this contribution, we have assumed that all individuals in a pedigree have been genotyped for all markers. We have work in progress on extending our methods to pedigrees with missing marker data (the haplotyping methods presented in this article were implemented in a C++ program, which is available upon request from the first author for academic research).
We thank Keying Ye and Nan Bing for helpful comments and suggestions. This research was supported by grant R01 GM66103-01 from the National Institutes of Health and a grant from the Monsanto Company to I. Hoeschele.

LITERATURE CITED


