Pedigree Data Analysis With Crossover Interference

Sharon Browning

Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina 27695-7566
and GlaxoSmithKline, Research Triangle Park, North Carolina 27709

Manuscript received May 27, 2002
Accepted for publication March 7, 2003

ABSTRACT

We propose a new method for calculating probabilities for pedigree genetic data that incorporates crossover interference using the chi-square models. Applications include relationship inference, genetic map construction, and linkage analysis. The method is based on importance sampling of unobserved inheritance patterns conditional on the observed genotype data and takes advantage of fast algorithms for no-interference models while using reweighting to allow for interference. We show that the method is effective for arbitrarily many markers with small pedigrees.

EXISTING methods for likelihood-based analysis of pedigree genotype data assume absence of crossover interference, even though such interference has been well documented in humans (Weeks et al. 1993; Broman and Weber 2000; Lin et al. 2001) and other species (Zhao et al. 1995b; Broman et al. 2002). Moreover, simulation studies have shown that failure to incorporate crossover interference into linkage analyses is inefficient (Goldgar et al. 1989; Goldstein et al. 1995; Lin and Speed 1999).

The major reason that crossover interference models are not used is the lack of computationally feasible methods for working with more than a handful of linked markers. Lin and Speed (1996) present a method for exact computation of probabilities for pedigree data with the chi-square models; however, computational time increases exponentially with both the number of meioses in the pedigree and the number of markers, severely limiting its applicability. Thompson (2000a,b) has a Markov chain Monte Carlo approach for incorporating crossover interference, but it is limited to at most 12 loci per chromosome.

We propose a new method for calculating probabilities for pedigree genetic data that does incorporate crossover interference. The method is based on importance sampling, which is a Monte Carlo technique for estimating the value of an integral or sum (in this case probabilities of the data, expressed as a sum over unobserved inheritance indicators). Importance sampling involves evaluating the integrand at independently sampled realizations from a probability distribution that is roughly proportional to the integrand. Correct weighting of the sampled values gives an unbiased estimate that converges to the true value of the integral as the number of sample repetitions is increased. Because the samples are independent, the method is not plagued by the difficulties in assessing convergence encountered in use of Markov chain Monte Carlo (MCMC) methods.

The probabilities produced by this method may be used for multipoint linkage mapping, genetic map construction, and relationship inference. The method can be used with arbitrarily many linked markers for small pedigrees. For small pedigrees the method is fast enough that it could be used on a routine basis in analysis of pedigree genetic data.

BACKGROUND

It has long been known that the locations of multiple crossovers on a chromosome at meiosis are not independent, but exhibit crossover interference whereby the existence of a crossover at one location suppresses the occurrence of crossovers in nearby regions. A common approach is to use the Kosambi map function with methods that assume independence between recombination events in nonoverlapping intervals. This approach does not have any real effect, as genetic distances are used only for reporting results, while recombination fractions are used at all steps in the calculations. That is, observed recombination fractions are converted to genetic distances with the map function and are converted back to recombination fractions with the same map function for use in the analysis. As Speed (1996) points out, the only adequate way to incorporate crossover interference is with a point process model. Such an approach can account for nonindependence of recombination between intervals.

A useful family of point process models for crossover interference is the chi-square models. These are renewal models for the occurrence of crossovers on the four-strand chromatid bundle. A parameter $m$ controls the strength of interference: $m = 0$ corresponds to no interference, while $m = 4$ corresponds to approximately the...
level of interference found in humans (Lin and Speed 1996). The genetic distance between crossovers on the four-strand bundle is modeled as the sum of $m + 1$ independent exponential random variables each with rate $2(m + 1)$. This sum follows a gamma, or scaled chi-square, distribution. A gamete receives only one of the four chromatids, and since each crossover involves four chromatids, the probability that a certain crossover on the chromatid bundle is seen in the gamete is one-half. Assuming no chromatid interference (Zhao et al. 1995a), whether or not one crossover on the bundle is transmitted to the gamete is independent of the outcome for other crossovers. McPeek and Speed (1995) and Zhao et al. (1995b) found that the chi-square models provide an excellent fit to available data. Zhao et al. (1995b) provide a method for calculating the probability of a pattern of recombination between consecutive genotypic data, which are used in genetic map (1995b) provide a method for calculating the probabilities of inheritance patterns used to calculate probabilities of inheritance patterns for other crossovers. McPeek and Speed (1995) and Zhao et al. (1995b) found that the chi-square models provide an excellent fit to available data. Zhao et al. (1995b) provide a method for calculating the probability of a pattern of recombination between consecutive markers for a single meiosis, which involves the order of $Lm^2$ calculations for $L$ marker loci. We reproduce the details of this method in the appendix since our method makes use of these probabilities.

In pedigree data, the recombination patterns are generally not directly observed, due to unobserved individuals and insufficiently polymorphic markers. Thus, to calculate probabilities for genotypic data requires summation over possible recombination patterns. For $K$ meioses and $L$ marker loci there are $2^{2L-1}$ possible recombination patterns, so this computation, if performed exactly, rapidly becomes infeasible for increasing numbers of markers and meioses.

When independence (i.e., no-interference) models for recombination are used, exploitation of the independence can reduce the number of computations to the order of $LK2^k$ (Kruglyak and Lander 1998), so that large numbers of markers may be analyzed, while the size of the pedigree (measured by the number of meioses) must be restricted. McPeek and Sun (2000) describe a related approach for the chi-square models; however, the number of computations required by such an approach is proportional to $L4^{L+1}$, so that for moderate values of $m$, calculation is feasible only with extremely small pedigrees. Thus in general to calculate exact probabilities for pedigree data with the chi-square or other interference models involves summation over all $2^L$ recombination patterns, resulting in a computational burden that increases exponentially with both the number of markers and the number of meioses (Lin and Speed 1996). This is the approach taken by Lin and Speed (1999) in their work showing the efficiency gains obtainable using the chi-square model in gene mapping with pedigree data. They were able to work with 12 meioses (a three-generation family with four grandparents, two parents, and four children) and seven marker loci.

To reduce the computational burden while incorporating crossover interference, we propose an importance-sampling approach. This approach takes advantage of the special algorithms for independence models, while using reweighting to allow for interference.

METHODS

We present a method for calculating probabilities for genotypic data under the chi-square models. The method is based on importance sampling of underlying unobserved inheritance patterns. We focus primarily on calculation of the likelihood, which is equal to the probability of the genotypic data under the assumed model. Likelihoods may be used to compare models, for example, to find the most likely genealogical relationship between individuals (Boehnke and Cox 1997). In addition, the methodology presented here may be used to calculate probabilities of inheritance patterns given the genotypic data, which are used in genetic map construction and multipoint linkage analysis, and we give details of those calculations at the end of this section.

Importance sampling of inheritance indicators: For each marker locus $l$ and meiosis $k$, let $X_{kl}$ be a zero-one inheritance indicator. The parent involved in the meiosis has two copies of the DNA, one maternal (0) and the other paternal (1). The indicator $X$ describes which of these two copies was transmitted to the offspring at this locus. If $X_{kl} = X_{k+1,l}$ no recombination occurred on meiosis $k$ over the interval between markers $l$ and $l + 1$, while recombination occurred if $X_{kl} \neq X_{k+1,l}$. The inheritance pattern $X$ represents the inheritance indicators over all meioses and loci.

We can write the probability of the genotype data $Y$ as a sum over inheritance patterns

$$P(Y) = \sum_X P(C(Y|X)P(C(X)),$$

where $P_C$ denotes probabilities under the chi-square crossover model, with a given choice of parameter $m$. Since $P(C(Y|X)P(C(X) = P(C(Y|X)$, only those terms with large values of $P(C(Y|X)$ contribute significantly to the sum. An importance-sampling approach aims to sample with high frequency those terms with significant contribution, while sampling with low frequency the terms with negligible contribution. Ideally we would like to sample from $P(C(Y|X)$, but we have no way to do so directly. Instead we sample from $P(C(Y|X)$, where $P_C$ denotes probabilities under the independence model. The probabilities $P(C(Y|X)$ are sufficiently close to $P(C(Y|X)$ to result in a useful importance sampler.

In what follows we assume that probabilities for genotypes $Y_l$ at a marker $l$ are conditionally independent of genotypes and inheritance patterns at other markers given the inheritance pattern $X_l$ at the marker. This assumption holds if the markers are in linkage equilibrium, which will be approximately true if all the genotyped individuals come from a single homogeneous population and the markers are not too closely spaced.
A consequence of this assumption is that probabilities of genotypes $Y$ given inheritance patterns $X$ do not depend on the crossover model, so $P_c(Y|X) = P(Y|X) = P(Y|X)$. Then $P(Y|X) = P(Y|X) = P(Y|X)/P(X)$ and we can write

$$P_c(Y) = \sum_X P(Y|X)P_c(X)$$

Thus if $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ are sampled from $P_c(X|Y)$, an unbiased estimate of $P_c(Y)$ is given by

$$\hat{P}_c(Y) = \frac{\sum_{i=1}^{n} P(Y|X)P_c(X^{(i)})}{P(X^{(i)})}.$$}

Thompson (2000b, Sect. 8.4) describes how to sample $X$ from $P_c(X|Y)$. Let $X = (X_0, X_2, \ldots, X_n)$ be the inheritance vector at locus $I$. Making use of the hidden-Markov structure (the inheritance vectors $X_0, \ldots, X_i$ form a hidden-Markov model for the genotypes under the independence model), we can recursively sample $X_i$ conditional on the genotype data and then sample $X_{i-1}$ given the realized value of $X_i$ and the genotype data and so on through to $X_1$ to obtain one realization of $X$. The sampling algorithm has computational order $LA^2$, which is the same as for the original algorithm of Lander and Green (1987) for the independence model. We give further details in the appendix.

**Improved performance through resampling:** This method is an example of sequential importance sampling (Liu 2001, Sect. 2.6.3), because the inheritance patterns $X$ are sampled sequentially along each chromosome. Improvement, in terms of reducing the standard error of the estimate for a fixed number of iterations $n$, may be obtained by resampling (Liu 2001, Sect. 3.4.4). The idea is to concurrently sample a number of inheritance patterns $X$. After sampling the inheritance indicators at a couple of loci, write $X_i^{(0)}$ for the partial inheritance pattern in realization $i$. Then we calculate the weight $P_c(X_i^{(0)})/P(X_i^{(0)})$ for each $i$. If this weight is small for a particular $i$ the sample is not worth continuing, and it is replaced by a copy of one of the samples with a larger weight. Sampling continues for further loci (punctuated by further resampling), and samples that had been duplicated tend to differ in the values sampled at the remaining loci. Thus the samples are correlated, but the final weights $P_c(X_i^{(0)})/P(X_i^{(0)})$ tend to be less variable than they would be without resampling, which reduces the standard error of the estimate $\hat{P}_c(Y)$.

We now give more details of the resampling algorithm, in which we apply the method of residual resampling described in Sect. 3.4.4 of Liu (2001). First we set a batch size $B$ and a resampling interval $T$ (choice of these values is discussed below). We start by sampling $X_i^{(0)}$, $X_i^{(2)}, \ldots, X_i^{(2T-1)}$ for $i = 1, 2, \ldots, B$, perform resampling as described below, and then sample $X_i^{(2T-1)}$.

At the $j$th resampling, let $W_i = \sum_{i=1}^{B} w_i^{(j)}$. Start by retaining $k = \lfloor BW_i/W_j \rfloor$ copies of $X_i^{(j)}$, where $\lfloor \cdot \rfloor$ is the floor function giving the largest integer less than or equal to its argument. The number of elements remaining to be resampled is $r = B - \sum_{i=1}^{B} w_i^{(j)}$. Add to the set of retained inheritance patterns $r$ independent random draws from the $X_i^{(j)}$ with probabilities proportional to $W_i/W_j - k$, $i = 1, 2, \ldots, B$. Now reset the weights $w_i^{(j)}$ to $W_i/B$ for $i = 1, 2, \ldots, B$ and continue with the sequential sampling of the inheritance patterns.

Let $J = \lfloor (L - 1)/T \rfloor$ be the total number of resampling points. After the samples are completed, the weight of the $j$th sample $X_i^{(0)}$ is

$$w_i^{(j)} = \frac{P_c(X_i^{(0)})/P_c(X_i^{(j)})}{P(X_i^{(0)})/P(X_i^{(j)})}.$$}

An unbiased estimate of $P(Y)$ is

$$\hat{P}_c(Y) = \frac{1}{n} \sum_{i=1}^{n} w_i^{(j)} P_c(Y).$$}

The values obtained from a single batch are correlated and cannot be used to estimate the standard error of $\hat{P}_c(Y)$. To estimate the standard error we divide our total number of iterations into batches of size $B$. Choice of $B$ has negligible effect on computing time. Small choices of $B$ are less efficient (i.e., give larger standard error) because resampling is less effective with a smaller pool. Thus it is best to choose $B$ to be reasonably large, subject to computer memory limitations and to having a sufficient number of batches to give a good estimate of standard error (say at least 20 batches). We used $B = 100$ in all the examples presented in results.

A couple of issues are involved in best choice of resampling interval $T$. First, resampling involves computational time, and thus frequent resampling can be computationally expensive. The computational time for resampling is essentially independent of the pedigree size; thus the expense of resampling is most notable in computations on small pedigrees and is negligible relative to total computation time for larger pedigrees (such as
the eight-meiosis pedigree for four full-sibs). Second, as the resampling interval increases, the standard error of the estimates tends to increase. This increase in standard error is particularly evident in larger pedigrees, which are also the pedigrees for which resampling is most beneficial. Thus small values of $T$ (say $1 \leq T \leq 5$) are best for large pedigrees, and large values of $T$ (say $T \geq 5$) are best for small pedigrees. We chose to use $T = 5$ for simplicity throughout the examples presented in RESULTS.

We note that it is not necessary (and is very inefficient) to calculate $P_t(X[i])$ and $R(X[i])$ from scratch at each resampling point. For the independence sampler, $P_t(X[i])$ equals $P_t(X[i-1])$ multiplied by the probabilities of the recombination pattern in $X[i-(j-1), j], X[i-(j-1), j]$, ... , $X[i-j+1]$. For the chi-square sampler it is necessary to save an $m$-dimensional vector of partial probabilities from calculation of $P_t(X[i-1])$ for use in calculation of $P_t(X[i])$.

Our implementation of the algorithm described above is available from the author on request.

Probabilities for linkage analysis and map construction: Linkage analysis and map construction are based on probabilities of inheritance indicators $X$ at a locus $l$ given the genotype information (Kruglyak et al. 1996) or the joint probabilities of inheritance indicators $X_i$ and $X_{i+1}$ at two neighboring loci given the genotype information (Lander and Green 1987). Consider estimating

$$P_t(X_i = x|Y) = \sum_x P(x|X, Y)P_t(X|Y)$$

$$\approx \sum_x P(x|X)\frac{P_t(X)}{P_t(X|Y)},$$

where $P(x|X, Y) = P(x|X)$ is one if $X_i = x$, and zero otherwise. An unbiased estimate of this probability is proportional to

$$\frac{1}{n} \sum_{j=1}^n P(x|X^{(j)})\frac{P_t(X^{(j)})}{P(X^{(j)})},$$

where $X^{(j)}, \ldots, X^{(n)}$ are independent realizations from $P_t(X|Y)$ and the constant of proportionality may be found by summing over all possible values of $x$. Probabilities $P_t(X_i = x, X_{i+1} = x_{i+1}|Y)$ may be estimated similarly. Thus these probabilities may be estimated under interference models using the machinery presented here and then used in linkage mapping or genetic map construction.

RESULTS

We present results from analysis of simulated data, to demonstrate the capabilities of the proposed approach.

Our simulated data were designed to represent human chromosome 1, with an estimated genetic length of 2.9 morgans (Broman et al. 1998). The other human chromosomes are shorter, and since our results (not shown) indicate that standard errors tend to be lower with shorter chromosomes because of smaller numbers of crossovers, chromosome 1 represents the “worst case.” We simulated sets of data with marker spacing of 1 cM (291 markers) and 10 cM (29 markers), with single-nucleotide polymorphisms (SNPs; allele frequencies 0.7 and 0.3), and with microsatellites (seven alleles with frequencies 0.4, 0.2, 0.2, 0.05, 0.05, 0.05, 0.05, and heterozygosity 0.75). Calculation was performed under the chi-square model with $m = 4$ [giving $P_t(Y)$] and, for comparison, under the independence model with the Kosambi map function [giving $P_t(Y)$]. We look at three relationships: two half-sibs (two meioses), an aunt-niece pair (five meioses), and four full-siblings (eight meioses). The data were simulated under the chi-square model with $m = 4$, although this choice has little impact on the results. Table 1 shows results and computing times for one simulated chromosome for each map/relationship combination.

From the results in Table 1, we see that the type of marker has little impact on the computing time or standard errors, but does affect the magnitude of the likelihoods. Marker spacing affects computing time, but has little impact on standard errors. The most computing-intensive part of the algorithm is sampling of the inheritance patterns from the distribution $P_t(X|Y)$, which is of computational order $2^{2^k}$, so that computing time doubles for each additional meiosis. Hence, with the current implementation, eight meioses is about the maximum that one would want to work with—$10^5$ iterations for chromosome 1 with a 1-cM map took $\sim 1$ hr (fewer iterations and hence shorter computing time would suffice in some cases). Standard errors also increase with the number of meioses, so one would generally want to increase the number of iterations as pedigree size increases.

In considering whether an estimate of $\ln P_t(Y)$ is sufficiently precise, one minimal requirement is that the standard error be less than the difference between $\ln P_t(Y)$ and $\ln R(Y)$, where $\ln P_t(Y)$ is the natural log of the probability of the data under the independence model using the Kosambi map function. If this requirement is not satisfied, then the exact answer obtained under the independence model is a better estimate of $P_t(Y)$ than that from the importance sampler. We note that the requirement is satisfied in all but 1 of the 12 examples looked at, and in most of the examples a much smaller number of iterations would have sufficed to meet this minimal requirement (in fact, $< 100$ iterations, at approximately one-thousandth of the computing time shown, would have sufficed in 7 out of 12 of the examples). Thus $10^5$ iterations are more than enough (based on this criterion) for this chromosome length, choice of model parameter $m$, and pedigree size, but more iterations may be required for longer chromosomes, larger pedigrees, or different values of $m$. We found that resampling was very helpful in reduc-
ising the amount of computing time required to achieve a given standard error, with benefit increasing with pedigree size. For the aunt-niece and four full-sib relationships, savings in computational time due to the resampling were at least twofold and typically around fivefold.

**DISCUSSION**

We have presented an algorithm for calculating probabilities for pedigree genetic data under the chi-square family of interference models. The method is based on importance sampling and thus gives an approximate calculation with precision depending on the number of importance sampling iterations. The results shown indicate that the algorithm is feasible for pedigrees with up to eight meioses and for any number of markers. Calculated probabilities of the data may be used as likelihoods for relationship inference, by repeating the calculation with differing pedigrees. In addition, the approach may be used to calculate probabilities of inheritance patterns given the data for linkage gene mapping.

Computational time is higher than that under a no-interference assumption—in general, the computing time without interference would be approximately the time to perform one iteration of the importance sampler used here; however, the important point is that the computation times for this method scale the same with increasing marker density and pedigree size as do those for the basic no-interference algorithm of Lander and Green (1987), so that with increasing computing speed, the size of the problem that can be addressed with interference will lag only a couple of steps behind that for no interference.

We have presented this algorithm for the chi-square model only, but it is generalizable to other classes of models. To directly apply the approach here, it is necessary to be able to calculate the probability of a pattern of recombinations over a series of markers. Such calculation may not be computationally feasible for some models. A more natural approach would involve sampling of not only the recombination pattern but also the actual locations of crossovers on the chromatid bundle under the independence model and conditional on the genotype data \( Y \). Such an approach has the potential to be very flexible, although it will add some noise, resulting in lower precision for a given number of iterations of the importance sampler.

The author thanks Reinhard Hopperger for programming assistance and Terry Speed and the referees for helpful comments. This work was supported in part by a Research Starter Grant in Informatics from the PhRMA Foundation.

**LITERATURE CITED**


Goldstein, D. R., H. Zhao and T. P. Speed, 1995 Relative efficien-


APPENDIX

Probability of recombination pattern for chi-square models: We reproduce the method given in Zhao et al. (1995b) for calculating the probability of a pattern of recombination between consecutive markers for a single meiosis. Following their notation, let \( p = m + 1 \) and \( y = 2px \), and let \( D(y) \) be the \( p \times p \) matrix with the \( i, j \)th entry \( e^{-(y)^{(i,j-i)/2}}/(ps + j - i)! \) (but with zero entries where \( ps + j - i < 0 \)). The \( i, j \)th entry of \( D(y) \) represents the probability that over genetic distance \( x \), \( s \) crossovers occurred on the four-strand bundle, and the number of intermediate events (the events implied by the sum of \( m + 1 \) independent exponential random variables) since the last crossover changed from \( i \) to \( j \) (hence the zero terms, since \( j \) cannot be less than \( i \) if \( s = 0 \) crossovers occurred). For interval \( l \) between markers \( l \) and \( l + 1 \), let \( x_l \) be the genetic distance of the interval, \( y_l = 2(m + 1)x_l \) and define \( N_l = D_l(y_l) + \frac{1}{2} \Sigma s \in 1 \Sigma s \neq 1 D_s(y_l) \) and \( R_l = \frac{1}{2} \Sigma s \in 1 D_s(y_l) \). Then the probability of a recombination pattern over the \( L - 1 \) intervals between \( L \) consecutive loci on a chromosome is given by \((1/p)1M_1M_2 \ldots M_{L-1}1\), where \( M_l = N_l \) whenever there is no recombination in the \( l \)th interval and \( M_l = R_l \) when there is recombination. See Zhao et al. (1995b) for the proof of this result.

Sampling inheritance patterns conditional on the genotype data under the independence model: Thompson (2000b, Sect. 8.4) describes how to sample the inheritance pattern \( X \) from \( P_i(X|Y) \), the probability distribution of the inheritance patterns given the genotype data \( Y \) under the independence model, and we give the details here. Let \( X_l = (X_{l1}, X_{l2}, \ldots, X_{lq}) \) be the inheritance vector at locus \( l \), \( Y_l \) the genotype at locus \( l \), and

\[
\alpha_l(s) = P(X_l = s, Y_l, Y_{<l} \ldots, Y_l).
\]

The terms \( \alpha \) can be calculated recursively as

\[
\alpha_{t+1}(s) = \sum_i P(Y_{t+1}|X_{t+1} = s)P(X_{t+1} = s|X_t = t)\alpha_t(t)
\]

and

\[
\alpha_l(s) = P(Y_l|X_l = s)P(X_l = s).
\]

This is an application of the Baum (1972) forward algorithm and gives \( P_i(Y) = \Sigma \alpha_l(s) \). Then

\[
\alpha_l(s) = P(X_l = s, Y) \propto P(X_l = s|Y)
\]

can be used to sample \( X_l \) from \( P(X|Y) \). Recursively, after sampling \( X_l \), sample \( X_{l+1} \) from

\[
P(X_{l+1} = s|X_l, X_{l+1}, \ldots, X_{l+1}, Y) \propto P(X|X_{l+1})\alpha_{l+1}(s).
\]