*Biostatistics* (2006), **7**, 3, *pp.* 374–386
doi:10.1093/biostatistics/kxj012
Advance Access publication on December 14, 2005

# Incorporating interference into linkage analysis for experimental crosses

NICOLA J. ARMSTRONG*

*Division of Radiotherapy, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands*
n.armstrong@nki.nl

MARY SARA MCPEEK

*Department of Statistics, University of Chicago, Chicago, IL 60637, USA*

TERENCE P. SPEED

*Department of Statistics, University of California at Berkeley, Berkeley, CA 94720, USA and Division of Genetics and Bioinformatics, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia*

SUMMARY

The phenomenon of interference in genetic recombination is well-known and studied in a wide variety of organisms. Multilocus linkage analysis, which makes use of recombination patterns among all genetic markers simultaneously, is routinely used with data on humans and experimental organisms to build genetic maps. It is also used to try to determine the genes involved in traits of interest, such as common diseases. Most linkage analyses performed today ignore the occurrence of genetical interference. We present an extension to the Lander–Green algorithm for experimental crosses (backcross and intercross) to incorporate crossover interference according to the $\chi^2$ model. Simulation results show the impact of using this model on the accuracy of estimated genetic maps.

*Keywords*: Crossover interference; Hidden Markov model; Linkage analysis.

## 1. INTRODUCTION

Genetical interference was first observed in Thomas Hunt Morgan's *Drosophila* laboratory early last century (Muller, 1916). During meiosis, chromosomes replicate, producing sister chromatids. The homologous chromosomes then pair and synapse, forming a four-strand bundle. Once pairing is complete, crossing over, the reciprocal exchange of chromosomal segments among nonsister chromatids, begins. As the chromosomes separate, the crossover positions become visible as chiasmata. Two types of interference are distinguished in this crossover process: chromatid interference, the nonrandom choice of chromatids involved in adjacent chiasma, and crossover interference, the nonrandom placement of chiasma along a chromosome. Experimental evidence in various organisms has shown that crossovers do not occur at

*To whom correspondence should be addressed.

random, rather they are more evenly spaced along a chromosome (e.g. Hultén, 1974; Blank *et al.*, 1988). Chromatid interference can only be detected if all four products of meiosis can be recovered. However, to date there has been little evidence of this type of interference in experimental organisms.

For experimental crosses, the most common method of linkage analysis in use is the Lander–Green (LG) algorithm (Lander and Green, 1987). This algorithm makes use of a hidden Markov model (HMM) to construct a genetic map using multilocus linkage analysis. The LG approach is based on the assumption that there is no interference, either crossover or chromatid (NI model). The crossover positions are assumed to follow a Poisson distribution and the nonsister chromatids are chosen at random for each crossover.

Various models for the crossover process have been introduced over the years (e.g. Karlin and Liberman, 1979; Risch and Lange, 1979; King and Mortimer, 1990; McPeek and Speed, 1995). The $\chi^2$ distribution was first used in a model for crossing over in Fisher *et al.* (1947). We consider here a four-strand version of this model. Foss *et al.* (1993) gave a simple biological motivation and interpretation for the model based on gene conversions. They proposed that gene conversions are intermediate events that can resolve in reciprocal crossing over but do not need to do so. The $\chi^2$ model assumes that $m$ noncrossover resolutions of gene conversion occur between each crossover. The gene conversion events follow a Poisson distribution, and the distance between actual crossovers follows a $\chi^2$ distribution with $2(m+1)$ degrees of freedom. This model has been found to fit observed data from a variety of organisms as well (Foss *et al.*, 1993; Zhao *et al.*, 1995; Broman *et al.*, 2002), indeed better than other models (McPeek and Speed, 1995; Broman and Weber, 2000). Yet, despite the literature on alternative crossover models, in practice the phenomenon of interference is routinely ignored in genetic-mapping exercises. At most, the Kosambi map function is used after recombination fractions have been estimated under the NI model.

Geneticists have also known for many years that the rate of recombination differs, often substantially, between males and females for organisms such as humans and the mouse. In other species, such as the Lepidoptera *Bombyx mori* and *Drosophila melanogaster*, recombination does not occur at all in one of the sexes. This is equivalent to the genetic distances between markers being zero since no recombination is observed between them. Unfortunately, due to lack of parameter identifiability it is not possible to estimate sex-specific distances from an $F_2$ cross. The two heterozygous genotypes are indistinguishable, so that if both recombination fractions are greater than zero, the chromosome on which the crossover occurred cannot be determined. In the case of no recombination in one sex, genetic distances can be estimated from a group of $F_2$ progeny. However, if a model of equal recombination between the sexes is used, the genetic distance estimates will be too small as the model produces, in effect, sex-averaged genetic distance estimates. When recombination is absent in one sex, these distances will be much smaller than the true distances in the recombinant sex.

In this paper, we present an extension of the LG algorithm to incorporate crossover interference according to the $\chi^2$ model. The following section presents the algorithm for both the backcross (BC) and $F_2$ cases. The special case of no recombination in one sex is also considered for the $F_2$ progeny. In the results section, the accuracy of this extended algorithm in estimating genetic maps is evaluated through simulation studies and compared to the original LG algorithm and the Kosambi transformation. The extra time required to obtain genetic maps when incorporating crossover interference is also considered.

## 2. Model

Let $\mathcal{M}_1, \ldots, \mathcal{M}_T$ denote T genetic loci listed in order along a chromosome. The observed genotypic information can be denoted $O = (O_1, \ldots, O_T)$, where $O_t$ contains the genotypes of all individuals at $\mathcal{M}_t$. Define $\lambda_t = 2(m+1)d_t$, where $d_t$ is the genetic distance, in Morgans, between $\mathcal{M}_t$ and $\mathcal{M}_{t+1}$ and $m$ is a defined constant. An important difference between the LG and $\chi^2$ algorithms is that LG estimates the recombination fraction while the $\chi^2$ algorithm estimates $\lambda_t$. If $m = 0$, this gives a direct estimate of

the genetic distance and no transformation is necessary. For $m > 0$, it estimates a fixed multiple of the genetic distance.

Under the $\chi^2$ model, there are assumed to be $m$ unobserved crossover intermediates between each crossover. The hidden chain $y_t$ for the $\chi^2$ model keeps track of both the number of crossover intermediates that have occurred since the last crossover and the origin of the DNA, grandmaternal (gm) or grandpaternal (gp). Specifically, the states $i = 0, \ldots, m$ represent the event of $i$ crossover intermediates and gm origin of the DNA, while states $i = m+1, \ldots, 2m+1$ represent the event of $i-(m+1)$ crossover intermediates and gp origin of the DNA. The process along a chromosome can be described by the embedded chain in Figure 1. Every $(m+1)$st event results in a crossover that involves the strand of interest with probability $\frac{1}{2}$. For example, starting in state 0 we must progress through $m$ steps to state $m$, the next intermediate which will result in a crossover. When in state $m$, there is either a crossover and switch to gp DNA (state $m+1$) or there is no crossover and we stay on gm DNA, returning to state 0. Note that if $m = 0$, the chain is that of the NI model with the hidden states simply representing regions of gm and gp DNA on the chromosome.

Conditional on parental genotypes, the genotypes of each offspring are independent. Hence, the likelihood may be calculated for each individual separately and then combined. In the discussion that follows, the HMM is defined for one individual. The superscript $k$ is used when necessary to denote the $k$th offspring and omitted otherwise. It is also assumed, without loss of generality, that the $F_1$ parent is crossed with individuals from the grandmaternal strain to produce BC progeny.

### 2.1 *Backcross*

Given the inbred strains AA and aa, at any locus the BC offspring will have genotype Aa or the genotype of the homozygous parent (either aa or AA depending on which line the homozygous parent came from). We define the observed genotypes at a given locus of a BC animal to be A if homozygous and H if

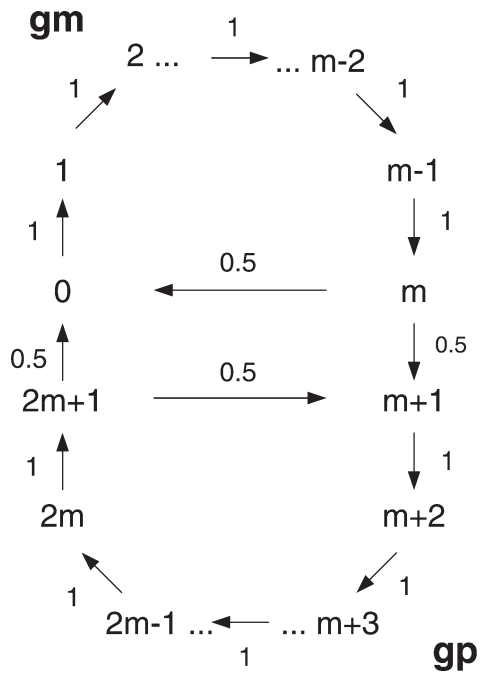


Fig. 1. Embedded chain along a chromosome for the $\chi^2$ model with parameter $m$.

heterozygous. Hence, for a BC experiment, the elements of $O$ are A, H and $-$, the three possible observed genotypes, where '$-$' denotes missing. In this case, we need consider only the chromosome inherited from the $F_1$ parent.

The initial-state distribution is uniform over all states. The transition probability matrices, $A(\lambda_t)$, at each marker locus have $2(m+1) \times 2(m+1)$ entries of which only $3m+2$ are distinct. The entries $a_{ij}(t)$ can be related to the distinct elements $a_l(t)$ by grouping the pairs $(i, j)$ into sets $I_l$:

1. If $0 \leqslant i, j \leqslant m$ or $i, j > m$, $(i, j) \in I_l$, where $l = j - i$.
2. If $0 \leqslant i \leqslant m$, $j > m$, then for $j > i + m$, $(i, j) \in I_l$, where $l = j - i$, else $(i, j) \in I_l$, where $l = j - i - (m+1)$.
3. If $i > m, 0 \leqslant j \leqslant m$, then for $j + (m+1) > i - 1$, $(i, j) \in I_l$, where $l = j - i + 2m + 2$, else $(i, j) \in I_l$, where $l = j - i + (m+1)$.

For example, consider $a_{12}(t)$ when $m = 2$. The chain progresses from state 1 to state 2 if there has been a single intermediate event, if there have been four intermediate events but no crossover has occurred ($1 \to 2 \to 0 \to 1 \to 2$), or if any multiple of four intermediate events and an even number of crossovers have occurred. This is also the same set of circumstances under which the chain could move from state 4 to state 5. In this case, the transition matrix is given by:

$$A(\lambda_t) = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & a_4 & a_5 \\ a_{-1} & a_0 & a_1 & a_{-1} & a_3 & a_4 \\ a_{-2} & a_{-1} & a_0 & a_{-2} & a_{-1} & a_3 \\ a_3 & a_4 & a_5 & a_0 & a_1 & a_2 \\ a_{-1} & a_3 & a_4 & a_{-1} & a_0 & a_1 \\ a_{-2} & a_{-1} & a_3 & a_{-2} & a_{-1} & a_0 \end{pmatrix}.$$

Defining

$$f_{m,s} = \frac{1}{2} \mathrm{e}^{-\lambda_t} \sum_{k=1}^{\infty} \frac{\lambda_t^{k(m+1)+s}}{(k(m+1)+s)!},$$

the distinct $a_l(t)$ are

$$a_l(t) = \begin{cases} f_{m,l}(\lambda_t), & -m \leqslant l \leqslant -1, \\ \frac{\mathrm{e}^{-\lambda_t} \lambda_t^l}{l!} + f_{m,l}(\lambda_t), & 0 \leqslant l \leqslant m, \\ f_{m,l-(m+1)}(\lambda_t), & m+1 \leqslant l \leqslant 2m+1. \end{cases}$$

The probability that a particular genotype is observed at a given marker is conditional on the hidden state at that marker, that is $b_i(O_t) = P(O_t | y_t = i)$. Random errors in the genotyping process of rate $\epsilon$ are allowed. If there is no genotyping error, $0 \leqslant y_t \leqslant m$ implies $O_t$ = A and $m+1 \leqslant y_t \leqslant 2m+1$ implies $O_t$ = H.

Likelihood calculations are done by way of the 'forward variables':

$$\alpha_t(i) = P(O_1, \ldots, O_t, y_t = i | \lambda).$$

For $n$ offspring, the likelihood is given by

$$P(O|\lambda) = \prod_{k=1}^{n} \sum_{i=0}^{2m+1} \alpha_T^k(i), \tag{2.1}$$

where $\alpha_T^k(i)$ denotes the value of $\alpha_T(i)$ for the $k$th offspring. The forward variables give the probability of observing the partial genotype sequence up to $\mathcal{M}_t$ and being in hidden state $i$ at $\mathcal{M}_t$. Likewise, there are corresponding 'backward variables', $\beta_t(i)$, giving the probability of observing the partial genotype sequence from $\mathcal{M}_{t+1}$ to $\mathcal{M}_T$ given the hidden state $i$ at $\mathcal{M}_t$. Both the forward and backward variables are computed recursively.

The expectation maximization (EM) algorithm is employed to determine the maximum likelihood estimates (MLEs) of the intermarker genetic distances, $d_t, t = 1, \ldots, T-1$. In particular, Baum's lemma is used, so that $Q(\lambda, \lambda')$ is explicitly maximized, where

$$Q(\lambda, \lambda') = \sum_{y} P(y, O|\lambda) \log P(y, O|\lambda')$$

and $y = (y_1, \ldots, y_T)$. In terms of $\lambda_t'$ this is equivalent to maximizing, for $n$ BC offspring,

$$\sum_{k=1}^{n} \sum_{y_t^k, y_{t+1}^k} P(y_t^k, y_{t+1}^k|O^k, \lambda) \log P(y_{t+1}^k|y_t^k, \lambda_t') = \sum_{k=1}^{n} \sum_{l=-m}^{2m+1} E(n_l^k(t)|O, \lambda) \log a_l(\lambda_t'),$$

where $n_l^k(t) \in \{0, 1\}$ is defined to be the number of transitions from state $i$ to state $j$ for a pair $(i, j) \in I_l$ for the $k$th offspring. The E step then involves calculating

$$E(n_l^k(t)|O, \lambda) = \sum_{i,j \in I_l} \xi_t^k(i, j),$$

where $\xi_t(i, j) = \frac{\alpha_t(i) a_{ij}(t) b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i,j} \alpha_t(i) a_{ij}(t) b_j(O_{t+1}) \beta_{t+1}(j)}$.

For general $m$ there is no closed-form solution to the maximization step. Instead, we must search for the maximum,

$$\underset{\lambda_t'}{\operatorname{argmax}} \sum_{k=1}^{n} \sum_{l=-m}^{2m+1} \log a_l(\lambda_t') E(n_l^k(t)|O, \lambda).$$

Brent's (1973) method is used to find a solution numerically, iterating the E and M steps until the scaled genetic distance estimates, $\lambda_t$, change by less than a prespecified tolerance level.

### 2.2 $F_2$ *intercross*

Given the two inbred strains AA and aa, at a given locus each $F_2$ offspring will be homozygous AA, homozygous aa, or heterozygous Aa. The observed genotypes are defined as A or B (homozygous AA or aa) and H if heterozygous. We can further define two incomplete genotypes C and D to be not B (i.e. AA or Aa) and not A (i.e. aa or Aa), respectively. For an $F_2$ intercross, both chromosomes must be considered because it is impossible to determine, at each heterozygous locus, which allele came from which parent. There are $4(m+1)^2$ hidden states, since on each chromosome the chain could be in any one of the states $0, \ldots, 2m+1$. For example, if $m = 1$, there are 16 hidden states, shown here with their corresponding

states on each individual chromosome:

| | | | |
|---|---|---|---|
| $0 \leftrightarrow 0, 0$ | $4 \leftrightarrow 1, 0$ | $8 \leftrightarrow 2, 0$ | $12 \leftrightarrow 3, 0$ |
| $1 \leftrightarrow 0, 1$ | $5 \leftrightarrow 1, 1$ | $9 \leftrightarrow 2, 1$ | $13 \leftrightarrow 3, 1$ |
| $2 \leftrightarrow 0, 2$ | $6 \leftrightarrow 1, 2$ | $10 \leftrightarrow 2, 2$ | $14 \leftrightarrow 3, 2$ |
| $3 \leftrightarrow 0, 3$ | $7 \leftrightarrow 1, 3$ | $11 \leftrightarrow 2, 3$ | $15 \leftrightarrow 3, 3.$ |

Again, the convention that states $0, \ldots, m$ represent gm DNA and $m+1, \ldots, 2m+1$ gp DNA on a single chromosome is used. The initial-state distribution is assumed uniform. The conditional distribution of the observed genotypes given the hidden state $i$ at $\mathcal{M}_t$ allows for the incomplete marker genotypes C and D, and incorporates a random error rate of $\epsilon$.

The transitions on the maternal and paternal chromosomes are independent; hence, the joint probability is simply the product of the marginal probabilities of the transitions on each chromosome. Formally, the transition probability matrix for the states is given by

$$C(\lambda_t^f, \lambda_t^p) = A(\lambda_t^f) \otimes A(\lambda_t^p),$$

where $\mathrm{A}(\lambda_t^f)$ and $\mathrm{A}(\lambda_t^p)$ are the transition matrices on the maternal and paternal chromosomes, respectively, and $\otimes$ the Kronecker product. It is generally assumed that the recombination rates for the two sexes are equal, i.e. $\lambda_t^f = \lambda_t^p = \lambda_t$ and $C(\lambda_t) = A(\lambda_t) \otimes A(\lambda_t)$. In general, the (nondistinct) $c_{ij}(t)$ are obtained by translating the indices $i$ and $j$ into the corresponding indices for the transition matrix $A(\lambda_t)$ for each individual chromosome. The algorithm detailed in the previous section is used to determine the distinct quantities in both cases ($a_{s_1}(t)$ and $a_{s_2}(t)$), giving

$$c_{ij}(t) = a_{s_1}(t) a_{s_2}(t), \quad -m \leqslant s_1 \leqslant s_2 \leqslant 2m+1.$$

For example, when $m = 1$ the states are given above and the probability of moving from state 1 to state 4 involves a transition from states 0 to 1 on the first chromosome and 1 to 0 on the second chromosome. Using the notation from Section 2.1, the marginal probabilities for these transitions are $a_1(t)$ and $a_{-1}(t)$, respectively. Therefore, in this case, $c_{14}(t) = a_1(t)a_{-1}(t)$.

The special case of no recombination in one sex only affects the transition matrix in the model. Without loss of generality, if $\lambda_t^f = 0$ for $t = 1, \ldots, T-1$ (i.e. no female recombination), then:

$$A(\lambda_t^f) = \mathcal{I}_{2m+1},$$

where $\mathcal{I}_{2m+1}$ is the $(2m+1) \times (2m+1)$ identity matrix. The joint transition matrix then simplifies to

$$C(\lambda_t^f, \lambda_t^p) = \begin{pmatrix} A(\lambda_t^p) & 0 & \cdots & 0 \\ 0 & A(\lambda_t^p) & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & A(\lambda_t^p) \end{pmatrix}.$$

In other words, only the block diagonal elements of $C(\lambda_t^f, \lambda_t^p)$ are nonzero.

Having defined the elements of the HMM, calculation of the forward, backward, and $\xi$ variables can then be carried out as in Section 2.1, substituting $c_{ij}(t)$ in place of $a_{ij}(t)$. In order to find the MLEs of genetic distance, we invoke the EM algorithm once more. For $n$ $\mathrm{F}_2$ offspring and equal recombination

rates for the sexes, i.e. $\lambda^f = \lambda^p$, $Q(\lambda, \lambda')$ in terms of $\lambda'_t$ becomes, in the general case,

$$\sum_{k=1}^{n} \sum_{y_t^k, y_{t+1}^k} P(y_t^k, y_{t+1}^k | O^k, \lambda) \log P(y_{t+1}^k | y_t^k, \lambda'_t)$$

$$= \sum_{k=1}^{n} \sum_{l_1, l_2 = -m}^{2m+1} \log[a_{l_1}(\lambda'_t) a_{l_2}(\lambda'_t)] E(n_{l_1, l_2}^k(t) | O, \lambda).$$

$n_{l_1, l_2}^k(t)$ is the number of transitions from $i$ to $j$ for a pair $(i, j)$ related to $I_{l_1}$ and $I_{l_2}$ on the individual chromosomes. The E step consists then of calculating $E(n_{l_1, l_2}^k(t) | O, \lambda)$ and the M step is carried out numerically using Brent's (1973) method. To avoid underflow problems, the forward and backward variables are scaled by way of logarithms for all values of $m$ and for all progeny numbers, when there are more than 100 markers in the data set for the BC HMM and, similarly, when there are more than 70 markers in the $F_2$ case.

## 3. RESULTS

To test the performance of the $\chi^2$ HMM, genotype data were simulated for both $F_2$ and BC breeding schemes. The data were simulated under the assumption of crossover interference. To mimic biological reality, the $\chi^2$ model with $m = 6$ was used. This model was previously estimated to provide the best fit for mouse data (Lin and Speed, 1996).

The programs were first verified by comparing all results (estimated genetic distances and likelihood values) obtained when $m = 0$ against both MapMaker and R/qtl (Broman *et al.*, 2003) which assume no crossover interference. Furthermore, for the BC case, the theoretical and empirical map functions were compared for a variety of $m$ values. For fixed $m$, 1000 BC mice were simulated and genotyped every centimorgan along a single chromosome. Using this genotype data, the empirical recombination fractions could then be computed. Plotting the empirical recombination fractions against the true genetic distance and the theoretical map function, given by

$$\theta = \frac{1}{2}\left[1 - e^{-2d} \sum_{i=0}^{m} \frac{(2d)^i}{i!}\left(1 - \frac{i}{m+1}\right)\right],$$

on the same set of axes found close agreement. Figure 2 depicts the results for $m = 0, 1, 2$, and 3.

### 3.1 *Bombyx mori*

A set of $F_2$ genotype data for *B. mori*, the domesticated silkworm, was obtained from D. G. Heckel (Shi *et al.*, 1995). In Lepidoptera, females are the heterogametic sex and have achiasmatic meiosis (no crossing over). The purpose of the original study was to create a genetic linkage map for *B. mori*. The data consist of the genotypes of 52 $F_2$ progeny at a total of 58 markers. The markers are both dominant and codominant. Fifty of the autosomal loci were grouped into a total of 15 primary linkage groups (PLGs), ranging from 2 to 8 loci each, by Shi *et al.* (1995). The PLGs were taken as originally assigned and the genetic maps were re-estimated using the HMM with $m = 0$ in order to verify the published map. In addition to estimating the intermarker genetic distances, the log-likelihood values for all possible map orders were calculated conditional on their PLG assignment to verify the published orders for all PLGs. The results, both estimated intermarker genetic distances and map orders, matched the published maps for all PLGs except PLG 9 and 15 (Figure 3).

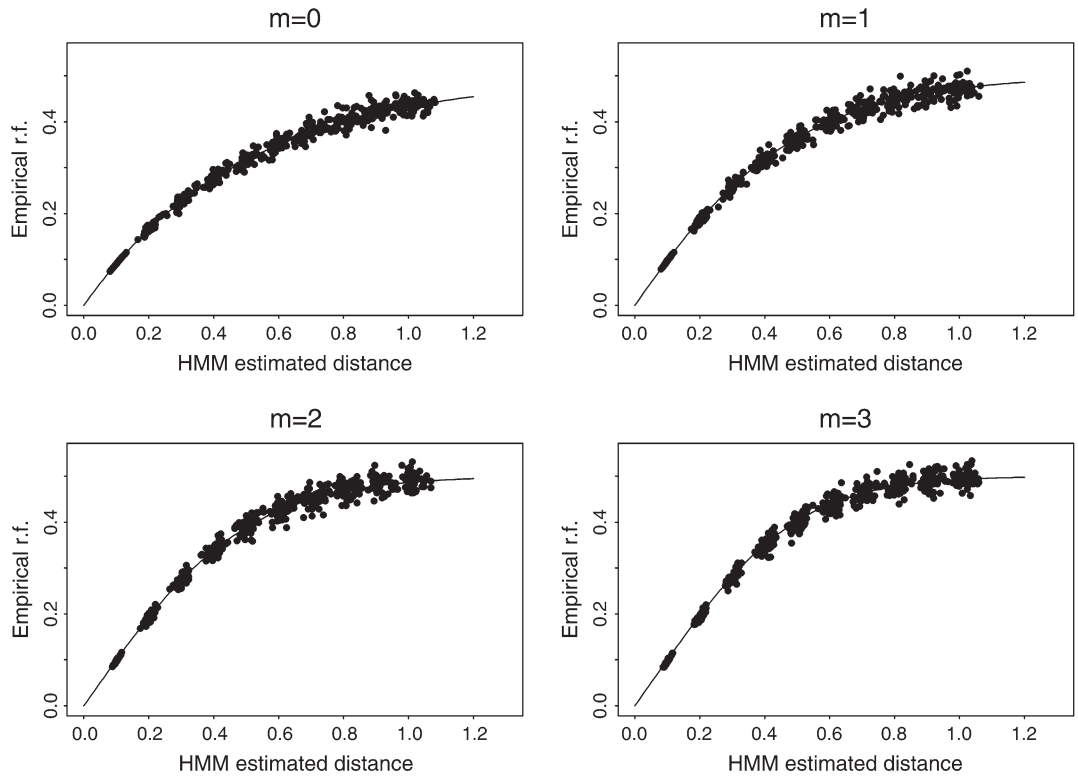


Fig. 2. Empirical and theoretical map functions for various values of $m$.

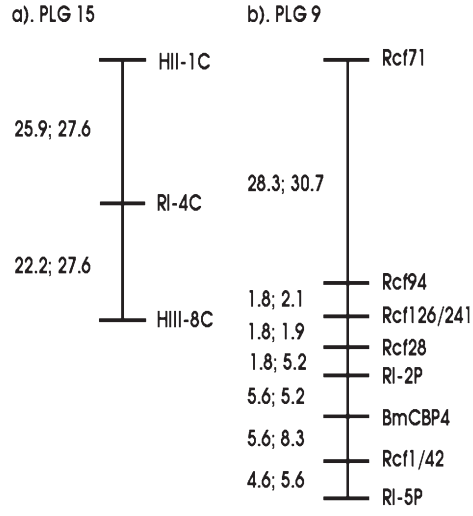


Fig. 3. Genetic maps for PLG 9 and 15, as estimated by Shi *et al.* (1995) (listed first) and by using the HMM (second).

PLG 15 consists of three dominant markers. The slight difference in estimated distances can be explained by the different estimation methods used. Shi *et al.* (1995) analyzed the linkage data under the assumptions of no crossing over in the female and no crossover interference, using maximum likelihood estimation and the EM algorithm. For some triple genotype combinations of dominant markers the Markov property of the observed genotype information $O$ is lost, e.g. $P(O_3 = C|O_1 = C, O_2 = C) \neq P(O_3 = C|O_1 = A, O_2 = C)$, making it incorrect to calculate the likelihood using a Markov chain, as was done by Shi *et al.* Hence, multilocus probabilities must be calculated some other way, such as by using an HMM. The inclusion of the hidden chain allows for the correct calculation of multilocus probabilities when there is dominance or missing data at some loci. Our method therefore makes more efficient use of the data for this PLG.

The largest PLG was PLG 9, consisting of eight loci. For this PLG, Shi *et al.* (1995) used MapMaker to estimate an approximation to the maximum likelihood map, as their program was too slow to calculate the maps for all possible orders in a reasonable time. Furthermore, the recombination fractions between loci were estimated from the MapMaker sex-averaged estimates using a linear regression (for details see Shi *et al.*, 1995). In contrast, using the HMM approach, the likelihood values for all possible orders were calculated and the genetic distances estimated under the assumption of no crossing over in the female. We concluded that the order of loci was the same as the published order but that the genetic distance estimates were different. Given their ad hoc method of estimating the distances for this PLG, it is not surprising that the HMM distance estimates differ from their published ones (Figure 3).

### 3.2 *Timing*

As $m$ increases, the number of states in the HMM also increases. Recall there are $2(m + 1)$ states for BC progeny and $4(m + 1)^2$ states for $F_2$ progeny. The increase in the number of states leads, necessarily, to an increase in the time taken to converge to the maximum likelihood estimates due to more computations being required to calculate the forward, backward, and $\xi$ variables. The time required to perform the E step also increases with $m$. However, finding the values of $\lambda_t$ which maximize $Q(\lambda, \lambda')$ at each iteration (the M step) requires only a slight increase in time. An important issue with regard to the time taken to find the estimated genetic distances is that $\lambda_t$, not $d_t$, is estimated. The shape of $Q(\lambda_t, \lambda_t')$ rapidly becomes very flat as the value of $\lambda_t$ approaches the maximum likelihood estimate, so that in order to get an accurate estimate of $d_t$ many iterations are often required. It is the increased number of iterations of the EM algorithm coupled with the increase in E step time which gives rise to the large overall increase in time.

The time taken to find the maximum likelihood estimates for varying $m$ values was investigated for both BC and $F_2$ cases. The impact of increasing the number of mice in the data set and, separately, the impact of increasing the number of genetic markers were investigated. The value of $m$ used for simulating the genotype data was equal to the value of $m$ used in the HMM. Initial values for the distances were taken to be the NI ($m = 0$) estimates.

All programs were run on 400-MHz UltraSparc machines with 256 MB RAM. The HMM does not require huge amounts of memory, but the time taken depends quite heavily on the speed of the machine. Under each set of conditions investigated, the differences in the time taken for each data set were directly attributable to the number of iterations of the EM algorithm required before the convergence tolerance was satisfied. The time taken rises linearly as both the number of markers and the number of mice rise, with times also increasing with $m$. Kinks in Figure 4 can be attributed to the fact that the machines used were not available solely for our purposes. However, the graphs give a rough indication of the increase in time required.

The relationship between number of mice, time taken, and $m$ is very similar for BC (Figure 4) and $F_2$ mice (not shown), except that the times involved are much larger in the $F_2$ case. For 1600 mice, it took

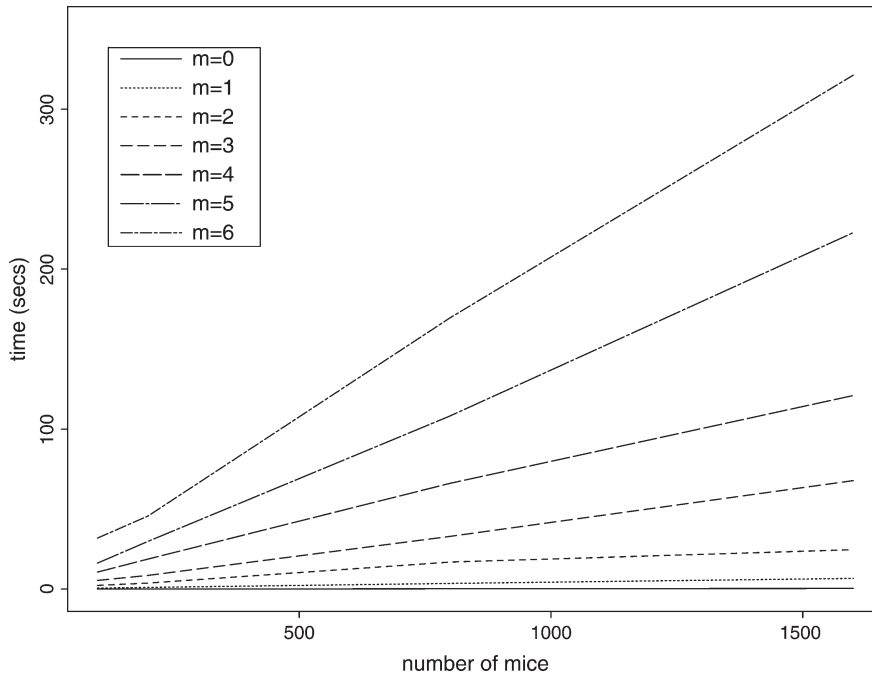


Fig. 4. Impact of increasing the number of mice on the time taken to find the distance MLEs for 10 intervals in the BC case.

the BC HMM 0.4 s, on average, to compute distance estimates for 10 intervals when $m = 0$, whereas when $m = 6$ the time taken rose to 321 s (5.5 min), an 800-fold increase. In the $F_2$ case, it took 12.2 s for $m = 0$, and 235 880 s (65.5 h) for $m = 6$, an almost 20 000-fold increase. When investigating the impact of the number of markers in a data set on time taken to find the distance MLEs, the number of mice was fixed at 400 in the BC case but reduced to 100 for $F_2$ data. Again, the relationship between the number of markers, $m$, and the amount of time taken to calculate the MLEs was very similar for $F_2$ and BC mice (data not shown). The time taken to calculate distance estimates rises linearly with an exponential increase in markers, similar to that shown in Figure 4 for the number of mice. For example, for 32 markers, the time taken to produce distance estimates when $m = 0$ was, on average, 4.4 s and increased to approximately 13 h if $m = 6$, for $F_2$ data, a 10 000-fold increase. In the BC case, under the no-interference model, estimates were obtained in 0.18 s compared to 163 s (2.7 min) if $m = 6$.

### 3.3 *Accuracy*

The most important gain that we hope to make by extending the LG algorithm to incorporate crossover interference is in the accuracy of the estimates. In order to investigate this, both BC and $F_2$ mice were simulated under the $\chi^2$ model with $m = 6$. In each instance, 1000 sets of 300 BC mice and 100 sets of 300 $F_2$ mice were simulated and the distance estimates obtained using both the NI model and the $\chi^2$ model recorded. The accuracies of distance estimates obtained using the Kosambi map function to transform the recombination fractions obtained under the NI model were also compared.

The entries in Table 1 demonstrate that as the distance between markers increases, the gains in precision from using the $\chi^2$ model also increase when the data are fully informative, even if the wrong level

Table 1. *Root mean squared error in centimorgans of the distance estimates obtained under various models for fully informative BC and $F_2$ data simulated under the $\chi^2$ model with $m = 6$*

| Model used | True distance (cM) | | | | | |
|---|---|---|---|---|---|---|
| | BC data | | | $F_2$ data | | |
| | 5 | 10 | 20 | 5 | 10 | 20 |
| NI | 1.38 | 2.28 | 3.90 | 0.87 | 1.58 | 3.04 |
| $\chi^2_{10}(m = 4)$ | 1.24 | 1.82 | 2.46 | 0.78 | 1.27 | 1.93 |
| $\chi^2_{14}(m = 6)$ | 1.23 | 1.81 | 2.38 | 0.78 | 1.26 | 1.87 |
| $\chi^2_{18}(m = 8)$ | 1.23 | 1.81 | 2.34 | | | |
| NI + Kosambi | 1.25 | 1.89 | 2.78 | 0.79 | 1.32 | 2.17 |

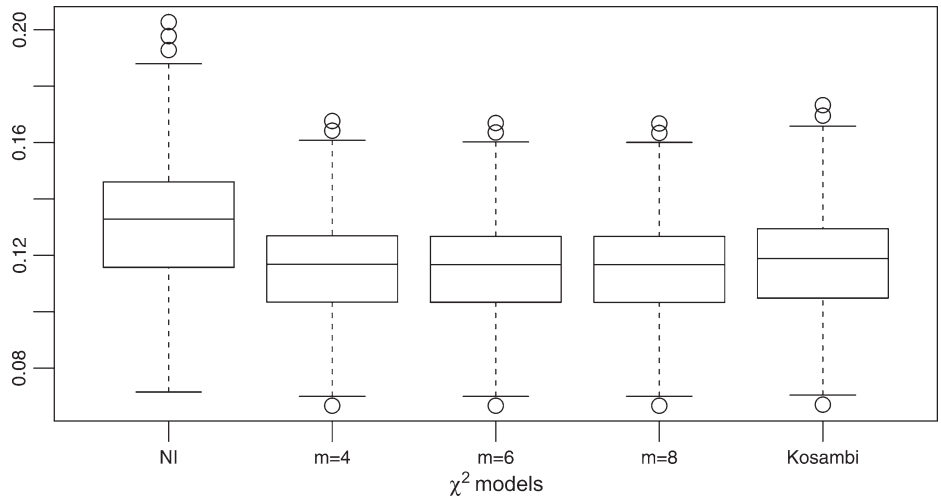


Fig. 5. Box plot of distance estimates obtained when true distance is 10 cM. In this case 300 BC meioses were simulated with 1% random genotyping error.

of interference is used in the HMM. Introducing a random genotyping error into the genotype data when either no error or the wrong amount of error is allowed for in the HMM produces similar results. For all distances the $\chi^2$ model is more accurate, but the gains in accuracy are sometimes slight, especially if the genetic distances involved are small. Figure 5 shows the results found under one scenario, the results being typical of those found under other scenarios also. Simulations with 5% and 10% missing genotype data, and both 2% and 4% incomplete genotypes for the $F_2$ case were also performed. We found that this conclusion holds true in those cases also (data not shown). The results presented in Figure 6 are representative of those found under the different conditions investigated, for both $F_2$ and BC data.

## 4. CONCLUSIONS

We have developed and tested an HMM for estimating genetic distances that incorporates crossover interference, specifically under the $\chi^2$ model. The HMM for $F_2$ data was extended to incorporate the situation where there is no crossing over in one of the sexes. The $\chi^2$ HMM produces more accurate estimates of

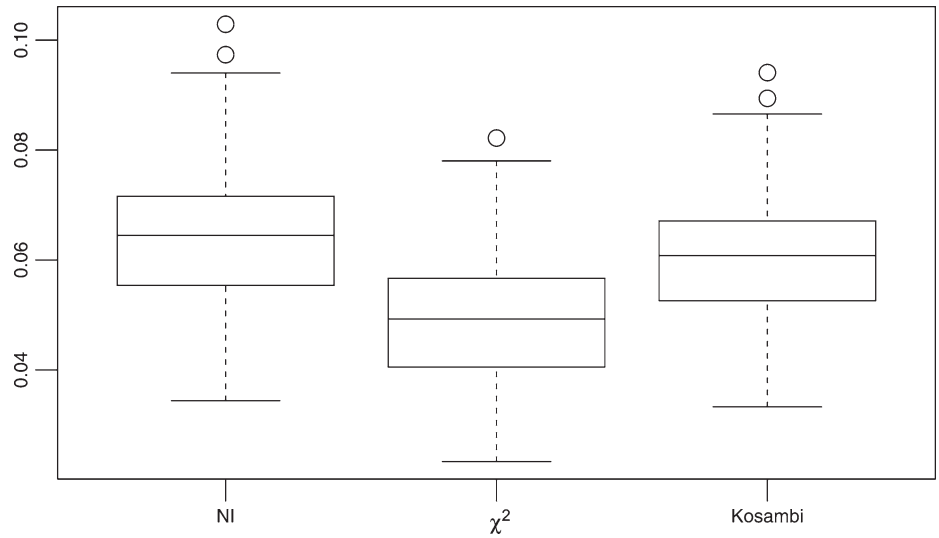


Fig. 6. Box plot of distance estimates obtained when true distance is 5 cM. In this case, 300 $F_2$ meioses were simulated with 5% missing and 4% incomplete genotyping data.

genetic distance, in the presence of crossover interference, than the NI model which is commonly implemented.

The gains in accuracy achieved by using the $\chi^2$ model, with $m = 6$, instead of the Kosambi map function, are small for distances under 10 cM. With the ever-increasing number of genetic markers available and the falling costs of genotyping, in future we can expect many markers to be typed for a particular experimental cross, leading to a reduction in the typical intermarker genetic distance. The use of single-nucleotide polymorphisms and microarray technology for genotyping purposes will also mean smaller distances between markers. Using a lower level of interference than true ($m = 4$ instead of $m = 6$) resulted in very little difference in the level of accuracy obtained, likewise if a slightly higher $m$ is used in the HMM. The problem of estimation of $m$ has been treated in detail by Zhao *et al.* (1995) and Goldstein *et al.* (1997). The method they use is to compute the maximum likelihood for different values of $m$ and compare them. The HMM presented in this paper may also be used to estimate the level of interference using the same approach of comparing likelihood values for different $m$ values (NJ Armstrong and TP Speed, in preparation). The computational cost is equal to the sum of the computational costs of maximizing the likelihood for the individual values of $m$. The accuracy of the estimate obtained is in large part dependent on the number of individuals in the cohort and the number of markers genotyped. A more realistic model for the distribution of crossovers along a chromosome is a mixture of the $\chi^2$ and NI models (Copenhaver *et al.*, 2002). For small-mixture probabilities, the results on accuracy of distance estimates are likely to be similar to those presented here.

The time taken to find estimates of genetic distance for $F_2$ data with either a large number of markers or progeny, for large levels of interference ($m = 6$), is at least 1000 times more than if no interference is assumed, and can take as much as 20 000 times longer. The increase in time required in conjunction with the small increases in accuracy suggests that this model will not become widely implemented in practice, even with the rapid increase in computing power. Development of the HMM allowing for crossing over in only one sex, for both the $\chi^2$ model and the NI model, will hopefully be of use for researchers involved with organisms such as those in the order Lepidoptera.

The software used in this paper is available on request from the first author.

## ACKNOWLEDGMENTS

We thank David Heckel for kindly allowing us access to the *Bombyx mori* data. This work was supported by National Institutes of Health grant GM59506-01 (to Speed).

## REFERENCES

BLANK, R., CAMPBELL, G., CALABRO, A. AND D'EUSTACHIO, P. (1988). A linkage map of mouse chromosome 12: localization of Igh and effects of sex and interference on recombination. *Genetics* **120**, 1073–1083.

BRENT, R. (1973). *Algorithms for Minimization Without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall.

BROMAN, K., ROWE, L., CHURCHILL, G. AND PAIGEN, K. (2002). Crossover interference in the mouse. *Genetics* **160**, 1123–1131.

BROMAN, K. AND WEBER, J. (2000). Characterization of human crossover interference. *American Journal of Human Genetics* **66**, 1911–1926.

BROMAN, K., WU, H., SEN, S. AND CHURCHILL, G. (2003). R/qtl: Qtl mapping in experimental crosses. *Bioinformatics* **19**, 889–890.

COPENHAVER, G., HOUSWORTH, E. AND STAHL, F. (2002). Crossover interference in Arabidopsis. *Genetics* **160**, 1631–1639.

FISHER, R., LYON, M. AND OWEN, A. (1947). The sex chromosome in the house mouse. *Heredity* **1**, 335–365.

FOSS, E., LANDE, R., STAHL, F. AND STEINBERG, C. (1993). Chiasma interference as a function of genetic distance. *Genetics* **133**, 681–691.

GOLDSTEIN, D., ZHAO, H. AND SPEED, T. (1997). The effects of genotyping errors and interference on estimation of genetic distance. *Human Heredity* **47**, 86–100.

HULTÉN, M. (1974). Chiasma distribution at diakinesis in the normal human male. *Hereditas* **76**, 55–78.

KARLIN, S. AND LIBERMAN, U. (1979). A natural class of multilocus recombination processes and related measures of crossover interference. *Advances in Applied Probability* **11**, 479–501.

KING, J. AND MORTIMER, R. (1990). A polymerization model of chiasma interference and corresponding computer simulation. *Genetics* **126**, 1127–1138.

LANDER, E. AND GREEN, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Science of the United States of America* **84**, 2363–2367.

LIN, S. AND SPEED, T. (1996). Incorporating crossover interference into pedigree analysis using the $\chi^2$ model. *Human Heredity* **46**, 315–322.

MCPEEK, M. AND SPEED, T. (1995). Modeling interference in genetic recombination. *Genetics* **139**, 1031–1044.

MULLER, H. (1916). The mechanism of crossing-over. *The American Naturalist* **50**, 193–221, 284–305, 350–366, 421–434.

RISCH, N. AND LANGE, K. (1979). An alternative model of recombination and interference. *Annals of Human Genetics* **43**, 61–70.

SHI, J., HECKEL, D. AND GOLDSMITH, M. (1995). A genetic linkage map for the domesticated silkworm, Bombyx mori, based on restriction fragment length polymorphisms. *Genetic Research (Cambridge)* **66**, 109–126.

ZHAO, H., SPEED, T. AND MCPEEK, M. (1995). Statistical analysis of crossover interference using the chi-square model. *Genetics* **139**, 1045–1056.

[*Received April 19, 2005; first revision August 12, 2005; second revision October 7, 2005; third revision November 21, 2005; accepted for publication December 7, 2005*]