Tree Genetics & Genomes (2010) 6:651–662
DOI 10.1007/s11295-010-0281-2

ORIGINAL PAPER

A hidden Markov model approach to multilocus linkage analysis in a full-sib family

Chunfa Tong · Bo Zhang · Jisen Shi

Received: 14 June 2009 / Revised: 16 December 2009 / Accepted: 9 February 2010 / Published online: 16 March 2010

Abstract Statistical packages for constructing genetic linkage maps in inbred lines are well developed and applied extensively, while linkage analysis in outcrossing species faces some statistical challenges because of their complicated genetic structures. In this article, we present a multilocus linkage analysis via hidden Markov models for a linkage group of markers in a full-sib family. The advantage of this method is the simultaneous estimation of the recombination fractions between adjacent markers that possibly segregate in different ratios, and the calculation of likelihood for a certain order of the markers. When the number of markers decreases to two or three, the multilocus linkage analysis becomes traditional two-point or three-point linkage analysis, respectively. Monte Carlo simulations are performed to show that the recombination fraction estimates of multilocus linkage analysis are more accurate than those just using two-point linkage analysis and that the likelihood as an objective function for ordering maker loci is the most powerful method compared with other methods. By incorporating this multilocus linkage analysis, we have developed a Windows software, FsLinkageMap, for constructing genetic maps in a full-sib family. A real example is presented for illustrating linkage maps constructed by using mixed segregation markers. Our multilocus linkage analysis provides a powerful method for constructing high-density genetic linkage maps in some outcrossing plant species, especially in forest trees.

Communicated by J. Davis

C. Tong · B. Zhang · J. Shi (✉)
Key Laboratory of Forest Genetics and Biotechnology of Ministry of Education, Nanjing Forestry University,
Longpan Road No.159,
210037 Nanjing, Jiangsu Province, China
e-mail: jshi@njfu.edu.cn

Keywords Multilocus linkage analysis · Hidden Markov model · Full-sib family · EM algorithm

Introduction

With rapid development of molecular technology, it is now easy to obtain many kinds of DNA markers in short time, such as RAPD, RFLP, SSR, AFLP, ISSR, and EST, etc. Genetic linkage maps have been constructed by using these markers in many species. They are useful for locating quantitative trait loci (QTL), marker-assisted selection, and map-based gene cloning. Some computer packages for constructing linkage maps in inbred line crosses, such as MAPMAKER (Lander et al. 1987) and JOINMAP (Stam 1993), have been well developed and applied extensively. However, statistical methods and algorithms for linkage mapping in outcrossing species are difficult to implement because the genetic structure is more complicated in these species than in inbred lines (Lu et al. 2004).

Constructing genetic linkage maps based on molecular markers involves several statistical and mathematical procedures, which include two-point linkage analysis, multiple locus linkage analysis, linkage grouping, and multiple locus ordering. Two-point linkage analysis is the basis of constructing genetic linkage maps, in which the recombination fraction between any two markers and the logarithm of the odds (LOD) score for testing whether the pair of markers is linked are calculated. Linkage grouping is a clustering analytical procedure to categorize all available markers into several linkage groups based on the recombination fractions or the LOD scores. Generally, it is expected that the number of linkage groups will be in agreement with the number of chromosome pairs in the studied organism. Multiple locus ordering is the key procedure for accurately constructing

genetic linkage maps. For *m* molecular markers in a linkage group, the optimal order is chosen based on an objective function among the *m*!/2 potential orders. Objective functions for ordering marker loci include sum of squares of the difference between observed distances of pairs of markers and their expected values (Jensen and Helms Jørgensen 1975; Lalouel 1977; Weeks and Lange 1987), sum of adjacent recombination fractions (SARF; Falk 1989), product of adjacent recombination fractions (PARF; Frary et al. 2000; Wilson 1988), sum of adjacent LOD scores (SALOD; Weeks and Lange 1987), and likelihood of Lander and Green algorithm (Lander and Green 1987). Multiple locus linkage analysis is a procedure of simultaneously estimating the recombination fractions between adjacent two marker loci for a given order of more than two-marker loci in a linkage group. Lander and Green (1987) proposed an approach based on hidden Markov models (HMM) for multiple locus linkage analysis, which can calculate simultaneously not only the recombination fractions between adjacent marker loci but also the likelihood for a given order of multiple marker loci. From a statistical point of view, Lander and Green's algorithm can more accurately infer marker loci order and estimate the recombination fractions between flanking marker loci than other methods because multiple markers are simultaneously analyzed and incomplete marker data can be utilized (Lander et al. 1987).

Although constructing genetic linkage maps are difficult in oubred species, increasing efforts have been made to develop statistical methods for linkage analysis in a full-sib family, which can be derived by hybridizing two high heterozygous individuals that have some traits differing substantially. Such a full-sib family is crucial for constructing genetic linkage maps in some plant species, especially in forest trees, in which inbred lines are almost impossibly obtained due to the long generation times. Grattapaglia and Sederoff (1994) proposed a pseudo-testcross strategy for constructing linkage maps using a full-sib family, which just makes use of the 1:1 segregation molecular makers. Maliepaard et al. (1997) presented maximum likelihood estimates of the recombination fraction and LOD score formulas for all possible pairs of markers that segregate in the ratio of 1:1, 3:1, 1:2:1, or 1:1:1:1 in a full-sib family. Recently, Wu and Ma (2002) developed a likelihood method for simultaneously estimating the recombination fractions and linkage phase between any pair of markers in a full-sib family. Following Wu and Ma (2002), Lu et al. (2004) constructed a unifying likelihood analysis which can not only simultaneously estimate linkage and linkage phases but also gene order for a group of markers that may segregate in any possible ratio in a full-sib family. However, their method for multiple linkage analysis is limited to a small number of markers in a linkage group and not appropriate for dozens of or even hundreds of markers because either the number of linkage phase configurations or the number of all possible orders increases exponentially with the number of markers.

In this article, we present a multilocus linkage analysis via HMM for a given order of a linkage group of markers that can segregate in any possible ratios in a full-sib family, which can be generated from two outbred parents. Our method can simultaneously estimate the recombination fractions between adjacent markers and calculate the likelihood value for an order of a large number of markers. The likelihood of a given marker order can be used as an objective function for ordering maker loci of a linkage group. When the number of markers decreases to two or three, the multilocus linkage analysis becomes traditional two-point or three-point linkage analysis, respectively, by which the linkage phases can be estimated in a full-sib family. Monte Carlo simulations are performed to show that the recombination fraction estimates between adjacent markers using multilocus linkage analysis are more accurate than those just using two-point linkage analysis and that the likelihood as an objective function for ordering maker loci is the most powerful method for choosing the true marker order of a linkage group compared with other methods, such as SARF, SALOD, and the regression method used by JOINMAP (Stam 1993). A real example is presented for illustrating linkage maps constructed by using mixed segregation markers. Also, we have developed a Windows software, FsLinkageMap, by incorporating this multilocus linkage analysis for constructing genetic maps using molecular data from a full-sib family. Our multilocus linkage analysis will be beneficial to constructing high-density genetic linkage maps in some outcrossing plant species, especially in forest trees.

Statistical model

Segregation types, hidden states of genotypes, and linkage phases

In a diploid full-sib family, the number of different alleles may be two, three, or four at an informative marker locus in two parents. Maliepaard et al. (1997) summarized that the combinations of two parental genotypes at a informative marker locus, which were called segregation types, may be *ab*×*aa*, *aa*×*ab*, *ab*×*ab*, *ab*×*cd*, *ao*×*ao*, *ab*×*ao*, or *ao*×*ab*, where *a*, *b*, *c*, and *d* denote different alleles at a marker locus; *o* denotes the null allele; the two characters left of the crossing symbol represents the maternal marker genotype and the two characters on the right the paternal marker genotype. Here, a hidden state of a genotype is defined as a genotype of an offspring that has two ordered alleles with the first one from the maternal grandmother or grandfather and the second from the paternal grandmother or grandfather.

Thus, an offspring may have one of the four kinds of hidden states at a marker locus in a full-sib family. For example, if the two parents both have genotype *a*/*b*, where the slash symbol segregates the maternal allele (the left one) from the paternal allele (the right one), then a hidden state of an offspring may be *aa*, *ab*, *ba*, or *bb*, which are called hidden state 1, 2, 3, or 4 from now on, respectively. It is obvious that hidden state 1 is the genotype with the first allele from maternal grandmother and the second one from paternal grandmother, hidden state 2 the genotype with the first allele from maternal grandmother and the second one from paternal grandfather, hidden state 3 the genotype with the first allele from maternal grandfather and the second one from paternal grandmother, and hidden state 4 with the first allele from maternal grandfather and the second one from paternal grandfather. The hidden states are not directly observed because we generally cannot discriminate maternal allele from paternal allele in a genotype. However, the hidden states and genotypes have certain relationships that can be described by the conditional probability of genotype, M, on the hidden state, y, which is denoted by $B_i(M) = P(M|y = i)$, where i takes the value of 1, 2, 3, or 4. Tables 1 and 2 show the conditional probabilities of genotypes of a progeny on hidden states at a locus for segregation types *ab*×*ab* and *ab*×*a0*, respectively, under all possible phases of parental genotypes. The conditional probabilities for the rest five segregation types are omitted because they can be easily expressed in the same way. When a marker genotype of an offspring is missing, which is denoted by symbol "–", we define $B_i(-)=1$ since the genotype can be arbitrary given any hidden state i.

Table 1 Conditional probabilities of genotypes of a progeny on hidden states for segregation type *ab*×*ab* under all possible phases

Segregation type	Hidden state	Phenotypes		
		aa	*ab*	*bb*
a/*b*×*a*/*b*	1(*aa*)	1	0	0
	2(*ab*)	0	1	0
	3(*ba*)		1	0
	4(*bb*)	0	0	1
a/*b*×*b*/*a*	1(*ab*)	0	1	0
	2(*aa*)	1	0	0
	3(*bb*)	0	0	1
	4(*ba*)	0	1	0
b/*a*×*a*/*b*	1(*ba*)	0	1	0
	2(*bb*)	0	0	1
	3(*aa*)	1	0	0
	4(*ab*)	0	1	0
b/*a*×*b*/*a*	1(*bb*)	0	0	1
	2(*ba*)	0	1	0
	3(*ab*)	0	1	0
	4(*aa*)	1	0	0

Table 2 Conditional probabilities of genotypes of a progeny on hidden states for segregation type *ab*×*ao* under all possible phases

Segregation type	Hidden state	Phenotypes		
		a-	*ab*	*bo*
a/*b*×*a*/*o*	1(*aa*)	1	0	0
	2(*ao*)	1	0	0
	3(*ba*)	0	1	0
	4(*bo*)	0	0	1
a/*b*×*o*/*a*	1(*ao*)	1	0	0
	2(*aa*)	1	0	0
	3(*bo*)	0	0	1
	4(*ba*)	0	1	0
b/*a*×*a*/*o*	1(*ba*)	0	1	0
	2(*bo*)	0	0	1
	3(*aa*)	1	0	0
	4(*ao*)	1	0	0
b/*a*×*o*/*a*	1(*b0*)	0	0	1
	2(*ba*)	0	1	0
	3(*ao*)	1	0	0
	4(*aa*)	1	0	0

Linkage phases between any two-marker loci are not prior known in a full-sib family, whereas they are fixed in populations derived from inbred lines, such as back cross and F_2 intercross populations. Maliepaard et al. (1997) first pointed out that, in a full-sib family, linkage phase combinations included: (1) coupling (*c*) in the maternal parent and uninformative in the paternal parent, or vice versa; (2) repulsion (*r*) in the maternal parent and uninformative in the paternal parent, or vice versa; (3) coupling in both parents (*c*×*c*); (4) repulsion in both parents (*r*×*r*); and (5) coupling in the maternal parent and repulsion in the paternal parent (*c*×*r*), or vice versa (*r*×*c*). Hence, there may be two or four possible linkage phases between two informative marker loci in a full-sib family. For example, when the segregation types are all *ab*×*aa* at two-marker loci, the maternal parent has joint genotype *abab* and the linkage phase is either coupling (*aa*/*bb*) or repulsion (*ab*/*ba*), while the paternal parent has joint genotype *aaaa* with no recombination information. Another example is that, when the segregation types are *ab*×*ab* and *ab*×*cd* at two-marker loci, respectively, the linkage phase between the two loci is either *aa*/*bb*×*ac*/*bd* (*c*×*c*), or *aa*/*bb*×*ad*/*bc* (*c*×*r*), or *ab*/*ba*×*ac*/*bd* (*r*×*c*), or *ab*/*ba*×*ad*/*bc* (*r*×*r*). With regard to inferring the linkage phases between any pairs of loci, several statistical methods have been well documented (Lu et al. 2004; Maliepaard et al. 1997; Wu and Ma 2002).

Multilocus linkage analysis

Assume that M_1, M_2, ..., M_T are an ordered marker loci with known linkage phases between adjacent markers in a linkage group in a full-sib family. Let r_t denote the recombination fraction between markers M_t and M_{t+1}, $t = 1, 2, \cdots, T-1$ and $r = (r_1, r_2, \cdots, r_{T-1})$. The observed marker genotypes of the kth individual in a full-sib family, corresponding to the listed marker loci are denoted by O_1^k, O_2^k, ..., O_T^k, and all of the marker genotypes of n individuals at locus t are denoted by O_t, i.e., $O_t = \left(O_t^1, O_t^2, \cdots, O_t^n\right)$, $t = 1, 2, \cdots, T$. We also denote by y_1^k, y_2^k, ..., y_T^k the corresponding hidden states of the kth offspring at the ordered marker loci, respectively. Thus, we have constructed a bivariate Markov chain $\left(y_t^k, O_t^k\right)$, $t = 1, 2, \cdots, T$for the kth individual, where y^k is a Markov chain on its own and O^k is a sequence of random variables conditionally independent given the chain y^k. In fact, y^k cannot be observed, so it is called the "hidden" chain.

For our HMM model, there are four possible hidden states: 1, 2, 3, or 4. When the process moves from one state at marker M_t to the other at marker M_{t+1}, there occur one or two recombination events between the two markers. The transition matrix of probabilities from marker M_t to marker M_{t+1} is then

$$A(r_t) = \begin{pmatrix} (1-r_t)^2 & r_t(1-r_t) & r_t(1-r_t) & r_t^2 \\ r_t(1-r_t) & (1-r_t)^2 & r_t^2 & r_t(1-r_t) \\ r_t(1-r_t) & r_t^2 & (1-r_t)^2 & r_t(1-r_t) \\ r_t^2 & r_t(1-r_t) & r_t(1-r)_t & (1-r_t)^2 \end{pmatrix} \tag{1}$$

where the element, $a_{ij}(r_t) = P\left(y_{t+1}^k = j \middle| y_t^k = i\right)$, $i,j = 1, 2, 3, 4$.

In order to estimate the recombination fractions, r_t s and calculate the likelihood of the observed marker data, we use so-called forward and backward algorithms (Rabiner 1989). Define the forward and backward variables for the kth individual in a full-sib family as

$$\alpha_t^k(i) = P\left(O_1^k, O_2^k, \cdots, O_t^k, y_t^k = i\right) \qquad i = 1, 2, 3, 4$$

and

$$\beta_t^k(i) = P\left(O_{t+1}^k, O_{t+2}^k, \cdots, O_T^k \middle| y_t^k = i\right) \qquad i = 1, 2, 3, 4$$

respectively. Then, we have following recursive relationships:

$$\begin{cases} \alpha_1^k(i) = P\left(O_1^k, y_1^k = i\right) = P\left(y_1^k = i\right) b_i\left(O_1^k\right) = \pi_i b_i\left(O_1^k\right) & i = 1, 2, 3, 4 \\ \alpha_{t+1}^k(j) = \sum_{i=1}^{4} \alpha_t^k(i) P\left(y_{t+1}^k = j \middle| y_t^k = i\right) b_j\left(O_{t+1}^k\right) = \sum_{i=1}^{4} \alpha_t^k(i) a_{ij}(r_t) b_j\left(O_{t+1}^k\right) & j = 1, 2, 3, 4 \end{cases}$$

and

$$\begin{cases} \beta_T^k(i) = 1 \\ \beta_t^k(i) = \sum_{j=1}^{4} \beta_{t+1}^k(j) P\left(y_{t+1}^k = j \middle| y_t^k = i\right) b_j\left(O_{t+1}^k\right) \\ \qquad = \sum_{j=1}^{4} \beta_{t+1}^k(i) a_{ij}(r_t) b_j\left(O_{t+1}^k\right) \quad i = 1, 2, 3, 4. \end{cases}$$

where π_i is the initial probability of the hidden chain being in state i at marker 1, which here is defined to be uniform, that is π_i=0.25, i=1,2,3,4.

It can be easily deduced that the probability of the process being in state i at marker t and in state j at marker t+1, given the observed marker data O^k, is

$$\xi_t^k(i,j) = P\left(y_t^k = i, y_{t+1}^k = j \middle| O^k\right) = \frac{\alpha_t^k(i)\beta_{t+1}^k(j) b_j\left(O_{t+1}^k\right) a_{ij}(r_t)}{\sum_{i'=1}^{4}\sum_{j'=1}^{4} \alpha_t^k(i')\beta_{t+1}^k(j') b_j\left(O_{t+1}^k\right) a_{ij}(r_t)} \qquad i,j = 1, 2, 3, 4 \tag{2}$$

and that the likelihood of marker data O^k is

$$P\left(O^k \middle| r\right) = \sum_{i=1}^{4} \alpha_T^k(i) = \sum_{i=1}^{4} \alpha_t^k(i)\beta_t^k(i) \tag{3}$$

Therefore, the likelihood of all the marker data O of the n individuals can be expressed as

$$P(O|r) = \prod_{k=1}^{n}\sum_{i=1}^{4} \alpha_T^k(i) = \prod_{k=1}^{n}\sum_{i=1}^{4} \alpha_t^k(i)\beta_t^k(i) \tag{4}$$

To calculate the maximum likelihood estimates for r, we use EM algorithm (Dempster et al. 1977). Let $n_t(i, j)$ denote the number of recombination events between $y_t^k = i$ and $y_{t+1}^k = j$, then the elements of $n_t(i, j)$ give the following matrix:

$$\begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{pmatrix}$$

The E step involves computing the expectation of $n_t\left(y_t^k, y_{t+1}^k\right)$ given the observed data O^k and recombination

r_t for the kth individual:

$$E\left(n_t\left(y_t^k, y_{t+1}^k\right)\middle|O^k, r_t\right) = \sum_{i=1}^{4}\sum_{j=1}^{4}\xi_t^k(i,j)n_t(i,j) \tag{5}$$

and the M step,

$$\hat{r}_t = \frac{1}{2n}\sum_{k=1}^{n}E\left(n_t\left(y_t^k, y_{t+1}^k\right)\middle|O^k, r_t\right)$$
$$= \frac{1}{2n}\sum_{k=1}^{n}\sum_{i=1}^{4}\sum_{j=1}^{4}\xi_t^k(i,j)n_t(i,j) \quad t = 1,\ 2,\ \cdots,\ T-1 \tag{6}$$

which is the ratio of the expectation number of recombination haplotypes to all $2n$ haplotypes between marker t and t+1 in the full-sib family. In fact, formula (6) can be directly obtained by Baum's lemma (Baum et al. 1970), which will be shown in Appendix A.

By iteratively computing formulas (5) and (6) until convergence, we can obtain the maximum likelihood estimates of recombination fractions between adjacent markers in the given order of the T markers. Finally, the estimates of recombination fractions for each interval are transformed to genetic distances using a map function, i.e., Haldane's or Kosambi's map function (Haldane 1919; Kosambi 1944).

Two-point linkage analysis

In the section above, if we let the number of markers T=2, then the multilocus linkage analysis becomes two-point linkage analysis. Thus, the formula for iteratively computing the estimate of recombination fraction between any two markers with an assumed linkage phase is given as

$$\hat{r}_1 = \frac{1}{2n}\sum_{k=1}^{n}\sum_{i=1}^{4}\sum_{j=1}^{4}\xi_1^k(i,j)n_1(i,j) \tag{7}$$

where

$$\xi_1^k(i,j) = \frac{b_i\left(O_1^k\right)a_{ij}(r_1)b_j\left(O_2^k\right)}{\sum_{i'=1}^{4}\sum_{j'=1}^{4}b_{i'}\left(O_1^k\right)a_{i'j'}(r_1)b_{j'}\left(O_2^k\right)} \quad i,j = 1,\ 2,\ 3,\ 4$$

and the likelihood for the two-loci marker data is

$$P(O|r_1) = \prod_{k=1}^{n}\sum_{j=1}^{4}\alpha_2^k(j) = \frac{1}{4^n}\prod_{k=1}^{n}\sum_{i=1}^{4}\sum_{j=1}^{4}b_i\left(O_1^k\right)a_{ij}(r_1)b_j\left(O_2^k\right) \tag{8}$$

by noting that $\alpha_1^k(i) = 1/4b_i\left(O_1^k\right)$ and $\beta_2^k(i) = 1$ for i=1, 2, 3, 4.

To test the null hypothesis $H_0 : r_1 = 0.5$ that there exists no linkage between the two markers versus the alternative H_1:r_1<0.5 that the two markers are linked, we often use LOD score as the test statistic,

$$LOD = \log_{10}\frac{P\left(O|\hat{r}_1\right)}{P(O|r_1 = 0.5)} \tag{9}$$

A LOD score of 3.0 is often used as the threshold for asserting that the pair of markers is linked.

As to inferring the linkage phase for a pair of markers, Maliepaard et al. (1997) suggested that the correct linkage phase should correspond to a significant LOD score and a legitimate estimate of recombination fraction (i.e., $\hat{r}_1 < 0.5$). Currently, a Bayesian method was proposed for inferring the most likely linkage phase between any two markers (Lu et al. 2004; Wu and Ma 2002). However, there is some uncertainty in characterizing linkage phase for a pair of dominant markers. In such a case, the true linkage phase of c×c can be correctly selected from the four possible linkage phases according to the likelihood criterion. Lu et al. (2004) found that the true linkage phase of r×r can also be discriminated from other three linkage phases since c×c would give an illegitimate estimation of recombination fraction and c×r or r×c would be related to a lower estimate of recombination fraction. Unfortunately, we cannot distinguish between the linkage phases of c×r and r×c because the marker genotypes share the same probability distribution under the two linkage phases. Therefore, two adjacent dominant markers are not recommended to use for multilocus linkage analysis if the linkage phase is possibly c×r or r×c.

Monte Carlo simulation

We perform simulation studies to investigate the accuracy of estimating the recombination fractions between adjacent markers using multilocus analysis based on our HMM and the ordering power using the likelihood as an objective function comparing with other methods, such as SARF, SALOD, and the regression mapping method used in JOINMAP 4.0 (JM-REG; Van Ooijen 2006). Assume that there are eight markers with a known order on a chromosome in a full-sib family. The segregation types of these eight markers are shown in the second column of Table 3, which includes three kinds of segregation ratios: 1:1, 1:2:1, and 1:1:1:1. The linkage phases and the recombination fractions between the ith and (i+1)th marker are assumed as in the third and fourth columns of Table 3. In order to examine the effects of sample size and marker quality on the estimates of recombination fractions and the ordering powers of several methods, we simulate a full-sib family with a small sample size 100 and a modest sample

size 300 and with different marker qualities: 0%, 5%, 10%, and 15% missing marker data. For each combination of sample size and marker quality, the simulation is repeated 1,000 times.

As we expected, the estimates of recombination fractions using multilocus linkage analysis based on our HMM are more accurate than those using just two-point linkage analysis. The means and standard deviations of recombination fractions between adjacent markers based on multilocus and two-point linkage analyses are shown in Table 4 with sample size 100 and in Table 5 with sample size 300. The means of the recombination fraction estimates of each marker interval by the two linkage analytical methods are almost the same and equal to the true value that we set previously, but the standard deviations of the recombination fraction estimates by the two methods give different results. The recombination fraction estimates of middle intervals by multilocus linkage analysis have smaller deviations than those by two-point linkage analysis, whereas the deviations of recombination fractions estimates for the most left and the most right intervals by multilocus linkage analytical method display no apparent advantages over those by two-point linkage analysis. The result is similar for any combination of sample size and marker data quality, and can be seen clearly by Figs. 1 and 2. The reason is that estimating recombination fractions for the middle intervals by multilocus linkage analysis can utilize the information of both the left-side and the right-side markers of those intervals, while the estimate of the most left interval or the most right interval can only use one side marker information of that interval. Therefore, estimates based on multilocus linkage analysis are more accurate than those based on two-point linkage analysis because the former can utilize much more information of markers than the latter.

As an ordering objective function, the likelihood of ordered markers based on HMM is the most powerful method for choosing the true marker order of a linkage group in a full-sib family, followed by SARF, SALOD, and JM-REG. The power here is defined as the number of runs out of 1,000 replicates in which the true order of the eight markers is estimated with an ordering method. We use exhausting algorithm to search the true order among the 8!/2=20,160 possible orders of the eight markers for each run by the methods of the likelihood, SARF and SALOD, respectively. For the JM-REG method, we directly use the software, JOINMAP 4.0, to obtain the estimated order of the eight markers for each run. Figure 3 shows the plots of powers of the four different ordering methods versus marker quality cases for the eight markers with sample sizes 100 and 300. It is no doubt that the power with modest sample size 300 is higher than that with small sample size 100 for each ordering method across marker data qualities and that the power of each method decreases as the marker data quality decreases. The power of the likelihood method is overwhelming advantage over other three methods. It is almost 100% with modest sample size 300 and more than 90% with small sample size 100 across different marker data qualities. The power of SARF is very close to that of the likelihood method when the sample size is modest, but the difference between the powers of the two methods is apparently large when the sample size is small. The performance of SALOD and JM-REG for ordering markers is very poor among the four ordering methods. The power of SALOD is relatively stable over marker data qualities, but has a little more improvement when the sample size is changed from 100 to 300. However, the performance of JM-REG is very sensitive to both the marker data quality and the sample size. When the sample size is 100, the power of JM-REG decreases abruptly from 0.7 to 0.45 as the marker data quality is changed from non-missing case to 15% missing case. Moreover, the power of JM-REG averagely increases almost 0.3 with the sample size changing from 100 to 300. This sensitive characteristic for JM-REG was reported by Van Os et al. (2005), where the ordering performances between the software JOINMAP and RECORD were compared by using simulation data in a backcross population.

A real example

We use an example of a forest tree to illustrate our statistical model for multilocus linkage analysis in a full-sib family.

Table 3 Assumed eight markers of a known order with different segregation types, linkage phases, and recombination fractions between adjacent markers on a chromosome in a full-sib family

Marker locus	Segregation type	Linkage phase	Recombination fraction
1	*ab*×*aa*	*r*	0.14
2	*ab*×*cd*	*c*	0.07
3	*aa*×*ab*	*c*	0.15
4	*aa*×*ab*	*c*	0.08
5	*ab*×*ab*	*r*	0.20
6	*aa*×*ab*	*c*	0.12
7	*ab*×*cd*	*r*×*c*	0.10
8	*ab*×*ab*	–	–

Table 4 Means and standard deviations of recombination fraction estimates between adjacent markers for the eight ordered markers on a chromosome with sample size 100 and different marker data qualities in a full-sib family based on 1,000 replicates

Marker interval	Non-missing data		5% missing data		10% missing data		15% missing data	
	Two-point	Multipoint	Two-point	Multipoint	Two-point	Multipoint	Two-point	Multipoint
1	0.1400±0.0344	0.1400±0.0344	0.1398±0.0367	0.1399±0.0367	0.1387±0.0379	0.1389±0.0379	0.1385±0.0411	0.1388±0.0407
2	0.0703±0.0254	0.0704±0.0245	0.0700±0.0268	0.0702±0.0258	0.0707±0.0286	0.0705±0.0272	0.0711±0.0295	0.0709±0.0280
3	0.1507±0.0355	0.1510±0.0327	0.1504±0.0381	0.1512±0.0372	0.1510±0.0396	0.1513±0.0352	0.1513±0.0413	0.1513±0.0361
4	0.0794±0.0378	0.0798±0.0299	0.0786±0.0399	0.0799±0.0319	0.0784±0.0423	0.0788±0.0340	0.0807±0.0459	0.0793±0.0370
5	0.2006±0.0563	0.1996±0.0357	0.2001±0.0593	0.1990±0.0372	0.2001±0.0638	0.1998±0.0389	0.2004±0.0655	0.1988±0.0398
6	0.1200±0.0317	0.1196±0.0302	0.1202±0.0334	0.1200±0.0317	0.1196±0.0355	0.1196±0.0331	0.1219±0.0382	0.1212±0.0354
7	0.1012±0.0226	0.1012±0.0226	0.1011±0.0242	0.1011±0.0241	0.1013±0.0250	0.1013±0.0248	0.1014±0.0268	0.1013±0.0265

The results of multilocus linkage analysis are compared with those of two-point linkage analysis

Table 5 Means and standard deviations of recombination fraction estimates between adjacent markers for the eight ordered markers on a chromosome with sample size 300 and different marker data qualities in a full-sib family based on 1,000 replicates

Marker interval	Non-missing data		5% missing data		10% missing data		15% missing data	
	Two-point	Multipoint	Two-point	Multipoint	Two-point	Multipoint	Two-point	Multipoint
1	0.1395±0.0197	0.1395±0.0197	0.1394±0.0206	0.1394±0.0206	0.1393±0.0219	0.1393±0.0218	0.1399±0.0234	0.1399±0.0233
2	0.0699±0.0146	0.0698±0.0140	0.0702±0.0151	0.0701±0.0145	0.0698±0.0162	0.0697±0.0155	0.0695±0.0174	0.0694±0.0165
3	0.1493±0.0201	0.1492±0.0184	0.1491±0.0210	0.1490±0.0187	0.1495±0.0224	0.1493±0.0198	0.1491±0.0238	0.1492±0.0211
4	0.0799±0.0223	0.0797±0.0176	0.0799±0.0238	0.0795±0.0187	0.0799±0.0255	0.0794±0.0199	0.0810±0.0264	0.0809±0.0205
5	0.2007±0.0313	0.2008±0.0214	0.2009±0.0331	0.2007±0.0221	0.2004±0.0352	0.2010±0.0227	0.2009±0.0371	0.2003±0.0242
6	0.1185±0.0196	0.1183±0.0185	0.1187±0.0208	0.1182±0.0194	0.1184±0.0216	0.1181±0.0203	0.1181±0.0228	0.1181±0.0216
7	0.1003±0.0125	0.1003±0.0125	0.1003±0.0128	0.1002±0.0128	0.1001±0.0142	0.1001±0.0141	0.1004±0.0148	0.1004±0.0147

The results of multilocus linkage analysis are compared with those of two-point linkage analysis

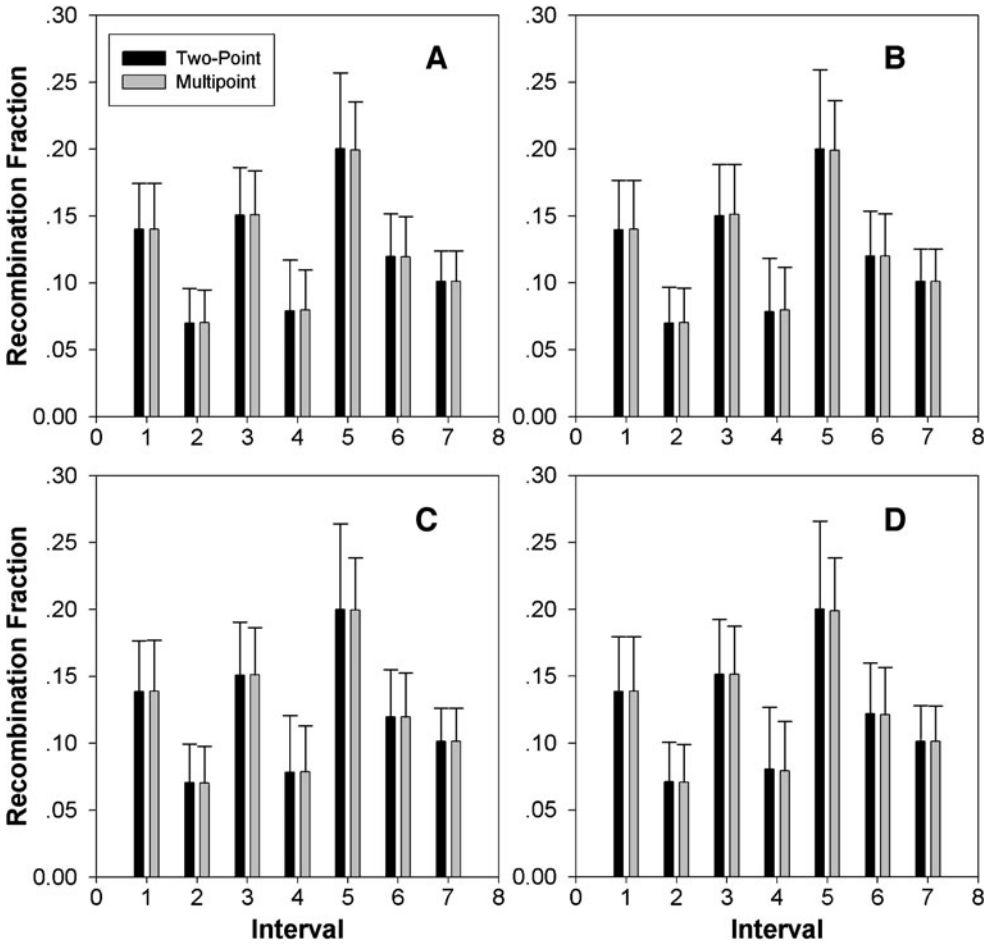

Fig. 1 Means of estimates of recombination fractions with *error bars* for the seven marker intervals based on sample size 100 and 1,000 replicates. The *left column* and *bar* denote the mean and SD of recombination fraction (RF) estimates based on two-point linkage analysis, and the *right column* and *bar* denote the mean and SD of RF estimates based on multilocus linkage analysis, for each marker interval. There are four cases for marker data qualities: **a** non-missing, **b** 5% missing, **c** 10% missing, and **d** 15% missing data

The material was an interspecific F_1 hybrid population between *Populus deltoides* and *Populus euramericana*. A *P. deltoides* clone (designated I-69) was used as a female parent to mate with a *P. euramericana* clone (designated I-45) as a male parent. Both *P. deltoides* I-69 and *P. euramericana* I-45 were selected at the Research Institute for Poplars in Italy in the 1950s and were introduced to China in 1972. In the spring of 1988, a total of 450 1-year-old rooted hybrid seedlings were planted at a spacing of 4× 5 m at a forest farm near Xuchou City, Jiangsu Province, China. A total of 93 genotypes randomly selected from the 450 genotypes were used for constructing linkage maps. An integrated genetic linkage map has been constructed using JoinMap, including 652 markers (225 SSRs, 288 AFLPs, 120 RAPDs, 16 ISSRs, two SNPs, and one sexual differentiation trait; Zhang 2005). Here, we choose a linkage group of 31 markers for linkage mapping. The segregation types are *aa*×*ab* for ten markers, *ab*×*aa* for 13 markers, *ab*×*ab* for one marker, and *ab*×*cd* for seven markers. First, the recombination fractions were estimated and linkage phases were predicted between any two markers by performing two-point linkage analysis. And then, the optimal order of the 31 markers was found and the recombination fractions between adjacent markers were simultaneously re-estimated over all possible orders by multilocus linkage analysis. Figure 4 shows the genetic linkage map for the 31 markers, where the recombination fractions are converted to genetic distances by Haldane's map function.

Discussion

Genetic linkage maps are important tools for locating QTL, and currently, there are several successful examples of finding QTLs with being isolated, cloned, and sequenced-based on QTL mapping (El-Din El-Assal et al. 2001; Frary et al. 2000; Li et al. 2006; Ren et al. 2005). Statistical methods for constructing genetic linkage maps in inbred lines have been well developed for the past 20 years, but these methods are difficult to establish in outbred species because of their complicated genetic structure. However, there are some species, such as forest trees, in which inbred lines are almost impossibly obtained due to the long generation times, whereas a full-sib family with large individuals can be easily obtained by hybridizing two individuals that have some traits of interest differing

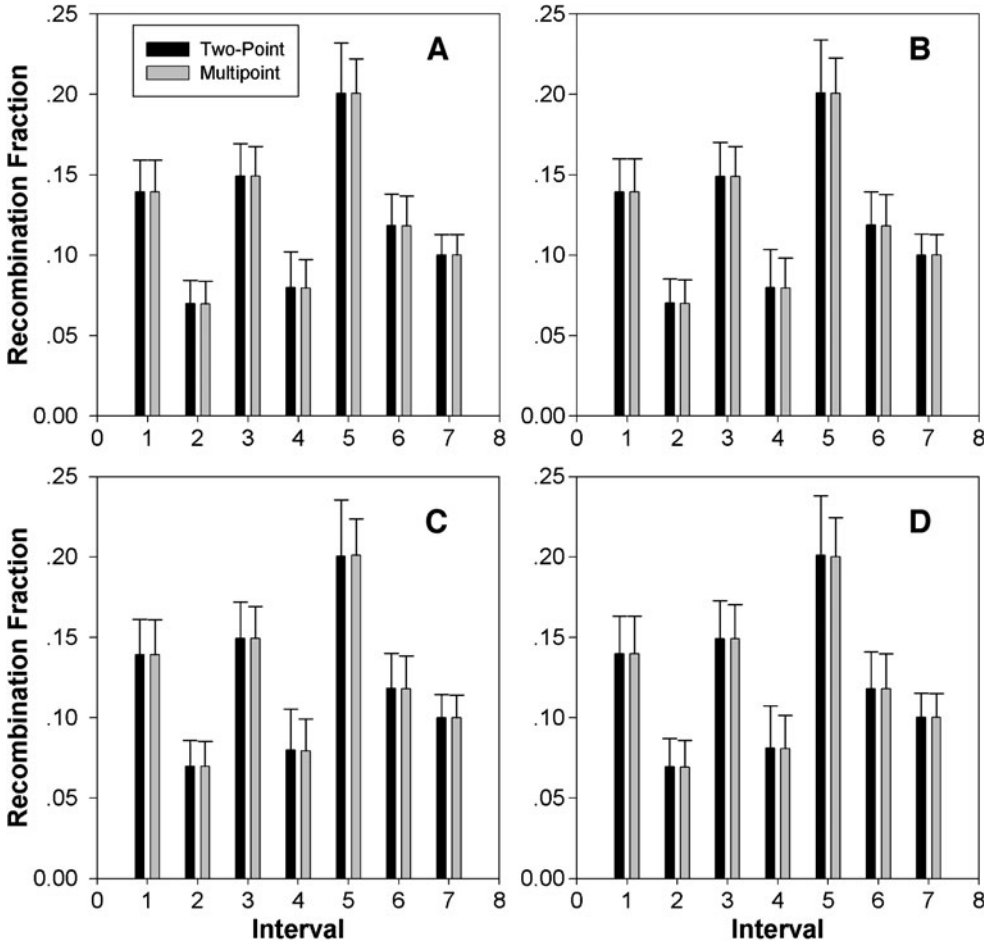

Fig. 2 Means of estimates of recombination fractions with *error bars* for the seven marker intervals based on sample size 300 and 1,000 replicates. The *left column* and *bar* denote the mean and SD of recombination fraction (RF) estimates based on two-point linkage analysis, and the *right column* and *bar* denote the mean and SD of RF estimates based on multilocus linkage analysis, for each marker interval. There are four cases for marker data qualities: **a** non-missing, **b** 5% missing, **c** 10% missing, and **d** 15 missing data

substantially. Such a large full-sib family is a good material for constructing genetic linkage maps in these species. Previous statistical work for linkage analysis in a full-sib family is limited to estimating the recombination fraction and inferring the linkage phase between any two markers (Maliepaard et al. 1997; Wu and Ma 2002). Lu et al. (2004) proposed a general model for simultaneously estimating linkage, parental diplotype, and marker order through multipoint analysis, but their method can only handle several markers in a linkage group because the computing time is prohibited as the number of markers becomes large. We proposed a hidden Markov model for multilocus

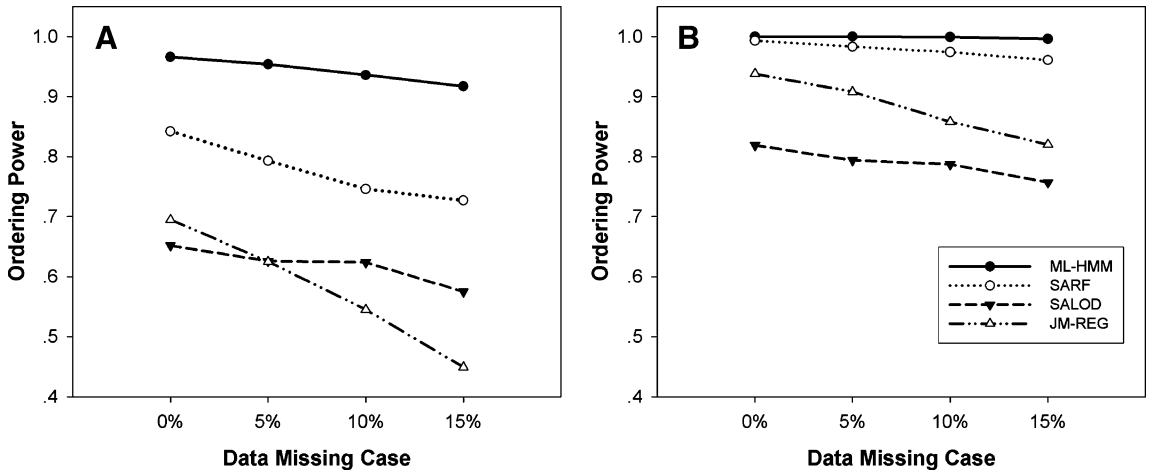

Fig. 3 Plots of ordering powers of the four different methods versus marker data quality cases for the eight markers on a chromosome in a full-sib family with sample size **a** $n=100$ and **b** $n=300$

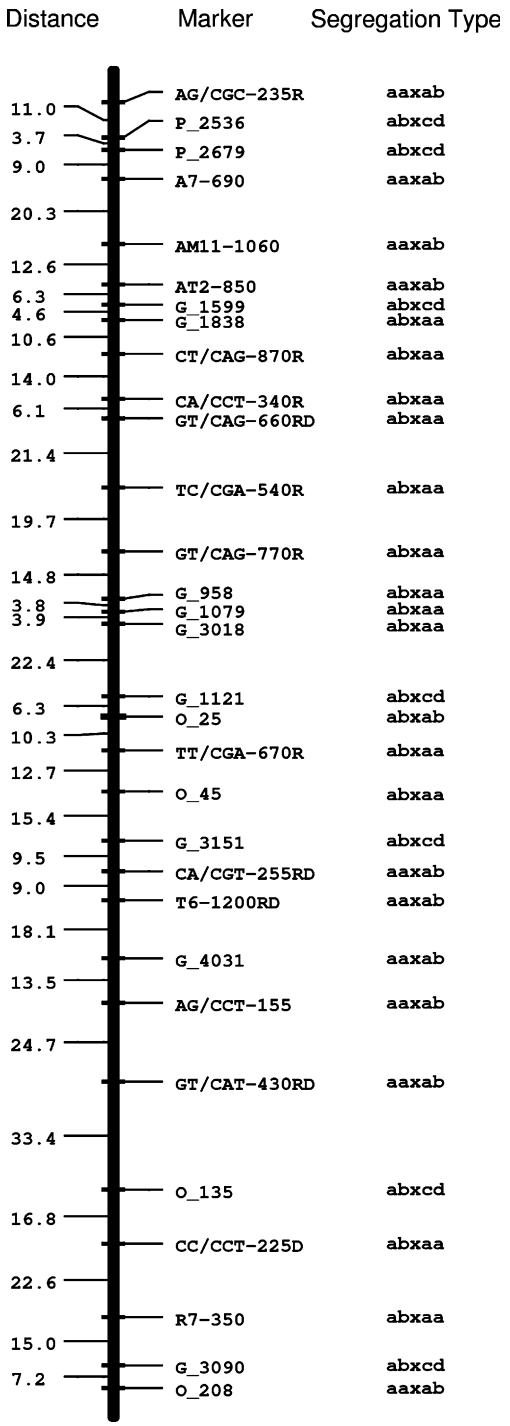

Fig. 4 A genetic map for a linkage group of 31 markers from a full-sib family that generated by "hybriding" *P. deltoides* and *P. euramericana*. There are four segregation types for these markers

linkage analysis in a full-sib family in this paper. Our approach has two advantage aspects compared with previous work and will be beneficial to accurately constructing genetic linkage maps with high density of any two parents that can be crossed to generate a full-sib family with large individuals.

First, our multilocus linkage analysis based on HMM in a full-sib family can provide much more linkage information than just two-point linkage analysis. Our approach cannot only deal with any type of segregation markers but also incomplete marker data in a full-sib family. Moreover, all markers in the linkage group, including missing markers, are incorporated in the multilocus linkage analysis. Therefore, the recombination fractions between adjacent markers can be estimated more accurately and stably by multilocus linkage analysis based on HMM than by just two-point linkage analysis, as shown in Figs. 1 and 2. In the meantime, we obtained a uniform formula (7) for two-point linkage analysis by just considering two markers in the multilocus linkage analysis. It is the simplest formula for estimating the recombination fraction between any two markers in a full-sib family. Contrarily, Maliepaard et al. (1997) summarized 17 formulas for estimating recombination fractions each for a pairwise combination of marker segregation types.

Second, and most importantly, the maximum likelihood of ordered markers can be as an objective function for ordering a linkage group of markers, which has much higher power than other methods even if the sample size is small. Marker ordering for a linkage group is a key step for constructing genetic linkage maps because marker order is more important than marker distances, especially when there are large markers in a linkage group (Van Os et al. 2005). However, there is a little attention to marker ordering for a linkage group. Wu et al. (2003) compared five algorithms, ML, SARF, SALOD, PARF, and seriation, for their efficiencies of ordering markers based on doubled haploid populations. The result indicated that marker-ordering powers for the five methods were almost identical. Van Os et al. (2005) compared their method, RECORD, with that of JOINMAP for ordering markers based on BC populations. RECORD, which is equivalent to SARF for perfect data, is much faster and less sensitive to missing data. We have compared four methods, ML-HMM, SARF, SALOD, and JM-REG, for their ordering powers by simulation data from a full-sib family.

By incorporating our multilocus linkage analysis, we have developed a Windows software, FsLinkageMap, for constructing genetic linkage maps with molecular markers generated from a full-sib family. For a linkage group of less than nine markers, FsLinkageMap performs exhaustive search method to find the optimal order that has the maximum likelihood value of all possible orders. For more than eight markers in a linkage group, FsLinkageMap has to resort to heuristic search method because of huge number of possible marker orders. As indicated in our simulation results, FsLinkageMap has an advantage over the extensively used software, JOINMAP, in linkage group ordering. Therefore, it provides a powerful tool for constructing high-

quality genetic linkage maps in some outcrossing species, especially in forest trees. The software is available and can be freely downloaded from its web page: http://fgbio.njfu.edu.cn/tong/FsLinkageMap/FsLinkageMap.htm.

Acknowledgements We thank the anonymous reviewer and the associate editor for their constructive comments on the manuscript. This work was supported by the National Natural Science Foundation of China (No. 30872051) and the Natural Science Foundation of Jiangsu Province, China (No. BK2008422).

Appendix A

Following Amstrong's deriving procedure (Armstrong 2001), we define

$$Q(r,\ r') = \sum_{y} P(y,\ O|r) \log P(y,\ O|r'). \tag{10}$$

Then, we have

$$\begin{aligned} Q(r,\ r') &= E_r[\log P(y,\ O|r')] \\ &= E_r\left[\sum_{k=1}^{n} \log P\left(y^k,\ O^k|r'\right)\right] \\ &= \sum_{k=1}^{n}\sum_{y^k} P\left(y^k|O^k,\ r\right) \log P\left(y^k,\ O^k|r'\right) \\ &= \sum_{k=1}^{n}\sum_{y^k} P\left(y_1^k,\cdots,y_T^k|O^k,\ r\right)\left[\log P\left(y_1^k\right)\right. \\ &\quad + \log P\left(y_2^k|y_1^k,r_1'\right) + \cdots + \log P\left(y_T^k|y_{T-1}^k,r_{T-1}'\right) \\ &\quad \left. + \log P\left(O_1^k|y_1^k\right) + \cdots + \log P\left(O_T^k|y_T^k\right)\right] \end{aligned} \tag{11}$$

In terms of r_t', the above can be expressed as

$$\begin{aligned} &\sum_{k=1}^{n}\sum_{y_t^k,y_{t+1}^k} P\left(y_t^k,y_{t+1}^k|O^k,\ r_t\right) \log P\left(y_{t+1}^k|y_t^k,r_t'\right) \\ &= \sum_{k=1}^{n}\sum_{i,j=1}^{4} P\left(y_t^k=i,\ y_{t+1}^k=j|O^k,\ r_t\right) \\ &\quad \log\left[r_t'^{n_t}(i,j)\left(1-r_t'\right)^{2-n_t(i,j)}\right] \\ &= \sum_{k=1}^{n}\left[\sum_{i,j}\xi_t^k(i,j)n_t(i,j)\log\frac{r_t'}{1-r_t'} + 2\log\left(1-r_t'\right)\right] \end{aligned} \tag{12}$$

By Baum's lemma, to maximum the likelihood (Eq. 4) is equivalent to maximizing (Eq. 12) with respect to r_t'. Therefore, differentiating (12) and setting it to zero, we obtain the likelihood estimate:

$$\hat{r}_t = \frac{1}{2n}\sum_{k=1}^{n}\sum_{i,j=1}^{4}\xi_t^k(i,j)n_t(i,j) \tag{13}$$

References

Armstrong N (2001) Incorporating interference into the linkage analysis of experimental crosses. Ph. D. thesis. Berkeley: University of California

Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Stat 41:164–171

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39:1–38

El-Din El-Assal S, Alonso-Blanco C, Peeters AJ, Raz V, Koornneef M (2001) A QTL for flowering time in *Arabidopsis* reveals a novel allele of CRY2. Nat Genet 29:435–440

Falk CT (1989) A simple scheme for preliminary ordering of multiple loci: application to 45 CF families. In: Elston RC, Spence MA, Hodge SE, MacCluer JW (eds) Multipoint mapping and linkage based upon affected pedigree members, Genetic Workshop 6. Liss, New York, pp 17–22

Frary A, Nesbitt TC, Frary A, Grandillo S, van der Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert KB, Tanksley SD (2000) Fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 289:85–88

Grattapaglia D, Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. Genetics 137:1121–1137

Haldane JBS (1919) The combination of linkage values and the calculation of distance between the loci of linked factors. J Genet 8:299–309

Jensen J, Helms Jørgensen J (1975) The barley chromosome 5 linkage map. Hereditas 80:17–26

Kosambi DD (1944) The esimation pf map distances from recombination values. Ann Eugen 12:172–175

Lalouel JM (1977) Linkage mapping from pair-wise recombination data. Heredity 38:61–77

Lander ES, Green P (1987) Construction of multilocus genetic maps in humans. Proc Natl Acad Sci USA 84:2363–2367

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174–181

Li CB, Zhou AL, Sang T (2006) Rice domestication by reducing shattering. Science 311:1936–1939

Lu Q, Cui YH, Wu RL (2004) A multilocus likelihood approach to joint modeling of linkage, parental diplotype and gene order in a full-sib family. BMC Genetics 5:20

Maliepaard C, Jansen J, Van Ooijen JW (1997) Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications. Genet Res 70:237–250

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286

Ren ZH, Gao JP, Li LG, Cai XL, Huang W, Chao DY, Zhu MZ, Wang ZY, Luan S, Lin HX (2005) A rice quantitative trait locus for salt tolerance encodes a sodium transporter. Nat Genet 37:1029–1030

Stam P (1993) Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. Plant J 3:739–744

Van Ooijen JW (2006) JoinMap 4, software for the calculation of genetic linkage maps in experimental populations. Kyazma BV, Wageningen, The Netherlands

Van Os H, Stam P, Visser RGF, Van Eck HJ (2005) RECORD: a novel method for ordering loci on a genetic linkage map. Theor Appl Genet 112:30–40

Weeks D, Lange K (1987) Preliminary ranking procedures for multilocus ordering. Genomics 1:236–242

Wilson SR (1988) A major simplification in the preliminary ordering of linked loci. Genet Epidemiol 5:75–80

Wu RL, Ma CX (2002) Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. Theor Popul Biol 61:349–363

Wu J, Jenkins J, Zhu J, McCarty J, Watson C (2003) Monte Carlo simulations on marker grouping and ordering. Theor Appl Genet 107:568–573

Zhang B (2005) Constructing genetic linkage maps and mapping QTLs affecting important traits in poplar. Ph. D. Dissertation, Nanjing Forestry University, Nanjing, China: Available at : http://fgbio.njfu.edu.cn/tong/zhang2005.pdf