

RESEARCH ARTICLE

A new strategy for estimating two-locus recombination fractions under some natural inequality restrictions

YING ZHOU^{1*}, WEIJUN MA¹, XIAONA SHENG² and HUAKUN WANG¹

¹*School of Mathematical Sciences, Heilongjiang University, Harbin 150080, People's Republic of China*

²*School of Science, Harbin University, Harbin 150086, People's Republic of China*

Abstract

Linkage analysis is now being widely used to map markers on each chromosome in the human genome, to map genetic diseases, and to identify genetic forms of common diseases. Two-locus linkage analysis and multi-locus analysis have been investigated comprehensively, and many computer programs have been developed to perform linkage analysis. Yet there exists a shortcoming in traditional methods, i.e., the parameter space of two-locus recombination fractions has not been emphasized sufficiently in the usual analyses. In this paper, we propose a new strategy for estimating the two-locus recombination fractions based on data of backcross family in the framework of some natural and necessary parameter restrictions. The new strategy is based on a restricted projection algorithm, which can provide fast reasonable estimates of recombination fraction, and can therefore serve as a superior alternative algorithm. Results obtained from both real and simulated data indicate that the new algorithm performs well in the estimation of recombination fractions and outperforms current methods.

[Zhou Y., Ma W., Sheng X. and Wang H. 2011 A new strategy for estimating two-locus recombination fractions under some natural inequality restrictions. *J. Genet.* **90**, 275–282]

Introduction

Linkage analysis is used in both human and other biological studies. Lathrop *et al.* (1984) discussed the strategies for multi-locus linkage analysis in humans. Lander and Green (1987) considered the problem of construction of multi-locus genetic linkage maps in humans. Ridout *et al.* (1998) investigated three-locus linkage in crosses of allogamous plant species. Among all these researches, the method of maximum likelihood estimate (MLE; see Fisher 1935) plays an important role. At the same time, to perform linkage analysis, many efficient algorithms have been proposed. Elston and Stewart (1971) developed a general algorithm for calculating the complex likelihood function; Ott (1976) wrote a useful computer program, LIPED, which allowed researchers to efficiently determine the recombination fraction between pairs of genetic loci.

Recombination fraction is a key parameter to be investigated in the process of linkage analysis (Ott 1976; Lathrop 1985; Li *et al.* 2006; Wu *et al.* 2007). Although the essence of the recombination fraction is a probability that describes the possibility of occurrence of a recombinational event between

two genetic loci, it can measure the distance between them to a certain degree (Haldane 1919). Therefore, recombination fraction is also called position parameter, the estimate of which will help to investigate mode of inheritance, gene mapping, etc.

A natural approach for the detection of linkage is two-locus linkage analysis, and the classical method of LOD score (Morton 1955) or relative techniques are still widely used. Later, two-locus analysis was extended to three-locus analysis, or multi-locus analysis (Meyers *et al.* 1976), in order to determine a genetic map from results of linkage analysis, or to map genes of interest. Although various methods are available for inferring recombination fraction, the parameter space of two-locus recombination fractions is not considered sufficiently in various analysis (Zhou *et al.* 2008). Even in three-locus analysis, the parameter space of two-locus recombination fractions is not simple. Besides a two-locus recombination fraction belonging to the interval [0, 0.5], there are still other necessary and natural restrictions that need to be satisfied. Therefore, three-locus (or multi-locus) analysis involves the constrained parameter problem (Kudo 1963; Royle and Dykstra 1984; Shi *et al.* 2005). Zhou *et al.* (2008) considered the estimation of the two-locus recombination fractions under some natural and necessary

*For correspondence. E-mail: zhouy577@yahoo.com.cn.

Keywords. constrained parameter problems; linkage analysis; restricted projection algorithm; recombination fraction.

restrictions and proposed a restricted EM algorithm (REM), which works well in application.

In this paper, using backcross (BC) families we reconsider the estimation problem, and developed a new restricted estimation strategy, called restricted projection algorithm (RPA). The new algorithm gives estimate results through taking account of the natural inequality restrictions on the two-locus recombination fractions, and the computing speed of the RPA is faster than the REM in some degree. Moreover, the new method is also applicable to various different cases. Simulation studies and an example are used to validate the application of our method in practice.

Statement of problem and methods

Suppose the data are available for three loci A, B and C in a phase-unknown backcross family ($abc/abc \times Aa/Bb/Cc$), and let the three two-locus recombination fractions be denoted by θ_{AB} , θ_{BC} and θ_{AC} , respectively. Assume the gene loci order is known, i.e., A-B-C. The heterozygous parent in the family has four possible linkage phases: (i) ABC/abc , (ii) ABc/abC , (iii) AbC/aBc , (iv) Abc/aBC ; and he/she may produce eight possible haplotypes: 1- ABC , 2- ABc , 3- AbC , 4- Abc , 5- aBC , 6- aBc , 7- abC , 8- abc . The conditional probability with which one offspring receives a certain haplotype given each parental phase can be easily calculated. The probabilities are functions of θ_{AB} , θ_{BC} and θ_{AC} , and also functions of joint recombination fractions g_{ij} 's ($i = 0, 1$, and the subscript 1 represents recombination, and 0 represents nonrecombination), since they have one-to-one relationship (Zhou et al. 2008).

Here we consider the situation in which there are two offspring in each observed family. The observed data can be grouped into four classes (Zhou et al. 2008). We list the details in table 1. In table 1, pair (i, j) presents the haplotype pair of two offspring in an observed family, which are received from the heterozygous parent, respectively, and the two haplotypes also correspond to two detailed phenotypes. Each class probability is also provided in table 1.

We denote the total number of families observed as n , and the number of families which are grouped into class

k as n_k ($k = 1, 2, 3, 4$), so $\sum_{k=1}^4 n_k = n$. We assume that (n_1, n_2, n_3, n_4) is multinomially distributed with probability (p_1, p_2, p_3, p_4) . Zhou et al. (2008) recalled the existing unrestricted method (Ott 1999) for estimating two-locus recombination fractions based on the above type of data, and then proposed a restricted method, since the following inequality restrictions on parameters should be considered in the process of parameter estimation:

$$\begin{cases} \theta_{AB} \leq \theta_{AC}, \\ \theta_{BC} \leq \theta_{AC}, \\ \theta_{AC} \leq \theta_{AB} + \theta_{BC}, \\ \theta_{AC} \leq 1/2. \end{cases} \quad (1)$$

Next we represent a new restricted algorithm to handle this situation, in order to solve the problem with better statistical properties.

MLEs under restrictions

In this section, we propose a new approach to calculate MLEs of two-locus recombination fractions under restriction (1). Suppose each phase of the heterozygous parent occurs with equal probability $\frac{1}{4}$. The log-likelihood function of parameters can be written as follows:

$$l(\theta) = \sum_{k=1}^4 n_k \ln(p_k(\theta_{AB}, \theta_{BC}, \theta_{AC})),$$

where vector $\theta = (\theta_{AB}, \theta_{BC}, \theta_{AC})$. Directly maximizing the log-likelihood function will lead to the unrestricted MLE $\hat{\theta}^U$ of recombination fractions (Ott 1999). However, only when restriction (1) is satisfied, can we obtain reasonable and restricted MLE $\hat{\theta}^R = (\hat{\theta}_{AB}^R, \hat{\theta}_{BC}^R, \hat{\theta}_{AC}^R)$.

Our new algorithm is a restricted projection algorithm (RPA), which is intended for fast and efficient calculation of parameter estimation. The detail of the RPA is as follows:

Table 1. Classification for data of phase-unknown triple backcross families with two offspring.

k	$(i, j)^a$	p_k
1	(1,1), (2,2), (3,3), (4,4), (5,5), (6,6), (7,7), (8,8), (4,5), (3,6), (2,7), (1,8)	$g_{11}^2 + g_{10}^2 + g_{01}^2 + g_{00}^2$
2	(1,2), (3,4), (3,5), (1,7), (4,6), (2,8), (5,6), (7,8)	$2(g_{11}g_{10} + g_{01}g_{00})$
3	(2,3), (1,4), (1,5), (2,6), (3,7), (4,8), (6,7), (5,8)	$2(g_{11}g_{01} + g_{10}g_{00})$
4	(1,3), (2,4), (2,5), (1,6), (4,7), (3,8), (5,7), (6,8)	$2(g_{11}g_{00} + g_{10}g_{01})$
Total		1

^a (i, j) , i and j refer to the code of haplotype, corresponding to a phenotype each.

The Taylor expansion of the log-likelihood about the fixed point $\hat{\theta}^U$ (i.e., the unrestricted MLE of θ which can be obtained by Ott's (1999) method) is;

$$\begin{aligned} l(\theta) &= l(\hat{\theta}^U) + (\theta - \hat{\theta}^U)' \cdot \frac{\partial l(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}^U} \\ &\quad + \frac{1}{2} (\theta - \hat{\theta}^U)' \cdot \frac{\partial^2 l(\theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}^U} \cdot (\theta - \hat{\theta}^U) \\ &\quad + o(\|\theta - \hat{\theta}^U\|^2), \end{aligned} \quad (2)$$

where $\frac{\partial l(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}^U} = 0$, and assuming that $o(\|\theta - \hat{\theta}^U\|^2)$ can be ignored when sample size n is appropriate, so we have,

$$\begin{aligned} -l(\theta) &\approx -l(\hat{\theta}^U) + \frac{1}{2} (\theta - \hat{\theta}^U)' \\ &\quad \cdot \left[-\frac{\partial^2 l(\theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}^U} \right] \cdot (\theta - \hat{\theta}^U), \end{aligned}$$

then the restricted estimation problem may be written as:

$$\min (\theta - \hat{\theta}^U)' \cdot \left[-\frac{\partial^2 l(\theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}^U} \right] \cdot (\theta - \hat{\theta}^U), \quad (3)$$

$$\text{subject to } \theta \in D = \bigcap_{i=1}^4 D_i = \bigcap_{i=1}^4 \{\theta \mid \sum_{j=1}^3 a_{ij} \theta_j \leq b_i\},$$

where a_{ij} are the elements of the matrix:

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 1 \\ 0 & 0 & 1 \end{pmatrix},$$

$b_1 = b_2 = b_3 = 0, b_4 = 1/2$, and $\theta_1 = \theta_{AB}, \theta_2 = \theta_{BC}, \theta_3 = \theta_{AC}$. So D is exactly the restriction (1). Let matrix $\Sigma_2 = -\frac{\partial^2 l(\theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}^U} = (\sigma_{ij})$, where σ_{ij} ($i, j = 1, 2, 3$) is the i th line and the j th column element of Σ_2 , which are given by:

$$\begin{aligned} \sigma_{11} &= (2\theta_{AB} - 1)^2 \left(\frac{n_1}{p_1^2} + \frac{n_2}{p_2^2} + \frac{n_3}{p_3^2} + \frac{n_4}{p_4^2} \right) \\ &\quad - 2 \left(\frac{n_1}{p_1} + \frac{n_2}{p_2} - \frac{n_3}{p_3} - \frac{n_4}{p_4} \right) \Big|_{\theta=\hat{\theta}^U}, \end{aligned}$$

$$\begin{aligned} \sigma_{12} &= \sigma_{21} \\ &= (2\theta_{AB} - 1)(2\theta_{BC} - 1) \\ &\quad \times \left(\frac{n_1}{p_1^2} - \frac{n_2}{p_2^2} - \frac{n_3}{p_3^2} + \frac{n_4}{p_4^2} \right) \Big|_{\theta=\hat{\theta}^U}, \end{aligned}$$

$$\begin{aligned} \sigma_{13} &= \sigma_{31} \\ &= (2\theta_{AB} - 1)(2\theta_{AC} - 1) \\ &\quad \times \left(\frac{n_1}{p_1^2} - \frac{n_2}{p_2^2} + \frac{n_3}{p_3^2} - \frac{n_4}{p_4^2} \right) \Big|_{\theta=\hat{\theta}^U}, \end{aligned}$$

$$\begin{aligned} \sigma_{22} &= (2\theta_{BC} - 1)^2 \left(\frac{n_1}{p_1^2} + \frac{n_2}{p_2^2} + \frac{n_3}{p_3^2} + \frac{n_4}{p_4^2} \right) \\ &\quad - 2 \left(\frac{n_1}{p_1} - \frac{n_2}{p_2} + \frac{n_3}{p_3} - \frac{n_4}{p_4} \right) \Big|_{\theta=\hat{\theta}^U}, \end{aligned}$$

$$\begin{aligned} \sigma_{23} &= \sigma_{32} \\ &= (2\theta_{BC} - 1)(2\theta_{AC} - 1) \\ &\quad \times \left(\frac{n_1}{p_1^2} + \frac{n_2}{p_2^2} - \frac{n_3}{p_3^2} - \frac{n_4}{p_4^2} \right) \Big|_{\theta=\hat{\theta}^U}, \end{aligned}$$

$$\begin{aligned} \sigma_{33} &= (2\theta_{AC} - 1)^2 \left(\frac{n_1}{p_1^2} + \frac{n_2}{p_2^2} + \frac{n_3}{p_3^2} + \frac{n_4}{p_4^2} \right) \\ &\quad - 2 \left(\frac{n_1}{p_1} - \frac{n_2}{p_2} - \frac{n_3}{p_3} + \frac{n_4}{p_4} \right) \Big|_{\theta=\hat{\theta}^U}. \end{aligned}$$

Obviously, the matrix Σ_2 is positive definite, since the log-likelihood has the maximum value at $\hat{\theta}^U$. Here we define an inner product and a norm as:

$$(x, y) = x' \Sigma_2 y \quad \text{and} \quad \|x\| = (x, x)^{1/2}, \quad (4)$$

for any $x, y \in R^3$, so equation (3) is equivalent to:

$$\min \|\theta - \hat{\theta}^U\|, \quad \text{subject to } \theta \in D = \bigcap_{i=1}^4 D_i. \quad (5)$$

Hence, the maximum likelihood estimate $\hat{\theta}^R \in R^3$ in the restricted situations is just the projection of $\hat{\theta}^U$ with respect to the inner product we define in equation (4). Von Neumann (1950), Wiener (1955), Kudo (1963), and Shi *et al.* (2005) have studied similar projection problems. Dykstra (1983) develops an algorithm for projecting an element in a finite-dimensional inner product space on a closed convex cone K , when K can be written as $K_1 \cap K_2 \cap \dots \cap K_r$ and each K_i is also a closed convex cone. Dykstra's (1983) idea is that it is often easy to project onto the individual K_i , and this fact is helpful in finding the solution to more complicated problem of projecting onto K . Royle and Dykstra (1984) proves that Dykstra's (1983) algorithm actually converges correctly in an infinite-dimensional Hilbert space setting even when the K_i 's are replaced by arbitrary closed convex sets D_i 's.

It can be verified that the D_i 's in equation (3) are all closed convex sets. So, the estimation problem can be resolved by using Dykstra's (1983) idea and making some detailed modifications. From theorem 1 of Royle and Dykstra (1984), we know that the restricted estimate to be obtained is unique when we solve equation (5).

The algorithm:

cycle 1:

- i) project $\hat{\theta}^U$ onto D_1 and obtain $\theta_{11} = \hat{\theta}^U + I_{11}$,
- ii) project θ_{11} onto D_2 and obtain $\theta_{12} = \theta_{11} + I_{12}$,
- iii) project θ_{12} onto D_3 and obtain $\theta_{13} = \theta_{12} + I_{13}$,
- iv) project θ_{13} onto D_4 and obtain $\theta_{14} = \theta_{13} + I_{14}$.

After the first cycle, instead of projecting θ_{14} onto D_1 , we firstly remove the initial increment I_{11} and then project.

Specifically, the steps for the second cycle proceed is as follows:

cycle 2:

- i) project $\theta_{14} - I_{11}$ onto D_1 to obtain $\theta_{21} = \theta_{14} - I_{11} + I_{21}$,
- ii) project $\theta_{21} - I_{12}$ onto D_2 to obtain $\theta_{22} = \theta_{21} - I_{12} + I_{22}$,
- iii) project $\theta_{22} - I_{13}$ onto D_3 to obtain $\theta_{23} = \theta_{22} - I_{13} + I_{23}$,
- iv) project $\theta_{23} - I_{14}$ onto D_4 to obtain $\theta_{24} = \theta_{23} - I_{14} + I_{24}$.

Continue this routine, and remove the increment in the previous cycle associated with D_i , then we will obtain array $\{\theta_{ni}\}$ and $\{I_{ni}\}$, where $n \geq 1$ and $i = 1, 2, 3, 4$. (see Appendix for the calculation of θ_{ni} 's). The above procedure is iteratively carried out until convergence, i.e., when $|\theta_{n,4} - \theta_{n+1,4}| < 10^{-t_0}$, for some positive number t_0 . From theorem 2 of Royle and Dykstra (1984), we have $\|\theta_{n+1,4} - \hat{\theta}^R\| \rightarrow 0$, as $n \rightarrow \infty$, where $\hat{\theta}^R$ is the MLE of θ under constraints equation (1). Therefore, $\theta_{n+1,4}$ maybe interpreted as the restricted MLE of θ .

Cases for more offspring and unequal prior probabilities of linkage phases

Now we turn to more general cases. As is known, greater sample size will provide more information to factual statistical analysis. Likewise, more offspring in each family will provide more information for linkage analysis (Lathrop et al. 1984; Thompson 1984). Fortunately, the RPA is easy to extend to fit the data with multiple offspring in each observed family. Let the number of offspring in each family be m , and the number of classification for the observed data be $f(m)$. Then the log-likelihood function becomes $l(\theta) = \sum_{k=1}^{f(m)} n_k \ln(p_k(\theta_{AB}, \theta_{BC}, \theta_{AC}))$. The matrix $\Sigma_m = -\frac{\partial^2 l(\theta)}{\partial \theta^2} \big|_{\theta=\hat{\theta}^U}$ can be obtained correspondingly, the elements of which have similar expressions as those in case of two offspring, where the subscript m of Σ_m indicates that m offspring in each family is considered. Therefore, we need to project $\hat{\theta}^U$ onto $D = \bigcap_{i=1}^4 D_i$ with respect to the inner product $(x, y) = x' \Sigma_m y$ at this time, and the other steps of the RPA are similar with those for the case of two offspring, except replacing Σ_2 by Σ_m . Cases with greater numbers of offspring are completely analogous.

In analysis, in section title 'MLEs under restriction', each linkage phase of the heterozygous parent is thought to be equally possible. In a more general case, we can assign different prior probability h_i ($i = 1, 2, 3, 4$) to each phase. Of course, the classification of the observed data and the likelihood function will change correspondingly, and the elements of matrix Σ_2 will be functions of both $\theta = (\theta_{AB}, \theta_{BC}, \theta_{AC})$ and h_i 's when the proposed RPA is utilized. Repeating the similar procedure for case of equal prior probabilities; we can obtain the restricted MLEs of the two-locus recombination fractions correspondingly.

Table 2. The estimate results of recombination fractions for mice data by the unrestricted method and the RPA.

Recombination fraction	Unrestricted method	RPA
θ_{AB}	0.3167	0.3162
θ_{BC}	0.3942	0.3744
θ_{AC}	0.3634	0.3744

An example

Zhou et al. (2008) analysed a real data set that comprises of 134 individuals from a backcross of mice (Clemens et al. 2000). Here we use the proposed RPA to revisit that data set and estimate three two-locus recombination fractions among marker loci D2Mit365, D2Mit272 and D2Mit456 on the linkage map of chromosome 2. After classification according to the regularity in table 1, we obtain $n = 67$ two-offspring families, and in detail, $(n_1, n_2, n_3, n_4) = (21, 17, 14, 15)$. We use the new algorithm to the data, and the restricted MLEs of two-locus recombination fractions are listed in table 2. For comparison, the corresponding unrestricted MLEs are also presented in table 2. It is easy to find that the RPA provides reasonable estimates, whereas the unrestricted method is ineffectual.

Simulation studies

In this section, simulated data of two-offspring BC family are used for analysis. Here we evaluate the performance of the proposed RPA, and compare it with the unrestricted method and REM.

To show the performance of the proposed RPA, we consider six scenarios by designing different true values of recombination fraction $\theta_0 = (\theta_{AB}, \theta_{BC}, \theta_{AC})$, which includes various cases in practice (see table 3). Then data of $n = 300$ two-offspring families are simulated for analysis in each scenario, and we calculate $\hat{\theta}^R$ by the RPA, $\hat{\theta}^U$ by the unrestricted method, and $\tilde{\theta}^R$ by the REM. Repeating the whole process for $M = 1000$ times, we obtain the averages of $\hat{\theta}^R$, $\hat{\theta}^U$ and $\tilde{\theta}^R$ over 1000 replicates by the three methods (see table 3).

To better compare the three methods, we also take advantage of the following two measures of accuracy: (i) the standard derivations (SDs) of each estimate; (ii) the mean absolute error (MAE) of the each estimate, where we define $MAE = \sum_{i=1}^M (|\hat{\theta}_{AB}^R - \theta_{AB}| + |\hat{\theta}_{BC}^R - \theta_{BC}| + |\hat{\theta}_{AC}^R - \theta_{AC}|) / 3M$. The results of SDs and MAEs corresponding to each estimate are presented in tables 4 and 5, respectively.

Results of comparisons between RPA and unrestricted method

First, we find in the process of simulation that the unrestricted method gives some unreasonable results in each scenario, and as expected, $\hat{\theta}^R$ by the RPA over 1000 repli-

Table 3. The averages of estimates over 1000 replicates for data of 300 two-offspring families by various methods.

Scenario	Parameters			RPA			Unrestricted method			REM		
	θ_{AB}	θ_{BC}	θ_{AC}	$\hat{\theta}_{AB}^R$	$\hat{\theta}_{BC}^R$	$\hat{\theta}_{AC}^R$	$\hat{\theta}_{AB}^U$	$\hat{\theta}_{BC}^U$	$\hat{\theta}_{AC}^U$	$\tilde{\theta}_{AB}^R$	$\tilde{\theta}_{BC}^R$	$\tilde{\theta}_{AC}^R$
1	0.05	0.05	0.06	0.050	0.049	0.061	0.050	0.050	0.060	0.050	0.050	0.060
			0.075	0.051	0.050	0.075	0.050	0.050	0.075	0.050	0.050	0.075
			0.09	0.051	0.050	0.090	0.050	0.050	0.090	0.050	0.050	0.090
2	0.05	0.15	0.16	0.050	0.149	0.161	0.050	0.149	0.160	0.050	0.149	0.161
			0.175	0.051	0.151	0.176	0.050	0.150	0.175	0.050	0.150	0.174
			0.19	0.051	0.150	0.188	0.050	0.151	0.191	0.050	0.151	0.190
3	0.05	0.35	0.36	0.050	0.339	0.357	0.050	0.330	0.371	0.050	0.353	0.369
			0.375	0.050	0.343	0.369	0.050	0.329	0.384	0.050	0.353	0.378
			0.39	0.050	0.342	0.375	0.050	0.334	0.399	0.050	0.353	0.394
4	0.15	0.15	0.16	0.148	0.148	0.164	0.150	0.151	0.161	0.149	0.150	0.164
			0.225	0.151	0.150	0.226	0.150	0.150	0.225	0.150	0.150	0.225
			0.29	0.150	0.149	0.284	0.150	0.151	0.292	0.150	0.151	0.289
5	0.15	0.35	0.36	0.150	0.335	0.363	0.150	0.325	0.374	0.151	0.349	0.375
			0.425	0.150	0.341	0.426	0.151	0.325	0.431	0.151	0.353	0.427
			0.49	0.151	0.342	0.455	0.150	0.333	0.454	0.150	0.348	0.450
6	0.35	0.35	0.36	0.337	0.337	0.372	0.357	0.331	0.368	0.344	0.347	0.360
			0.425	0.346	0.345	0.413	0.358	0.332	0.427	0.353	0.352	0.431
			0.49	0.345	0.345	0.458	0.358	0.326	0.451	0.353	0.352	0.455

cates agree better with θ_0 than the averages of $\hat{\theta}^U$. Second, table 4 also shows that our RPA outperforms the unrestricted method for estimating two-locus recombination fractions in each simulated scenario. The estimates obtained by the RPA have smaller SDs than the unrestricted method, which is more obvious especially when at least one of the intervals of AB and BC is loosely linked. This indicates that the accuracy of estimates by the RPA is higher than the unrestricted

method. Third, compared to $\hat{\theta}^U$, $\hat{\theta}^R$ is closer to the true value θ_0 (each $\text{MAE}(\hat{\theta}^R)$ is less than the corresponding $\text{MAE}(\hat{\theta}^U)$ in table 5).

Results of comparisons between RPA and REM

Through comparing the results by the REM and the RPA methods (see tables 3 to 5), we find that the performance of

Table 4. The standard derivations of various estimates over 1000 replicates for data of 300 two-offspring families.

Scenario ^a	RPA			Unrestricted method			REM		
	$\hat{\theta}_{AB}^R$	$\hat{\theta}_{BC}^R$	$\hat{\theta}_{AC}^R$	$\hat{\theta}_{AB}^U$	$\hat{\theta}_{BC}^U$	$\hat{\theta}_{AC}^U$	$\tilde{\theta}_{AB}^R$	$\tilde{\theta}_{BC}^R$	$\tilde{\theta}_{AC}^R$
1	0.0089	0.0092	0.0097	0.0094	0.0095	0.0106	0.0089	0.0088	0.0095
	0.0093	0.0093	0.0115	0.0100	0.0102	0.0116	0.0090	0.0092	0.0114
	0.0093	0.0091	0.0125	0.0100	0.0093	0.0130	0.0091	0.0093	0.0127
2	0.0094	0.0180	0.0190	0.0097	0.0191	0.0206	0.0093	0.0180	0.0177
	0.0093	0.0183	0.0208	0.0100	0.0185	0.0227	0.0091	0.0182	0.0195
	0.0094	0.0185	0.0211	0.0094	0.0185	0.0237	0.0094	0.0183	0.0209
3	0.0094	0.0391	0.0431	0.0095	0.1964	0.0708	0.0095	0.0463	0.0481
	0.0093	0.0401	0.0447	0.0099	0.1989	0.0778	0.0090	0.0464	0.0482
	0.0093	0.0399	0.0454	0.0093	0.1721	0.0813	0.0093	0.0445	0.0467
4	0.0161	0.0162	0.0176	0.0197	0.0200	0.0204	0.0156	0.0168	0.0168
	0.0175	0.0173	0.0249	0.0182	0.0177	0.0260	0.0181	0.0176	0.0239
	0.0175	0.0177	0.0272	0.0176	0.0189	0.0431	0.0174	0.0187	0.0261
5	0.0176	0.0389	0.0485	0.0177	0.2317	0.0761	0.0177	0.0452	0.0514
	0.0176	0.0391	0.0571	0.0180	0.2313	0.0791	0.0179	0.0459	0.0504
	0.0176	0.0412	0.0838	0.0182	0.2042	0.0863	0.0179	0.0410	0.0584
6	0.0365	0.0369	0.0472	0.0783	0.2406	0.0725	0.0390	0.0373	0.0454
	0.0382	0.0377	0.0543	0.0661	0.1887	0.0777	0.0454	0.0436	0.0498
	0.0384	0.0385	0.0825	0.0629	0.2114	0.0853	0.0460	0.0465	0.0577

^aThe scenarios are same with those in table 3.

Table 5. The mean absolute errors of various estimates over 1000 replicates for data of 300 two-offspring families.

Scenario	RPA	Unrestricted method	REM
1	0.0074	0.0075	0.0072
	0.0079	0.0079	0.0078
	0.0081	0.0084	0.0083
2	0.0124	0.0124	0.0119
	0.0129	0.0130	0.0124
	0.0131	0.0132	0.0128
3	0.0246	0.0352	0.0272
	0.0250	0.0363	0.0272
	0.0253	0.0359	0.0267
4	0.0131	0.0144	0.0131
	0.0160	0.0162	0.0159
	0.0166	0.0185	0.0166
5	0.0279	0.0395	0.0298
	0.0313	0.0429	0.0311
	0.0387	0.0389	0.0303
6	0.0309	0.0449	0.0319
	0.0358	0.0489	0.0378
	0.0381	0.0432	0.0375

Note, the mean absolute error of estimate is defined as $MAE = \sum_{l=1}^M (|\hat{\theta}_{ABl} - \theta_{AB}| + |\hat{\theta}_{BCl} - \theta_{BC}| + |\hat{\theta}_{ACl} - \theta_{AC}|) / 3M$.

the two methods are comparable, since each has its advantage in some scenarios. However, the RPA proposed in this study has another significant advantage, i.e., the computing speed of which is much faster than that of the REM. For the same simulated data generated above, we also compared the computing time per 1000 replicates for the two restricted methods RPA and REM with the results listed in table 6. We found that the average computing time per 1000 replicates is 121.06 ms for the RPA, but 236.5 ms for the REM. In each replicate, the REM needs 10 steps to converge in average, but the RPA needs only four or five steps to reach convergence. This is expected because the REM needs to calculate many conditional expectations, and then produce the restricted parameter value in each cycle. In contrast, the RPA directly projects the $\hat{\theta}^U$ onto the restricted region in order to obtain the restricted estimate.

In addition, we performed another simulation to evaluate the impact of ignoring the higher ordered terms in the Taylor expansion of $l(\theta)$ in equation (2). Let $\varepsilon(\theta) = o(\|\theta - \hat{\theta}^U\|^2)$ denote the error of the Taylor expansion. We calculated the values of $|\varepsilon(\theta)|$ at $\hat{\theta}^R$ under various scenarios in table 3. At the same time, we consider different cases of sample sizes. From the results we find when the sample size $n = 300$, $|\varepsilon(\hat{\theta}^R)| < 10^{-3}$ in most of the scenarios; when n reaches 1000, $|\varepsilon(\hat{\theta}^R)| < 10^{-6}$ in almost all scenarios; and the error will be much smaller when the sample size increases more. In fact, when the sample size is appropriate, the restricted and unrestricted MLE of θ are all closer to the true value of θ (Royle and Dykstra 1984), so $\|\hat{\theta}^R - \hat{\theta}^U\|$ and the error $|\varepsilon(\hat{\theta}^R)|$ should be smaller. Therefore, ignoring higher

Table 6. Comparison of the computing time per 1000 replicates for the two restricted methods.

Scenario	Parameters			Time (ms) ^a	
	θ_{AB}	θ_{BC}	θ_{AC}	RPA	REM
1	0.05	0.05	0.06	78	156
			0.075	78	156
			0.09	93	171
2	0.05	0.15	0.16	93	171
			0.175	93	171
			0.19	93	187
3	0.05	0.35	0.36	109	281
			0.375	125	265
			0.39	109	250
4	0.15	0.15	0.16	93	203
			0.225	93	218
			0.29	109	203
5	0.15	0.35	0.36	125	281
			0.425	156	296
			0.49	171	265
6	0.35	0.35	0.36	156	296
			0.425	187	343
			0.49	218	344

^aTime, the computing time in milliseconds per 1000 replicates on a Pentium PC with 3.0 GHz and 512 MB RAM.

ordered terms in our method has little impact on the estimates of recombination fractions.

These above results indicate that the natural restriction (1) should be taken into account when estimating recombination fractions in linkage analysis, otherwise it can significantly affect the accuracy of inference. The use of the RPA can yield better performance than the current unrestricted method, and at the same time, the RPA can serve as an alternative or even better estimation method than the REM.

Discussion

Three-locus linkage analysis can potentially offer more advantages over two-locus analysis, and it is also an important case of multi-locus problems (Ridout et al. 1998). Nowadays, solutions can be obtained analytically for the case of usual data type. Ott (1999) first presented a direct estimation method of two-locus recombination fractions in a three-locus analysis. However, as the author mentioned, the parameter space of two-locus recombination fractions, i.e., the inequality restriction (1) in Methods was not considered thoroughly, so that unreasonable estimates may come forth frequently. Based on the same type of data, Zhou et al. (2008) developed a restricted algorithm REM to estimate two-locus recombination fractions, and obtained more precise estimate results in the three-locus analysis. Yet as an alternative or even better method, the restricted projection algorithm RPA proposed in this paper can also deal with this problem. We carry out comprehensive simulations to evaluate the performance of

the RPA, and compare it with the existing unrestricted method and the REM. Results show that the new method has its special advantages.

From the derivation process of the new algorithm, one can find that it is a quadratic algorithm, which determines that the proposed RPA must have a fast computing speed. Although the estimate results of the REM is also completely accepted, the computing speed of it is a little slower than the new one, which may be a common character of those algorithms related to EM algorithm. In addition, the RPA is easy to extend to various usual cases, and can be used a unified method. Moreover, the advantage of the the computing speed of the new algorithm will be more apparent in the course of processing large scale of family or pedigree data.

In practice, the RPA can also be extended to families other than backcross. Taking F_2 family ($Aa/Bb/Cc \times Aa/Bb/Cc$) for example, we still need to obtain the phenotype classification for this family. At this time, families with one offspring can provide corresponding linkage information, and the phenotypes should be grouped into 27 classes. Then the likelihood function $l(\theta)$ is expressed into the function of recombination fractions, and the matrix $\Sigma_1 = -\frac{\partial^2 l(\theta)}{\partial \theta^2} \big|_{\theta=\hat{\theta}^U}$ can be further obtained. Projecting $\hat{\theta}^U$ onto D given in equation (5) according to the detailed steps of the RPA will provide the restricted MLE $\hat{\theta}^R$.

Of course, it is worth emphasizing that the proposed restricted methods need to be investigated further to fit multi-locus linkage analysis. Because when the numbers of loci and alleles on each locus increase, the inference problem of recombination fractions will become complex. For example, if every adjacent three loci are subject to a three-locus analysis in multi-locus case, it will give arise to two different estimates of the same two-locus recombination fraction. How to integrate these estimates (Ridout *et al.* 1998), and how to embed the restricted methods into current popular interval mapping (Lander and Botstein 1989; Kao *et al.* 1999; Chen 2005; Zhou 2010) are worth discussing and studying further. All in all, we should insist that it is the reasonable parameter estimates that can make the further statistical inference more reliable.

Appendix

Calculation of θ_{ni} s

Here, we present the method to obtain θ_{ni} , $n \geq 1$ and $i = 1, 2, 3, 4$. Recalling problem (3), it is easy to make the transformation $\beta = T'\theta$, where T is an orthogonal matrix whose columns are eigenvectors of Σ , and the nonnegative eigenvalues of Σ are λ_i , $i = 1, 2, 3$. It is clear that $T'\Sigma_2 T = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$. Then problem (3) is transformed to:

$$\min \sum_{j=1}^3 \lambda_j (\beta_j^* - \beta_j)^2, \quad (6)$$

$$\text{subject to } \beta \in \bigcap_{i=1}^4 \left\{ \beta \mid \sum_{j=1}^3 c_{ij} \beta_j \leq b_i \right\},$$

where $\beta^* = T'\hat{\theta}^U$, $c_i = a_i T$, and a_i is the i th line of matrix A .

For illustrative purposes, here we give only the projection $\theta_{11} = T\beta_{11}$ of $\hat{\theta}^U = \theta_{10}$ onto D_1 , where β_{11} solves the following equation (7). The projection θ_{12} of θ_{11} onto D_2 can be obtained by the same method, and so on. The projection problem is:

$$\min \sum_{j=1}^3 \lambda_j (\beta_j^* - \beta_j)^2, \quad (7)$$

$$\text{subject to } \beta \in \left\{ \beta \mid \sum_{j=1}^3 c_{1j} \beta_j \leq b_1 \right\},$$

where $\sum_{j=1}^3 \lambda_j (\beta_j^* - \beta_j)^2$ is strictly convex for $\beta = (\beta_1, \beta_2, \beta_3)'$.

The Kuhn–Tucker conditions (see Mokhtar and Shetty 1979; Anthony *et al.* 1992) are usually used to deal with problem (7), and it is easy to verify that there is a unique solution for equation (7).

The Lagrangian is:

$$\sum_{j=1}^3 \lambda_j (\beta_j^* - \beta_j)^2 + \lambda \left(\sum_{j=1}^3 c_{1j} \beta_j - b_1 \right). \quad (8)$$

Differentiating equation (8) to obtain the equation

$$2\lambda_j (\beta_j - \beta_j^*) + \lambda c_{1j} = 0, \quad j = 1, 2, 3,$$

$$\sum_{j=1}^3 c_{1j} \beta_j = b_1.$$

These equations yield:

$$\beta_{11}^{(j)} = \begin{cases} \beta_j^*, & \text{if } \sum_{j=1}^3 c_{1j} \beta_j^* \leq b_1, \\ \beta_j^* - \lambda c_{1j} / 2\lambda_j, & \text{otherwise,} \end{cases}$$

where $\lambda = 2(\sum_{j=1}^3 c_{1j} \beta_j^* - b_1) / (\sum_{j=1}^3 c_{1j}^2 / \lambda_j) > 0$, and $\beta_{11}^{(j)}$ the j th element of vector β_{11} . Then $\theta_{11} = T\beta_{11}$ can be obtained correspondingly. In the analogous manner, sequence $\{\theta_{ni}\}$ all can be obtained.

Acknowledgements

This research was supported by the Mathematical Tianyuan Foundation of China (No. 10926174), the Scientific Research Foundation of Department of Education of Heilongjiang Province of China (nos. 11544032, 11551367), and the Scientific Foundation of Heilongjiang University for Distinguished Young Scholars (no. JCL201003).

References

- Anthony L. P., Francis E. S. and Uhl Jr J. J. 1992 *The mathematics of nonlinear programming*. Springer-Verlag, New York, USA.
- Chen Z. 2005 The full EM algorithm for the MLEs of QTL effects and positions and their estimated variance in multiple-interval mapping. *Biometrics* **61**, 474–480.
- Clemens K. E., Churchill G., Bhatt N., Richardson K. and Noonan F. P. 2000 Genetic control of susceptibility to UV-induced immunosuppression by interacting quantitative trait loci. *Genes Immun.* **1**, 251–259.
- Dykstra R. L. 1983 An algorithm for restricted least squares regression. *J. Am. Statist. Assoc.* **78**, 837–842.
- Elston R. C. and Stewart J. 1971 A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21**, 523–542.
- Fisher R. A. 1935 The detection of linkage with 'dominant' abnormalities. *Ann. Eugen.* **6**, 187–201.
- Haldane J. B. S. 1919 The recombination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet.* **8**, 299–309.
- Kao C. H., Zeng Z. B. and Teasdale R. D. 1999 Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203–1216.
- Kudo A. 1963 A multivariate analogue of the one-side test. *Biometrika* **50**, 403–418.
- Lander E. S. and Green P. 1987 Construction of multilocus linkage maps in human. *Proc. Natl. Acad. Sci. USA* **84**, 2363–2367.
- Lander E. S. and Botstein D. 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.
- Lathrop G. M. 1985 Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *Am. J. Hum. Genet.* **37**, 482–498.
- Lathrop G. M., Lalouel J. M., Julier C. and Ott J. 1984 Strategies for multilocus linkage analysis in humans. *Proc. Natl. Acad. Sci. USA* **81**, 3443–3446.
- Li Z. H., Qin H. and Zhang H. 2006 *Statistical methods in genetics*. Science Press, Beijing, P. R. China.
- Meyers D. A., Conneally P. M., Lovrien E. W., Magenis R. F., Merritt A. D., Norton J. A. et al. 1976 *Birth defects: original article series*, pp. 335–339. The National Foundation, New York.
- Mokhtar S. B. and Shetty C. M. 1979 *Nonlinear programming: theory and algorithms*. John Wiley, New York, USA.
- Morton N. 1955 Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7**, 277–318.
- Ott J. 1976 A computer program for linkage analysis of general human pedigrees. *Am. J. Hum. Genet.* **28**, 528–529.
- Ott J. 1999 Phase-unknown triple backcross with two offspring. In *Analysis of human genetic linkage*, 3rd edition, pp. 122–124. The Johns Hopkins University Press, Baltimore, USA.
- Ridout M. S., Tong S., Vowden C. J. and Tobutt K. R. 1998 Three-point linkage analysis in crosses of allgamous plant species. *Genet. Res.* **72**, 111–121.
- Royle J. P. and Dykstra R. L. 1984 A method for finding projection onto the intersection of convex sets in hilbert space. In *Advance of ordered restricted statistical inference*, (ed. R. L. Dykstra, T. Robertson and F. T. Wright). Springer-Verlag, New York, USA.
- Shi N. Z., Zheng S. R. and Guo J. H. 2005 The restricted EM algorithm under inequality restrictions on the parameters. *J. Multivariate Anal.* **92**, 53–76.
- Thompson E. A., 1984. Information gain in joint linkage analysis. *IMA J. Math. Appl. Med. Biol.* **1**, 31–49.
- Von Neumann J. 1950 *Functional operators*, volume 2. Princeton University Press, Princeton, USA.
- Wiener N. 1955 On factorization of matrices. *Comm. Math. Helv.* **29**, 97–111.
- Wu R. L., Ma C. X. and Casella G. 2007 *Statistical genetics of quantitative traits: linkage, maps, and QTL*. Springer, New York, USA.
- Zhou Y. 2010 Multiple interval mapping for quantitative trait loci via EM algorithm in the presence of crossover interference. *Commun. Stat. Part A* **39**, 3041–3057.
- Zhou Y., Shi N. Z., Fung W. K. and Guo J. H. 2008 Maximum likelihood estimates of two-locus recombination fractions under some natural inequality restrictions. *BMC Genet.* **9**, 1.

Received 5 April 2010, in revised form 12 October 2010; accepted 8 March 2011

Published on the Web: 19 August 2011