ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.9, September 2007, pp.2090−2099
DOI: 10.1360/jos182090

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

三元家庭基因数据的单体分型和单体型频率估计*

张强锋[1,2+], 徐 云[1,2], 陈国良[1], 车皓阳[2]

[1](中国科学技术大学 计算机系,安徽 合肥 230027)

[2](中国科学院 软件研究所 计算机科学国家重点实验室,北京 100080)

Haplotyping and Haplotype Frequency Estimates on Trio Genotype Data

ZHANG Qiang-Feng[1,2+], XU Yun[1,2], CHEN Guo-Liang[1], CHE Hao-Yang[2]

[1](Department of Computer Science, University of Science and Technology of China, Hefei 230027, China)

[2](State Key Laboratory of Computer Science, Institute of Software, The Chinese Academy of Sciences, Beijing 100080, China)

+ Corresponding author: Phn: +86-551-3602441, E-mail: qfzhang@mail.ustc.edu.cn

Zhang QF, Xu Y, Chen GL, Che HY. Haplotyping and haplotype frequency estimates on trio genotype data. *Journal of Software*, 2007,18(9):2090−2099. http://www.jos.org.cn/1000-9825/18/2090.htm

Abstract: The problems of haplotyping and haplotype frequency estimation on trio genotype data under the Mendelian law of inheritance and the assumption of Hardy-Weinberg equilibrium are studied in this paper. Since most past efforts only focused on haplotyping on genotype data of unrelated individuals and data with general pedigrees, but gave insufficient efforts to the special case of trio genotype data, there is coming an increasing demand in analyzing them in particular, especially when taking into account that part of HAPMAP database is exactly trio data. This paper presents a two-staged method to estimate haplotype frequencies in trios: i) haplotyping stage, find haplotype configurations without recombinant for each trio; ii) frequency estimation stage, use the expectation-maximization (EM) algorithm to estimate haplotype frequencies based on these inferred haplotype configurations. Both the haplotyping algorithm and the EM algorithm are implemented in software package TRIOHAP using C language. Its effectiveness and efficiency and tested on simulated and real data sets as well. The experimental results show that, TRIOHAP runs much faster than a popular frequency estimation software which discards trio information. Moreover, because TRIOHAP utilizes such information, its estimation is more reliable.

Key words: genotype; haplotype; SNP; haplotyping; haplotype frequencies estimate; trio; EM algorithm

摘 要: 研究了在门德尔遗传定理和哈代-维恩伯格平衡假设下,三元家庭基因型数据的单体分型和单体型频率估计问题.过去的研究仅仅关注个体间没有联系或者含有一般家系信息的基因型数据,而对这种特殊的三元家庭关注得不够.考虑到 HAPMAP 数据库中有一部分数据就基于这种三元家庭,现在有越来越多的需求要求直接分析这种特殊的家系结构.提出一个两段式的三元家庭中单体型频率的估计方法:i) 分型阶段,找出每一个三元家庭零重组单体构型;ii) 频率估计阶段,在前一阶段得到的单体构型基础上,应用 EM 算法来估计单体型频率.在程序包 TRIOHAP 中用 C 语言实现了单体分型算法和 EM 算法,并且使用模拟和实际数据测试了 TRIOHAP 的有效性和效率.实验结果表明,TRIOHAP 要比其他那些忽略了三元家庭信息的常见单体型频率估计软件运行快很多.进一步地,

* Supported by the National Natural Science Foundation of China under Grant No.60533020 (国家自然科学基金)
Received 2004-12-21; Accepted 2006-03-31

由于 TRIOHAP 利用了这些信息,其估计结果更加可靠.

关键词: 基因型;单体型;SNP;单体分型;单体型频率估计;三元家庭;EM 算法

中图法分类号: TP301 **文献标识码:** A

Haplotype analysis has gained increasing attention in the context of association studies of disease gene and drug responsivity over the last years. In diploid organisms, such as humans, chromosomes come in pairs, and experiments often yield *genotype* information, which blend haplotype information for chromosome pairs. There is growing evidence that, in order to better characterize the role of a candidate gene, full haplotype information should be exploited instead of unclear genotype information. Unfortunately, it is both time-consuming and expensive to derive haplotype information experimentally. This explains the increasing interest in inferring haplotype information, or haplotyping, computationally. Actually, the potential use of haplotypes has led to the initiation of the HapMap project (http://www.hapmap.org/) to investigate haplotype patterns in the human genome in different populations. Haplotyping and haplotype frequency estimates are essential components of this endeavor.

The input genotype data can be with or without any *pedigree* information. Genotype data without pedigree information are also called population genotype data where only unrelated individuals are available for analysis[1−5]. The information obtained by haplotyping pedigree genotype data is believed to be more reliable than haplotyping population genotype data: the constraint provided by other members in a pedigree would force one genotype to settle on a unique haplotype configuration as being most probable.

There are also various algorithms and programs developed for haplotyping pedigree data[6−10], which have to deal with the huge number of consistent haplotype configurations derived from even a data set of moderate-size. There are many pedigrees which are simple families with only a pair of parents and one child. Such simple pedigrees are called *trios*. For example, the data samples for the HapMap include data from 30 trio families of Yoruba and 30 trio families of U.S. However, few emphases have been given to these special cases.

We study the problem of haplotyping and haplotype frequency estimation in such trios in this paper. Instead of estimating frequencies directly, we first determine the plausible haplotype configurations. In general, haplotyping pedigrees need to consider the entire solution space of all consistent haplotype configurations. However, genomic DNA can be partitioned into blocks such that recombinants within each block are rare or even nonexistent[12]. Thus it is believed that haplotype configurations with fewer recombinations should be preferred. When the region of interest is so small that the expected number of recombinations in the pedigree data is very close to zero, the solution space of all consistent haplotype configurations can be replaced by that of zero recombinant (provided it is non-empty) when estimating haplotype frequencies. Because the contribution of the solutions with recombinants to the overall likelihood becomes small compared with those containing no recombinant, they bring considerable complexity to the computation. So when we focus on a small region of haplotypes consisting of tightly linked markers, we usually assume that there is no recombination. Thus the number of haplotypes whose frequencies need to be estimated in the latter step is greatly reduced.

In the second stage, we generalize the EM algorithm that was originally designed to estimate haplotype frequencies based on unrelated genotype individuals so that it can handle genotypes with trios. We use the Hardy-Weinberg equilibrium to compute the probabilities of parent genotypes in each trio and use a general genetic model[13] to compute those of the child ones.

Both the haplotyping algorithm and the EM algorithm are implemented using C language in a software package named TRIOHAP. We test its effectiveness and efficiency on simulated and real data sets as well. The experimental results show that, TRIOHAP runs much faster than the other popular haplotype frequency estimation software

which discards trio information. Moreover, because TRIOHAP utilizes such information, our estimation is more reliable.

The paper is organized as follows. We describe the background in Section 1, the haplotyping algorithm and the EM algorithm in Section 2. Section 3 presents the experimental results, followed by concluding remarks in Section 4.

1 Preliminary Definitions

Haplotypes and *genotypes* consist of linked genetic markers which are small segments of DNA with some specific features. The physical position of a marker on a chromosome is called a *locus* and its state is called an *allele*, which can be denoted a letter in alphabet Σ. A haplotype h with m loci can be presented as a string of length m over Σ^m. A genotype g is a string of length m where each component is an un-ordered duet over Σ. A locus of g is said to be *homozygous* when the two alleles of that locus are the same, else it is *heterozygous*. Haplotype pair $\langle h_1,h_2\rangle$ is *consistent* with a genotype g if and only if the two alleles of h_1 and h_2 equals those of g for each locus. For example, h_1='aacg', h_2='atca' are two haplotypes of length 4. g='{a,a}{a,t}{c,c}{g,a}' is a genotype, and $\langle h_1,h_2\rangle$ is *consistent* with g. Locus 1 and locus 3 of g are *homozygous* sites while Locus 2 and locus 4 are *heterozygous*.

A *pedigree* is a fundamental structure used in genetics which describes the relationship within a family. A *trio* is a simple pedigree consists of only two parents and one child. Figure 1 shows a pictorial illustration of a trio.

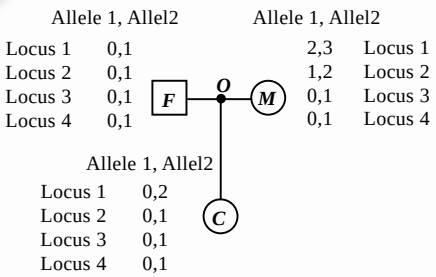


Fig.1 A pictorial illustration of a trio. Each node in the trio is associated with a genotype

Usually, each individual node in a trio is associated with a genotype. In the absence of genetic mutation, the child must inherit one allele from its father and the other from its mother at each locus. This is known as the *Mendelian law of inheritance*. A *haplotype configuration* for a trio T with genotype information then refers to assigning a pair of consistent haplotypes for the genotype of each node. In the case of no missing alleles, it is trivial to check the Mendelian consistency, so from now on, we assume that the given genotype information of pedigree T is always Mendelian consistent.

Genetic research shows that recombinations are rare in human genome. And in the case of no mutation and no recombination, each haplotype of a child is inherited as a whole from one of the two haplotypes of a parent.

2 Haplotyping and Haplotype Frequency Estimates

Excoffier, *et al.*[3] first used the EM algorithm to estimate haplotype frequencies based on unrelated individual genotype data. His method was then widely applied in genetic studies[13,14]. However, it should calculate the frequencies of all possible haplotypes consistent to the given genotypes, which is unbearable when their lengths grow to more than 20 O'Connell[15] showed that genetic information from relatives in a general pedigree could be used to resolve haplotype ambiguity. However, O'Connell's method focused on the general pedigree and had an exponential time complexity. Here we present a two-staged method to estimate haplotype frequencies in trios.

2.1 Haplotyping stage

Let us consider a single trio T and the genotype information at a single locus i. Haplotyping locus i means determining the paternal allele and maternal allele for the child node. Suppose that the alleles of the father F, the mother M, and the child C are $\{a,b\}$, $\{c,d\}$ and $\{e,f\}$ ($\{e,f\}\subset(\{a,b\}\cup\{c,d\})$) at locus i. If locus i of node C is homozygous ($e=f$), then it is clear that its paternal allele and the maternal allele are the same (e or f). The situation becomes complicated if it is heterozygous ($e\neq f$). Given the genotype information at locus i for all of the three members, it may or may not be possible to determine the paternal allele and the maternal allele for the child. Table 1 lists out all cases, and locus i is called *unambiguous* in the trio if we can; otherwise it will be called *ambiguous*.

Table 1 Determining the paternal allele and the maternal allele for the child at a single locus

Conditions	Paternal	Maternal	Conditions	Paternal	Maternal
$e=f$	e	f	$e\neq f$, $a\neq b$, $c\neq d$		
$e\neq f$, $a=b$			$a\neq c$, $b=d$		
$e=a$	e	f	$e=a$ or $f=c$	e	f
$f=a$	f	e	$e=c$ or $f=a$	f	e
$e\neq f$, $a\neq b$, $c=d$			$a=d$, $b=c$	Ambiguous	
$e=c$	f	e	$a=d$, $b\neq c$		
$f=c$	e	f	$e=b$ or $f=c$	e	f
$e\neq f$, $a\neq b$, $c\neq d$			$e=c$ or $f=b$	f	e
$a=c$, $b=d$	Ambiguous		$a\neq d$, $b=c$		
$a=c$, $b\neq d$			$e=a$ or $f=d$	e	f
$e=b$ or $f=d$	e	f	$e=d$ or $f=a$	f	e
$e=d$ or $f=b$	f	e	$a\neq c\neq b\neq d$		
			$e=a$ or $e=b$	e	f
			$e=c$ or $e=d$	f	e

Theorem 1. (1) The haplotype configuration for a trio with no ambiguous loci is unique; (2) The number of haplotype configurations for a trio with h ambiguous loci is 2^h.

Proof: (1) Obviously, we can determine the paternal allele and the maternal allele of the child for each unambiguous locus. Then for a trio without ambiguous loci, we can determine those for all loci. For the child, the alleles from the father form a haplotype (paternal haplotype) and the alleles from the mother form another (maternal haplotype). In the absence of recombinations, each haplotype of the child must be inherited as a whole from one of the two haplotypes of a parent. The child's haplotypes then can be used to resolve its parents' genotypes uniquely. So the haplotype configuration for a trio without ambiguous loci is unique.

(2) Firstly, ignoring all ambiguous loci, from (1), the (partial) haplotype configuration for the remaining loci is unique. For each ambiguous locus, we can select one node and arbitrarily set the two alleles on the two partial haplotypes to be 'e' and 'f'. Once the two alleles of one node have been set, the alleles of the other two parental nodes can be set consequently. There are two ways to set the two alleles of an ambiguous locus, i.e. set the allele on the haplotype to 'e' or to 'f' and the allele on the other haplotype to 'f' or to 'e'. Because each of the h ambiguous loci can be set independently, there are 2^h possible haplotype configurations for a trio with h ambiguous loci. □

Note that the proof also gives a way to find a haplotype configuration in linear time.

Figure 2 shows an example. There is only one ambiguous locus (locus 3) in trio (F,M,C). Fig.2(a) shows its genotype configuration, Fig.2(b) is the unique haplotype configuration for the unambiguous loci, Fig.2(c) and Fig.2(d) are two haplotype configurations for the genotype configuration in Fig.2(a). In Fig.2(c), the paternal haplotype is '0101' and the maternal haplotype is '1110'; in Fig.2(d) the paternal haplotype is '0111' and the maternal haplotype is '1100'.

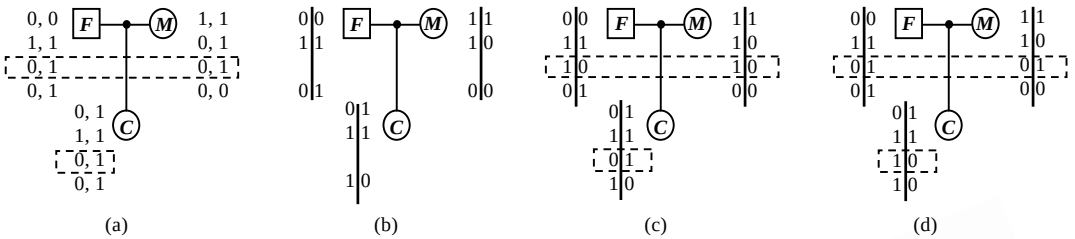


Fig.2 An example of two consistent haplotype configurations for a trio with genotype information

2.2 Haplotype frequency estimation stage

Suppose that we are given a set of trios $\boldsymbol{T}=\{T_1,T_2,\ldots,T_K\}$. Each T_i consists of a father F_i, a mother M_i and a child C_i. Suppose that there are π_i feasible haplotype configurations for trio T_i found by the haplotyping algorithm, and the j-th one is: $S_{j,F_i}=\langle \alpha_{j,F_i}, \beta_{j,F_i}\rangle$, $S_{j,M_i}=\langle \alpha_{j,M_i}, \beta_{j,M_i}\rangle$, and $S_{j,C_i}=\langle \alpha_{j,C_i}, \beta_{j,C_i}\rangle$ ($1\le j\le r_i$). All haplotypes appeared in these configurations form a list $H=\{h_1,h_2,\ldots,h_l\}$ with frequencies $\Theta=\{\theta_1,\theta_2,\ldots,\theta_l\}$.

The likelihood of haplotype frequencies given the observed trio data is,

$$\begin{aligned} L\{\theta_1,\theta_2,...,\theta_l\} &= \Pr(T_1,T_2,...,T_K\mid\Theta)\\ &= c\times\prod_{i=1}^{K}\Pr(T_i\mid\Theta)\\ &= c\times\prod_{i=1}^{K}\Pr((F_i,M_i,C_i)\mid\Theta)\\ &= c\times\prod_{i=1}^{K}\left(\sum_{j=1}^{\pi_i}\Pr(\langle S_{j,F_i},S_{j,M_i},S_{j,C_i}\rangle\mid\Theta)\right) \end{aligned} \tag{1}$$

Under the assumption of random mating, the parents' haplotype configurations are independent, and the child's haplotype configuration is transmitted from its parents,

$$\Pr(\langle S_{j,F_i},S_{j,M_i},S_{j,C_i}\rangle\mid\Theta)=\Pr(S_{j,F_i}\mid\Theta)\cdot\Pr(S_{j,M_i}\mid\Theta)\cdot\Pr(S_{j,C_i}\mid\langle S_{j,F_i},S_{j,M_i}\rangle) \tag{2}$$

Where $\Pr(S_{j,F_i}\mid\Theta)$ (and $\Pr(S_{j,M_i}\mid\Theta)$) can be computed using the Hardy-Weinberg equilibrium: $\Pr(S_{j,F_i}\mid\Theta)=\theta^2_{\alpha_{j,F_i}}$ if $\alpha_{j,F_i}=\beta_{j,F_i}$ and $\Pr(S_{j,F_i}\mid\Theta)=2\theta_{\alpha_{j,F_i}}\theta_{\beta_{j,F_i}}$ if $\alpha_{j,F_i}\ne\beta_{j,F_i}$. $\Pr(S_{j,C_i}\mid\langle S_{j,F_i},S_{j,M_i}\rangle$ is the gamete transmission probabilities of haplotype configuration S_{j,C_i} with the parental haplotype configurations of S_{j,F_i} and S_{j,M_i}. It can be computed by a genetic model presented by Elston and Stewart[11] in 1971.

The EM algorithm is an iterative method to compute successive sets of frequencies $\Theta=\{\theta_1,\theta_2,\ldots,\theta_l\}$, starting with initial arbitrary values $\Theta^{(0)}=\{\theta_1^{(0)},\theta_2^{(0)},...,\theta_l^{(0)}\}$. These initial values are employed as if they were the unknown true frequencies to estimate genotype frequencies $\Pr(S_{j,F_i}\mid\Theta)$, $\Pr(S_{j,M_i}\mid\Theta)$, and $\Pr(S_{j,C_i}\mid\langle S_{j,F_i},S_{j,M_i}\rangle)$ (the expectation step). These expected genotype frequencies are used in turn to estimate haplotype frequencies at the next iteration $\Theta^{(1)}=\{\theta_1^{(1)},\theta_2^{(1)},...,\theta_l^{(1)}\}$ (the maximization step), and so on, until convergence is reached.

Suppose that in the s-th iteration, $\Theta=\Theta^{(s)}$ and we want to estimate $\Theta^{(s+1)}$. We calculate $\Pr(H_{j,F_i},H_{j,M_i},H_{j,C_i}\mid\Theta)$ in the expectation step. In the maximization step, we firstly normalize $\Pr(\langle S_{j,F_i},S_{j,M_i},S_{j,C_i}\rangle\mid\Theta)$, then the haplotype frequencies can be computed using a gene-counting method. Let $\delta_{i,j,t}$ be an indicator variable equaling to the number of haplotype h_t that appears in the solution $\langle S_{j,F_i},S_{j,M_i},S_{j,C_i}\rangle$, we have,

$$\tilde{\Pr}(\langle S_{j,F_i},S_{j,M_i},S_{j,C_i}\rangle\mid\Theta)^{(s)}=\frac{1}{K}\cdot\frac{\Pr(\langle S_{j,F_i},S_{j,M_i},S_{j,C_i}\rangle\mid\Theta)^{(s)}}{\sum_{j'=1}^{\pi_i}\Pr(\langle S_{j',F_i},S_{j',M_i},S_{j',C_i}\rangle\mid\Theta)^{(s)}} \tag{3}$$

$$\theta_t^{(s+1)}=\frac{1}{2}\cdot\sum_{i=1}^{K}\sum_{j=1}^{\pi_i}\delta_{i,j,t}\cdot\tilde{\Pr}(\langle S_{j,F_i},S_{j,M_i},S_{j,C_i}\rangle\,|\,\Theta)^{(s)}) \tag{4}$$

The flow chart depicting the implementation of the EM algorithm is shown in Fig.3.

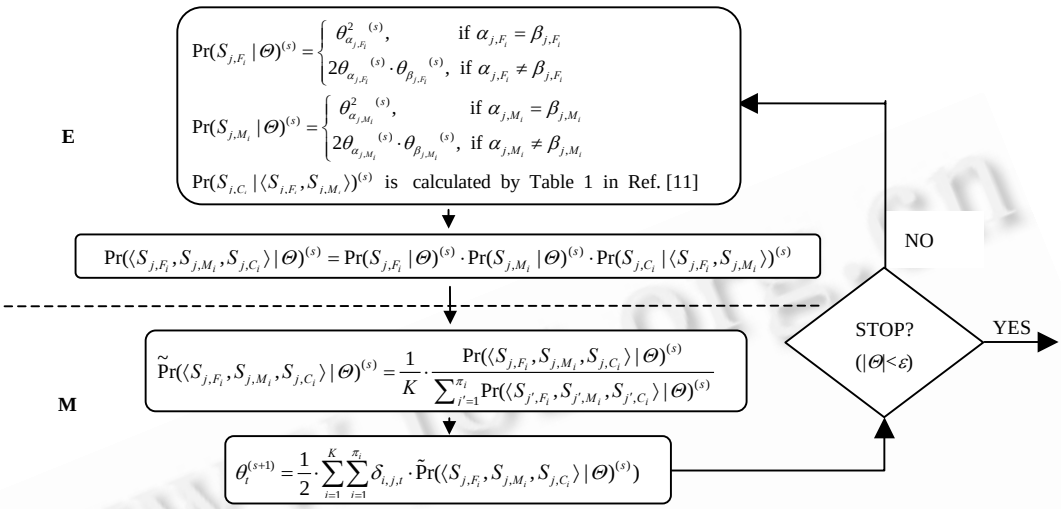


Fig.3 A flow chart depicting the implementation of the EM algorithm

There are several ways to initialize the haplotype frequencies $\Theta=\{\theta_1,\theta_2,\ldots,\theta_l\}$. For instance, the initial haplotype frequencies can be chosen at random, or all haplotypes are equally frequent, *i.e.* $\theta_t^{(0)}=1/l$ (t=1,2,…,l). Or that all initial haplotype frequencies are equal to the product of the corresponding single-locus allele frequencies (*i.e.*, a complete linkage equilibrium). Also, we can set all feasible solutions for each trio to be equally likely, *i.e.* $\Pr(\langle S_{j,F_i},S_{j,M_i},S_{j,C_i}\rangle\,|\,\Theta)^{(0)}=1/\pi_i$, ($j$=1,2,…,$\pi_i$). We can even initialize the haplotype frequencies by counting their occurrence in all the feasible solutions. Since in practical applications the EM algorithm could be trapped in some local maximum, we recommend to restart the algorithm several times with different initial haplotype frequencies and it is better with a randomized additive perturbation.

The stopping (convergence) criterion is defined as the absolute value of the difference of Θ between the consecutive iterations being less than some small value $\varepsilon>0$.

2.3 Discussion

Excoffier, *et al*.[3] used the EM algorithm to estimate haplotype frequencies while ignoring the relatives' information. Here we adopt a two-staged method, which tries to reduce the number of haplotypes to be considered in the estimation stage by utilizing the relatives' information to do haplotyping at first.

Suppose that each haplotype consists of m biallelic SNP loci. For locus i, suppose that i happens to be one state with a probability of p_i, and to be the other state with a probability of $(1-p_i)$. Then locus i of the genotype is heterozygous with the probability of $2p_i\,(1-p_i)$. Suppose that the expected value of p_i is p, then the genotype is expected to have $2p(1-p)\cdot m$ heterozygous loci. As a consequence, a number of $2^{2p(1-p)\cdot m-1}$ possible haplotype pairs are expected to be considered if we use the EM algorithm directly. However, the probability that locus i in a trio is ambiguous is $2p_i(1-p_i)\cdot 2p_i(1-p_i)\cdot 2/4=2\,p_i^2\,(1-p_i)^2$. So the expected number of possible haplotype configurations for the trio is $2^{2p^2(1-p)^2\cdot m}$. If p=1/2, our method can handle $\lambda=(2p(1-p))/(2p^2(1-p)^2)=4$ times longer genotypes than Excoffier's methods. Moreover, in practice, the more frequent allele often appears with a probability of more than

0.9, so our method usually can handle $\lambda=1/p(1-p)>10$ times longer genotypes.

If each locus of the haplotype is a micro-satellite locus of l different alleles: $a_1,a_2,\ldots,a_l$, where each appears with the probability of $p_1,p_2,\ldots,p_l$. Then the expected number of consistent haplotype pairs for a genotype is $2^{\sum_{i\neq j} p_i p_j \cdot m-1}$, and the expected number of haplotype configurations for a trio is $2^{\sum_{i\neq j} p_i^2 p_j^2 \cdot m}$, so our method usually can handle $\lambda=\sum_{i\neq j} p_i p_j / \sum_{i\neq j} p_i^2 p_j^2$ times longer genotypes. For example, when l=8, and $p_1=p_2=\ldots=p_8=1/8$, λ=64, *i.e*. our method can be applied to cases of much larger scale.

3 Experimental Results

Our algorithms have been implemented in a C software package named TRIOHAP. To evaluate its performance, we test it on both simulated and real data in terms of efficiency and accuracy. All experiments are conducted on a Linux server with 1.7GHz CPU and 256MB RAM.

3.1 Number of solutions

In order to demonstrate the effectiveness of the haplotyping algorithm, we use the simulation data to examine the number of haplotypes in set H.

We first generate a population of haplotypes H^*. All locus of each haplotype are set to certain alleles according to the probability distribution function P. In our simulation, we generate haplotypes of SNP loci and micro-satellite loci as well. For SNP loci, we set P_1 ($p_1=p_2=0.5$) and P_2 ($p_1=0.9$, $p_2=0.1$). For micro-satellite loci, we set l=4, and P_3 ($p_1=p_2=p_3=p_4=0.25$) and P_4 ($p_1=0.5$, $p_2=p_3=0.2$, $p_4=0.1$). We let $|H^*|$=20, and $\theta_1^*=0.2$, $\theta_2^*=\theta_3^*=\theta_4^*=0.1$, $\theta_5^*=\theta_6^*=\theta_7^*=\theta_8^*=0.05$, $\theta_9^*=\theta_{10}^*=\ldots=\theta_{20}^*=0.025$. In each trio, we independently select a pair of haplotypes for each parent node according to θ_i^*. The two haplotypes of the child are transmitted from those of its parents according to Elston's model. At last, a pair of haplotypes of the same node is blended to form a genotype. For each parameter setting (m,K,P) (here m is the length of haplotypes and K is the number of trios), 100 copies are generated.

The average numbers of haplotypes that should be estimated are recorded in Table 2.

Table 2 Comparison of number of haplotypes ($|H|$) on simulation data

Parameter settings (m,K)	Directly				TRIOHAP			
	P_1	P_2	P_3	P_4	P_1	P_2	P_3	P_4
(20,20)	3.18e4	3.43e2	2.06e6	1.02e6	5.83e2	72.2	1.12e2	76.4
(20,100)	1.49e5	1.68e3	1.02e7	5.02e6	2.67e3	3.55e2	5.72e2	3.61e2
(20,200)	2.56e5	3.30e3	2.02e7	0.98e7	4.45e3	5.96e2	1.14e3	6.02e2
(100,20)	N/A	8.13e6	N/A	N/A	5.90e5	2.60e2	6.91e2	2.74e2
(100,100)	N/A	1.62e7	N/A	N/A	2.68e6	1.31e3	3.42e3	1.52e3
(100,200)	N/A	3.21e7	N/A	N/A	4.65e6	2.55e3	6.91e3	2.89e3
(200,20)	N/A	N/A	N/A	N/A	N/A	7.92e2	6.08e3	1.07e3
(200,100)	N/A	N/A	N/A	N/A	N/A	4.09e3	3.21e4	5.17e3
(200,200)	N/A	N/A	N/A	N/A	N/A	7.79e3	6.22e4	1.03e4

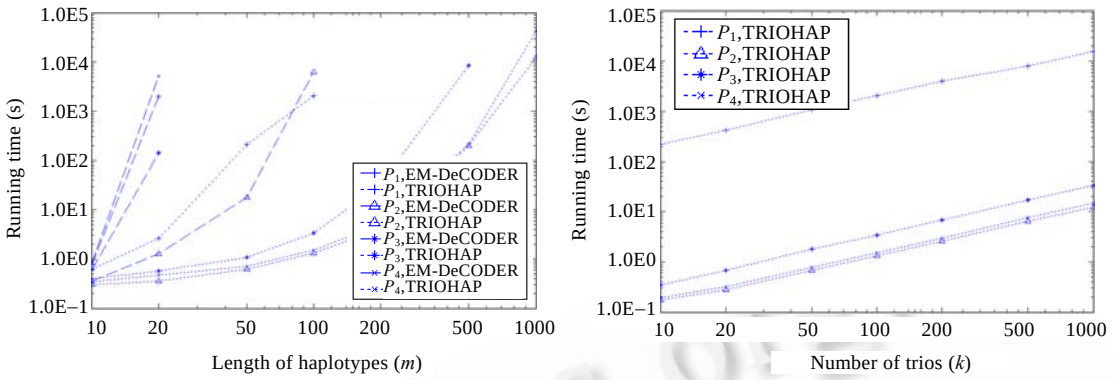
We can see from the table that the numbers of haplotypes are greatly reduced after applying the haplotyping algorithm (TRIOHAP *vs*. directly), which will immediately bring improvement on the running time.

3.2 Running time

We compare the running time of TRIOHAP with EM-DeCODER, a popular software using the EM algorithm to estimate the haplotype frequencies. Fig.4 shows the running time of TRIOHAP and EM-DeCODER over different parameter setting (m,K,P).

We can learn from the figure that TRIOHAP runs much quicker than EM-DeCODER, and thus can be applied to larger instances. We also notice that the running times of both TRIOHAP and EM-DeCODER increase

exponentially with the length of haplotypes while increasing near-linearly with the number of trios (the running time of EM-DeCODER is not plotted in Fig.4(b) because haplotypes of length 100 are out of its capability).



(a) Running time vs. length of haplotypes (K=100) (b) Running time vs. number of trios (m=100)

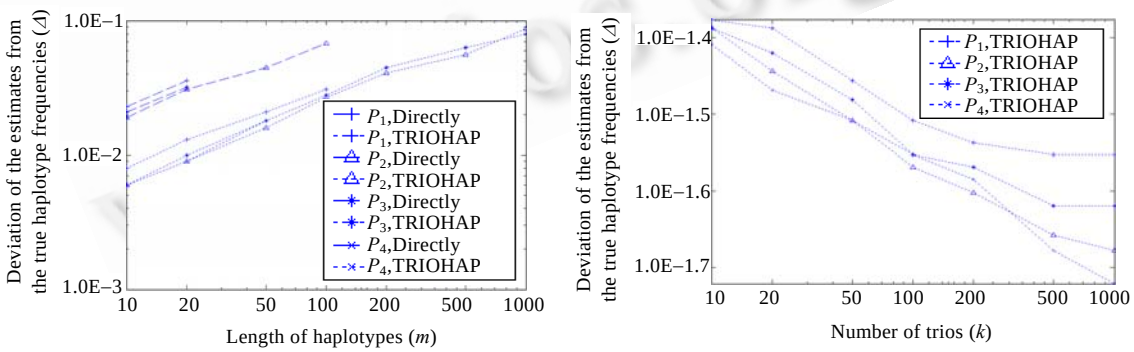
Fig.4 The running time of TRIOHAP and EM-DeCODER

3.3 Accuracy rate

We define a parameter Δ to incarnate the deviation of the estimated haplotype frequencies from the underlying true ones. Because the simulation data is generated according to the Θ^*, we recognize that as the true frequencies. Suppose the estimated haplotype set is H with frequencies Θ. Suppose the estimated frequencies of the 20 haplotypes in H^* are $\theta_1,\theta_2,\ldots,\theta_{20}$. Let

$$\Delta=\sqrt{\frac{\sum_{i=1}^{20}(\theta_i-\theta_i^*)^2}{20}} \tag{5}$$

Fig.5 shows the deviation of the estimate of TRIOHAP and EM-DeCODER over different parameter setting (m,K,P). We can learn from the figure that the deviation of TRIOHAP is smaller than that of EM-DeCODER. We also notice that the deviation of the estimate increases with the length of haplotypes while decreasing with the number of trios.



(a) Estimate deviation vs. length of haplotypes (K=100) (b) Estimate deviation vs. number of trios (m=100)

Fig.5 The deviation of the estimate of TRIOHAP and EM-DeCODER

3.4 A real data set

We also test the efficiency and the accuracy of TRIOHAP on a real data set, which is a set of 122 trios (366

genotypes) coming from dbMHC|ABDR. Each genotype contains 31 markers of the same positions on chromosome 6, among which 10 are micro-satellite markers and the other 21 are SNPs.

We run TRIOHAP to find the underlying haplotypes. It only takes TRIOHAP 0.97 second to find the 20 most frequent haplotypes (with frequencies larger than 0.01) while it is out of the capability of EM-DeCODER.

4 Concluding Remarks

We present a two-staged method to do haplotyping and to estimate haplotype frequencies for trio genotype data. Given a set of trios, it firstly determines all haplotype configurations for each trio, and then uses an expectation-maximization (EM) algorithm to estimate the frequencies of haplotypes that appear in these configurations. Because a large number of haplotypes have been eliminated from the possible haplotype list, our method greatly speeds up the estimation stage.

We implement both algorithms using C language in software package TRIOHAP and test its effectiveness and efficiency on simulated and real data sets as well. The experimental results show that, TRIOHAP runs much faster than EM-DeCODER, and thus can be applied to much larger scale of instances. Moreover, the deviation of the estimate of TRIOHAP is smaller than that of EM-DeCODER, which means it is more accurate.

However, there remain problems which should be taken into account in the future. Firstly, we assume that there are no recombinations between generations, but this is not always true. We should make a tradeoff between the probability of bringing in recombinations and the probability of introducing infrequent haplotypes in estimation. In addition, sometimes in practice a significant part of alleles in the genotype data are lost. It is interesting to deduce these missing alleles, whether determinately or probabilistically.

References:

[1] Clark AG. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol., 1990,7(2):111−122.

[2] Zhang QF, Che HY, Chen GL, Sun G. A practical algorithm for haplotyping by maximum parsimony. Journal of Software, 2005,16(10):1699−1707 (in English with Chinese abstract). http://www.jos.org.cn/1000-9825/16/1699.htm

[3] Excoffier L, Slatkin M. Maximum-Likelihood estimation of molecular haplotype frequencies in a diploid population. Molecular Biology and Evolution, 1995,12(5):921−927.

[4] Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction for population data. American Journal of Human Genetics, 2001,68(4):978−989.

[5] Niu T, Qin ZS, Xu X, Liu JS. Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. American Journal of Human Genetics, 2002,70(1):157−169.

[6] Lin S, Speed TP. An algorithm for haplotype analysis. Journal Computational Biology, 1997,4(4):35−46.

[7] Tapadar T, Ghosh S, Majumder PP. Haplotyping in pedigrees via a genetic algorithm. Human Heredity, 2000,50(1):43−56.

[8] Qian D, Beckmann L. Minimum-Recombinant haplotyping in pedigrees. American Journal of Human Genetics, 2002,70(6): 1434−1445.

[9] Li J, Jiang T. Efficient rule-based haplotyping algorithms for pedigree data. In: Proc. of the RECOMB 2003. 2003. 197−206.

[10] Li J, Jiang T. An exact solution for finding minimum recombinant haplotype configurations on pedigrees with missing data by integer linear programming. In: Proc. of the RECOMB 2004. 2004. 20−29.

[11] Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. Human Heredity 1971,21(6):523−542.

[12] Griffiths A, Gelbart W, Lewontin R, Miller J. Modern Genetic Analysis: Integrating Genes and Genomes. New York: W.H. Freeman and Company, 2002.

[13] Fallin D, Schork NJ. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. American Journal of Human Genetics, 2000,67(4):947−959.

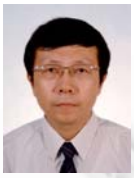
[14] Zhao HY, Zhang SL, Merikangas KR, Trixler M, Wildenauer DB, Sun FZ, Kidd KK. Transmission/Disequilibrium tests using multiple tightly linked markers. American Journal of Human Genetics, 2000,67(4):936−946.

[15] O'Connell JR. Zero-Recombinant haplotyping: applications to fine mapping using SNPs. Genetic Epidemiology, 2000,19(Suppl. 1): s64−s70.

附中文参考文献:

[2] 张强锋,车皓阳,陈国良,孙广中.最大节约原则下单倍型推导问题的实用算法.软件学报,2005,16(10):1699−1707. http://www.jos.org.cn/ 1000-9825/16/1699.htm

ZHANG Qiang-Feng was born in 1979. He is a Ph.D. candidate at the University of Science and Technology of China. His current research areas are bioinformatics and combinatorial optimization.



XU Yun was born in 1960. He received the Ph.D. degree from the University of Science and Technology of China in 2002. Now he is an associate professor at USTC. His current research areas are parallel computing and combinatorial optimization.

CHEN Guo-Liang was born in 1938. He is a fellow of Chinese Academy of Science and a professor in the Department of CS at USTC and a CCF senior member. His current research ares are parallel computing, computer architecture and combinatorial optimization.

CHE Hao-Yang was born in 1977. He is a Ph.D. candidate at the Institute of Software, the Chinese Academy of Sciences. His current research areas are optimization algorithms, recommendation system, trust management, etc.