



HAPLORE: a program for haplotype reconstruction in general pedigrees without recombination

Kui Zhang¹, Fengzhu Sun² and Hongyu Zhao^{3,*}

¹Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, AL 35294, USA, ²Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California 1042 W. 36th Place, Los Angeles, CA 90089, USA and ³Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, New Haven, CT 06520, USA

Received on July 28, 2003; revised and accepted on December 4, 2003

Advance Access publication July 1, 2004

ABSTRACT

Motivation: Haplotype reconstruction is an essential step in genetic linkage and association studies. Although many methods have been developed to estimate haplotype frequencies and reconstruct haplotypes for a sample of unrelated individuals, haplotype reconstruction in large pedigrees with a large number of genetic markers remains a challenging problem.

Methods: We have developed an efficient computer program, HAPLORE (HAPLOtype REconstruction), to identify all haplotype sets that are compatible with the observed genotypes in a pedigree for tightly linked genetic markers. HAPLORE consists of three steps that can serve different needs in applications. In the first step, a set of logic rules is used to reduce the number of compatible haplotypes of each individual in the pedigree as much as possible. After this step, the haplotypes of all individuals in the pedigree can be completely or partially determined. These logic rules are applicable to completely linked markers and they can be used to impute missing data and check genotyping errors. In the second step, a haplotype-elimination algorithm similar to the genotype-elimination algorithms used in linkage analysis is applied to delete incompatible haplotypes derived from the first step. All superfluous haplotypes of the pedigree members will be excluded after this step. In the third step, the expectation-maximization (EM) algorithm combined with the partition and ligation technique is used to estimate haplotype frequencies based on the inferred haplotype configurations through the first two steps. Only compatible haplotype configurations with haplotypes having frequencies greater than a threshold are retained.

Results: We test the effectiveness and the efficiency of HAPLORE using both simulated and real datasets. Our

results show that, the rule-based algorithm is very efficient for completely genotyped pedigree. In this case, almost all of the families have one unique haplotype configuration. In the presence of missing data, the number of compatible haplotypes can be substantially reduced by HAPLORE, and the program will provide all possible haplotype configurations of a pedigree under different circumstances, if such multiple configurations exist. These inferred haplotype configurations, as well as the haplotype frequencies estimated by the EM algorithm, can be used in genetic linkage and association studies.

Availability: The program can be downloaded from <http://bioinformatics.med.yale.edu>

Contact: hongyu.zhao@yale.edu

INTRODUCTION

With the completion of the Human Genome Project, millions of single nucleotide polymorphisms (SNPs) have become available through the coordinated efforts to identify genetic variants in the human genome. These markers will greatly facilitate the identification of genetic variants underlying complex diseases. Although methods based on individual SNPs may lead to significant findings, haplotypes may provide additional power to map disease genes (Akey *et al.*, 2001; Kruglyak, 1999; Zhang *et al.*, 2002; Zhao *et al.*, 2000). In addition, haplotypes may lead to insights on the factors influencing the dependences among genetic markers, i.e. linkage disequilibrium (LD), and such insights may prove essential to understand human evolutions (Daly *et al.*, 2001; Goldstein, 2001). The potential use of haplotypes has led to the initiation of the HapMap project to investigate haplotype patterns in the human genome in different populations (<http://www.genome.gov/10001688>). Haplotype inference and frequency estimates are an essential component of this endeavor.

*To whom correspondence should be addressed.

Molecular methods, such as allele-specific long-range PCR (Michlataos-Beloin *et al.*, 1996) or diploid-to-haploid conversion (Douglas *et al.*, 2001), can be used to directly assay haplotypes in diploid individuals. However, these methods are technologically demanding and expensive, making them impractical for large-scale studies. Instead, genotypes from individual markers, not haplotypes, are routinely collected in genetic studies. An alternative strategy for haplotype inference is through close relatives. Although this strategy may reduce haplotype ambiguity and improve the efficiency for haplotype frequency estimates (Becker and Knapp, 2003; Rohde and Fuerst, 2001; Schaid, 2002), haplotype ambiguity may still persist when the number of markers is moderately large (Hodge *et al.*, 1999), especially in the presence of missing data. Therefore, there is a great need to develop efficient and accurate statistical methods for haplotype inference from genotype data of unrelated individuals as well as of general pedigrees.

There is a growing number of articles on haplotype inference for unrelated individuals (Excoffier and Slatkin, 1995; Gusfield, 2001; Hawley and Kidd, 1995; Lin *et al.*, 2002; Long *et al.*, 1995; Niu *et al.*, 2002; Qin *et al.*, 2002; Stephens *et al.*, 2001). Similarly, many statistical and algorithmic methods have been developed for haplotype reconstruction for related individuals. Some of these methods are based on exact-likelihood computations (Du *et al.*, 1998; Lander and Green, 1987; Sobel *et al.*, 1995; Weeks *et al.*, 1995) or based on approximate-likelihood computations (Kruglyak *et al.*, 1996; Sobel *et al.*, 1995; Weeks *et al.*, 1995), while others rely on rule-based strategies (Haines, 1992; Li and Jiang, 2003; Nejati-Javaremi and Smith, 1996; O'Connell, 2000; Qian and Beckman, 2002; Tapadar *et al.*, 2000; Wijsman, 1987). All of these haplotyping methods and programs have elements in common and have their own weaknesses and strengths. The likelihood-based methods are limited to a small number of markers and small pedigrees, owing to the extensive computations required. Additional information and assumptions, such as recombination rates among the markers and Hardy-Weinberg equilibrium, are generally required to calculate the likelihood. The rule-based methods are *ad hoc* but they rely on fewer assumptions and generally run faster than likelihood-based methods. However, there is no assessment of the reliability of the results for rule-based methods. In addition, the rule-based methods can be computationally intensive if recombination events are allowed among markers and there are missing data. In fact, Li and Jiang (2003) showed that the problem of finding a minimum-recombinant haplotype configuration is in general NP-complete. This is to say that there is no polynomial time algorithm that guarantees to reconstruct the minimum-recombinant haplotype configuration for any input.

When no recombination events are allowed among the markers in the pedigrees, the complexity of the above algorithms can be substantially reduced. In the absence of

missing data, Li and Jiang (2003) proposed a polynomial-time exact algorithm to reconstruct all compatible haplotype configurations without recombination. These haplotype configurations can serve as preliminary data for estimating haplotype frequencies using likelihood-based methods (O'Connell, 2000). In addition, there is no need to require information on recombination rates in likelihood-based methods under the assumption of no recombination. This assumption may be valid due to a shift of focus on association studies that generally involve many tightly linked markers in a small region (e.g. Cox *et al.*, 2002; Daly *et al.*, 2001; Patil *et al.*, 2001), because recombination is an unlikely event for tightly linked genetic markers. Moreover, recent studies have shown that the human genome can be partitioned into large blocks with high LD and relatively low recombination, separated by short regions of low LD. Therefore, if the markers within the same haplotype block are analyzed together, it is reasonable to assume that there is no recombination among these markers across the pedigrees studied (Wang *et al.*, 2002).

Cox *et al.* (2002) first compiled zero-recombinant haplotype sets in families for using a simple rule-based program combined with a genotype elimination algorithm, then estimated the haplotype frequencies by the standard expectation-maximization (EM) algorithm. Generally, methods for haplotype reconstruction in pedigrees can (1) identify the partial or complete haplotypes carried by each individual and check for genotyping error and impute the missing data; (2) list all compatible haplotype pairs carried by each individual; (3) provide more reliable and accurate estimates of haplotype frequencies along with a set of unrelated individuals. However, existing methods are not ideally suited for these purposes involving many tightly linked genetic markers in large pedigrees. To meet this need, we have developed a three-step algorithm to identify all compatible haplotype sets in a pedigree and estimate haplotype frequencies using both pedigree data and unrelated individuals. In our algorithm, we first employ a set of logic rules, which are generalizations of those developed by Wijsman (1987) and Qian and Beckman (2002), to deduct all possible haplotypes for an individual in the pedigree. In the second step, we use the genotype elimination technique (Lange and Goradia, 1987; O'Connell and Weeks, 1999) to exclude inconsistent haplotypes in a pedigree. As a result, the algorithm will provide all the compatible haplotype configurations of a pedigree if such multiple configurations exist. These haplotype configurations can be used in association mapping to increase the statistical power (e.g. Zhao *et al.*, 2000). In our third step, the PL-EM algorithm (Qin *et al.*, 2002) is used to estimate haplotype frequencies based on the compatible haplotype configurations. Genotypes from unrelated individuals can be easily incorporated into this step. In the following, we first describe our methods in detail in the Methods section and then demonstrate the usefulness of our method through its application to simulated as well as real datasets.

METHODS

Basic assumptions

The current version of HAPLORE is developed to analyze genotype data from autosomes under the following assumptions: (1) markers are tightly linked, so that recombination among these loci is unlikely for the observed meioses; (2) there are no mutations; and (3) there are no genotyping errors.

Logic rules

In this section, we describe our logic rules in detail. A total of 13 rules are used with the first two rules to initialize the haplotypes in an individual and the other rules to update each individual's haplotypes using genotype information from its relatives. An extended pedigree is scanned by examining all the nuclear families in the pedigree using these rules sequentially.

Before we describe these 13 rules one by one, we define some notation and terminology used in our rules. For a given nuclear family, we use H, F, M and P to represent a haplotype in an offspring, the father, the mother and a parent where no distinction is made between the father and the mother. We call a pair of haplotypes anonymous if the parental origins of the haplotypes are unknown. The two haplotypes of an individual are denoted by the haplotype symbol followed by 1 and 2. For example, we use H1 and H2 to represent the two haplotypes that are anonymous in an offspring.

For a given haplotype, we use the symbol '−1' at a given locus to indicate that the alleles at this locus have not been assigned to the two haplotypes when the genotype is available at this locus. It is easy to see that if there is '−1' in one haplotype, the other haplotype must have '−1' at the same locus. With this notation, the exact haplotypes can be easily restored from such haplotypes in combination with the genotypes. For example, if (1,2), (1,3), (2,2), (1,1) denote the genotypes of an individual at four loci, then haplotypes H1 = (−1, 1, 2, 1) and H2 = (−1, 3, 2, 1) correspond to two possible haplotype pairs: {(1,1,2,1), (2,3,2,1)} and {(2,1,2,1), (1,3,2,1)}. Therefore, by using '−1' in a haplotype, we can use only two haplotypes to represent a set of compatible haplotype pairs. However, it should be noted that a set of compatible haplotype pairs might not always be able to be represented this way using one '−1'. More than one '−1' in the haplotype has to be used to make such a representation valid. For example, if (1,2), (3,4), (1,2), (3,4) denote the genotypes of an individual at four loci, and there are two compatible haplotype pairs, one of them is {(1,3,1,3), (2,4,2,4)}, the other is {(1,4,1,4), (2,3,2,3)}. Any pair of haplotypes with just one '−1' term cannot represent these two haplotype pairs. As a compromise, these two haplotype pairs are a subset of the following representations: {(1, −1, 1, −1), (2, −1, 2, −1)} or {(−1, 3, −1, 3), (−1, 4, −1, 4)}. If we use {(1, −1, 1, −1), (2, −1, 2, −1)} to represent them, the possible haplotype pairs are {(1,3,1,3), (2,4,2,4)}, {(1,4,1,4), (2,3,2,3)}, {(1,3,1,4),

(2,4,2,3)} or {(1,4,1,3), (2,4,2,3)}. The third and fourth haplotype pairs are inconsistent with the original haplotype pairs. However, it is necessary to use such representation to save computer memory and running time, which is often the bottleneck in haplotype inference. This compromise does sacrifice accurate haplotype reconstruction and may result in inconsistent haplotypes in our analysis, but such inconsistency will be resolved in the haplotype-elimination step and the EM algorithm step.

As mentioned above, the first two rules initialize haplotype assignment through parent–offspring pairs and trios using genotype data from all these individuals. This procedure deals with one locus at a time but is repeated for all the loci considered. It, thus, will produce a list of assignments (or unassignments) of marker alleles at each locus along the haplotype. The objective of these two rules is to determine the parental origin of alleles at each locus for the offspring.

Rule 1 (R1): For an individual and ONE of its parents, assign haplotypes at a locus if their genotypes satisfy one of the following conditions: (1) one of them is homozygous and (2) all are heterozygous, but their genotypes are not identical. Otherwise, assign '−1' to this locus in the haplotypes of these two individuals.

Rule 2 (R2): For an individual and BOTH of its parents, assign haplotypes at a locus if their genotypes satisfy one of the following conditions: (1) one of them is homozygous and (2) all are heterozygous, but their genotypes are not identical. Otherwise, assign '−1' to this locus in the haplotypes of these three individuals.

After applying these two rules, the haplotypes of each individual in the pedigree will be assigned. Some individuals have haplotypes with known origin, while others may only have anonymous haplotypes. It is important to note that an individual may be assigned different haplotypes through different pairs or trios when this individual has several offspring, or it has both parents and offspring. In this situation, we choose a haplotype assignment with the minimum number of '−1's through all these possible pairs and trios. The assigned haplotypes correspond only to the genotypes of individuals and the pedigree structure. It can avoid finding an optimal or nearly optimal individual order in the pedigree when we perform rules R1 and R2. It is also important to note that almost all individuals' haplotypes cannot be determined unambiguously by using such two simple rules only, especially for a large number of diallelic loci (Hodge *et al.*, 1999). In the example given in Table 1, the haplotypes of father and mother can be assigned through the trio of father, mother and offspring 1 or through another trio of father, mother and offspring 2. As there is only one '−1' in the haplotypes of the parents by using the first trio, we assigned them through this trio. Obviously, when we assign haplotypes of offspring 2 through its parents, we can find two '−1's in its haplotypes. It is easy to determine the origin of the haplotypes in these offspring.

Table 1. The haplotype assignments for a family by using rules R1 and R2

Locus	Genotype					Haplotype after rules R1 and R2				
	1	2	3	4	5	H1	H2	F	M	
Father	12	11	12	22	12	H1	H2	F	M	
Mother	12	12	12	12	11	H1	H2	F	M	
Offspring 1	11	12	12	22	12	F	M	F	M	
Offspring 2	12	12	12	22	12	F	M	F	M	

Once the haplotypes of the individuals in the pedigree are assigned, we use a set of rules to infer their inheritance patterns: where they are from and where they transmit. R3 and R4 are used to identify the parental origins of the haplotypes in an offspring and R5 and R6 are used to identify which parental haplotype is transmitted to the offspring. These rules jointly utilize the information of each individual's haplotypes and genotypes. Thus, they are more general than the rules developed by Wijnsman (1987). We define two haplotypes to be different if one haplotype contains an allele that is different from the allele in the other haplotype at a certain locus. For example, $(-1, -1, 2, 2)$ and $(1, 2, -1, 2)$ are not different, but $(-1, -1, 1, 2)$ and $(1, 2, -1, 1)$ are different because the alleles at the fourth locus differ.

Rule 3 (R3): For an offspring with haplotypes {H1, H2} and one of its parents, H1 can be identified to be inherited from this parent if one of the following conditions is satisfied: (1) H1 is known from this parent; (2) at a certain locus, the allele in H2 is not in this parent's genotype; and (3) H2 is different from the two haplotypes in this parent.

Rule 4 (R4): For an offspring with haplotypes {H1, H2} and both of its parents, H1 can be identified to be of paternal origin if one of the following conditions is satisfied: (1) H1 is known to be paternal; (2) at a certain locus, the allele in H2 is not in the father's genotype; (3) at a certain locus, the allele in H1 is not in the mother's genotype; (4) H2 is different from the two haplotypes in the father; and (5) H1 is different from the two haplotypes in the mother.

Rule 5 (R5): For an offspring and one of its parents with haplotypes {P1, P2}, P1 must be transmitted to this offspring if one of the following conditions is satisfied: (1) at a certain locus, the allele in P2 is not in the offspring's genotype and (2) P2 is different from the two haplotypes in the offspring.

Rule 6 (R6): For an offspring and both of his parents, if the father's haplotypes are {F1, F2}, then F1 must be transmitted to the offspring if one of the following conditions is satisfied: (1) at a certain locus, the allele in F2 is not in the offspring's

Table 2. Determination of haplotypes' origins by using rules R3–R6

Locus	Genotype					Haplotype after rules R3–R6				
	1	2	3	4	5	H1	H2	F	M	
Father	12	11	12	12	22	H1	H2	F	M	
Mother	12	12	12	12	11	H1	H2	F	M	*
Offspring	12	11	12	12	12	H1	H2	F	M	

genotype; (2) F2 is different from the two haplotypes in the offspring; and (3) by assuming F2 being transmitted to the offspring, the other haplotype in the offspring, namely H and derived from F2 as one of the offspring's haplotypes and this offspring's genotypes, does not have a maternal origin from R3 or R4.

In the example given in Table 2, we have trios that are from a large pedigree and their haplotypes are already assigned by the other relatives of these individuals. We do not know which haplotype of the offspring is inherited from its father at first. But there is an allele '1' at the 5th locus in haplotype H2 and '1' is not found in father's genotype, so H1 must be from father by using R3 and H2 from mother at the same time. Further more, by using R5, we can identify that haplotype H1 (indicated by an asterisk in the Table) of the mother is transmitted to offspring H2, because haplotype H2 of the mother is different from the two haplotypes of the offspring.

After applying R3–R6, we use R7–R9 in conjunction with R3–R6 to change the haplotypes in an individual. R7 only assigns an allele to an unassigned locus and will not change the assigned locus to unassigned. R8 and R9 can reduce the number of unassigned loci in the haplotypes, equivalent to reducing the number of '-1's in the haplotype. So during the use of these rules, some unassigned loci become assigned, and some assigned loci become unassigned again simultaneously.

Rule 7 (R7): For an offspring and one of its parents, suppose we can identify that haplotype P in the parent is haplotype H in the offspring using R3–R6. (1) If there exists a locus that P is assigned but H is not assigned, assign the allele at this locus in P–H. (2) If there exists a locus that P is not assigned but H is assigned, assign the allele at this locus in H to P.

Rule 8 (R8): For an offspring and one of its parents, suppose we can identify that haplotype H in the offspring is inherited from this parent, but cannot identify which haplotype in the parent produced H. (1) If the number of '-1's in H is smaller than that in the parental haplotypes, anonymously replace one of the parental haplotypes by H. (2) If the number of '-1's in H is more than that in the parental haplotypes and the parent is homozygous at those loci which are unassigned

Table 3. Change of haplotype assignments in a nuclear family by using rules R7–R9

Locus	Genotype					Haplotype before rules R7–R9					Haplotype after rules R7–R9						
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5		
Father	12	11	12	12	22	F	1	1	–1	–1	2	→H1	–1	1	1	1	2
						M	2	1	–1	–1	2	→H2	–1	1	2	2	2
Mother	12	11	12	12	11	F	2	1	–1	–1	1	→H1	–1	1	2	2	1
						M	1	1	–1	–1	1	→H2	–1	1	1	1	1
Offspring	12	11	12	12	12	F	–1	1	1	1	2	→F	–1	1	1	1	2
						M	–1	1	2	2	1	→M	–1	1	2	2	1

in H but assigned in the parent, replace H by one of the parental haplotypes. (3) If the number of ‘–1’s in H is more than that in the parental haplotypes and the parent is heterozygous at those loci which are unassigned in H but assigned in the parent, anonymously replace H by one of the parental haplotypes.

Rule 9 (R9): For an offspring and one of its parents, suppose we can identify that haplotype P in the parent is transmitted to the offspring, but cannot identify which haplotype in the offspring corresponds to P. (1) If the number of ‘–1’s in P is smaller than that in the offspring’s haplotypes, replace the offspring’s haplotypes by P. (2) If the number of ‘–1’s in P is more than that in the offspring’s haplotypes and the offspring is either all homozygous or all heterozygous at those loci which are unassigned in P but assigned in the offspring’s haplotypes, anonymously replace P by one of the offspring’s haplotypes.

In the example given in Table 3, we have trios that are in a large family and their haplotypes have been assigned through the other relatives of these individuals. We know the origins of the haplotypes in the father, the mother and the offspring. For the haplotype F and M in the father, we cannot identify which one transmits to the offspring by using rules R3–R6. As the number of ‘–1’s in the haplotypes is more than that in the haplotypes of the offspring, the haplotypes of the father are anonymously replaced by the haplotype F in the offspring using R8. The haplotypes of the mother are similarly changed.

Rules R3–R6 are very effective to determine the complete or partial inheritance pattern of the haplotypes. Therefore, we can use rules R7–R9 to change the assignment of the haplotypes. However, sometimes R3–R6 may fail to identify such patterns. Even under such situation, we can still change their haplotype assignment using the information of relation between haplotypes among offspring and its parents. Rules R10 and R11 are similar to R8 and R9 to attempt to reduce the number of unassigned loci in the haplotypes for an individual under some subtle conditions.

Rule 10 (R10): For an offspring and one of its parents, suppose we cannot identify the parental origins of the offspring’s haplotypes and we cannot identify which of the parental haplotypes is transmitted to the offspring through R3–R6. We further assume that the number of ‘–1’s in the offspring’s haplotype is smaller than that in the parental haplotypes. Then one of the parental haplotypes can be anonymously replaced by one of the offspring’s haplotypes if all of the following conditions are satisfied: (1) the offspring is either all homozygous or all heterozygous at those loci which are assigned in the parental haplotypes but unassigned in the offspring’s haplotypes and (2) their four haplotypes have the same allele at every locus that are all assigned.

Rule 11 (R11): For an offspring and one of its parents, suppose we cannot identify the parental origins of the offspring’s haplotypes and cannot identify which of the parental haplotypes is transmitted to the offspring through R3–R6. We further assume that the number of ‘–1’s in the offspring’s haplotypes is more than that in the parental haplotypes. Then one of the offspring’s haplotypes can be replaced by one of the parental haplotypes if all of the following conditions are satisfied: (1) the parent is either all homozygous or all heterozygous at those loci which are assigned in the parental haplotypes but unassigned in the offspring’s haplotypes and (2) their four haplotypes have the same allele at every locus that are all assigned.

Rules R8–R11 play important roles in our haplotype reconstruction. The purpose of these rules is to reduce the number of ‘–1’s in the haplotypes as much as possible by using the relative’s haplotype. The underlying idea for these rules is that the decrease in the number of unassigned loci will increase the information to identify haplotypes’ inheritance pattern. The fewer the number of ‘–1’s in the haplotypes, the more likely that the complete and partial inheritance pattern of haplotypes can be determined. This is why we allow the change of an assigned locus to an unassigned locus in our rules, which is forbidden in rules developed by Wijnsman (1987). The flexibility of our rules can greatly increase the efficiency in haplotype

Table 4. An example of the usefulness of rules R10 and R11

People ID	Father ID	Mother ID	Genotype locus				
			1	2	3	4	5
101	0	0	12	11	12	12	11
102	0	0	12	11	12	12	11
201	101	102	12	11	12	12	11
202	0	0	12	12	12	12	12
301	201	202	11	12	11	22	12

reconstruction. As an example, we apply HAPLORE to the family shown in Table 4. The haplotypes of individuals 201, 202, 301 are easily identified and the haplotypes of individuals 101, 102 have several haplotype configurations in the initial analysis, which are shown in Table 5. After rule R10 is used, all of the five individuals have a unique haplotype configuration as that shown in Table 5. While PATCH, developed by Wijsman (1987) on the basis of her rules, can only determine that individuals, 201, 202 and 301 have unique haplotype pairs. It cannot reduce the four haplotype pairs in individuals 101, 102 to 1.

After rules R7–R11, some loci can be changed from assigned to unassigned, we then use rules R12–R13 to change them back. Although these two rules seem trivial, they are quite useful in haplotype reconstruction (Wijsman, 1987).

Rule 12 (R12): For the haplotypes in an individual, if a certain locus is unassigned but this person is homozygous at this locus, assign the homozygous allele to this locus in the haplotypes.

Rule 13 (R13): For the haplotypes in an individual, if all of the assigned loci are homozygous and there exists an unassigned heterozygous locus, anonymously assign an allele to this haplotype at this locus.

Consider the example given in Table 6, all of the assigned loci are homozygous and there exists an unassigned heterozygous locus, so we assign ‘1’ at the 5th locus to the haplotype and the origin of this becomes anonymity.

To summarize our rules, we use R1 and R2 to initialize the haplotypes in an individual and sequentially apply the other rules to update each individual’s haplotypes by scanning all nuclear families in the pedigree. The change of the haplotype assignment of any individual in a family could possibly affect the assignment for the other individuals of this family and the individuals of neighborhood families in the pedigree. We repeat the process until no further changes to the haplotypes can be made. Our program is not sensitive to the order of the individuals in the pedigree although the logic rules must follow a certain order. In Figure 1, we show a logic flow of the rules used in HAPLORE.

The above rules are developed under the assumption that the genotypes of all individuals in the pedigree are known. They can easily be extended to incorporate missing data. When genotype data are completely or partially missing for some

individuals, a special allele ‘0’ is introduced in our analysis to represent the missing allele in genotypes as well as in haplotypes. By this notation, the allele ‘0’ in a haplotype of an individual indicates that we cannot assign an allele to this haplotype and there is at least one missing allele in its genotype at this locus. Therefore, it is straightforward to generalize the logic rules to incorporate missing data: (1) rules R1 and R2 are generalized to assign the haplotypes and impute the missing alleles in genotypes according to the rules developed elsewhere (Wijsman, 1987; Qian and Beckman, 2002). (2) In rules R8–R11, the total number of ‘–1’ and ‘0’ is used as a criterion to change the haplotype assignment for each parent-offspring trios. (3) Before performing rules R12 and R13, a simple rule is developed to impute the missing data by comparing the two haplotypes and the genotype of an individual.

The haplotype-elimination algorithm

It is common practice in linkage analysis to use the genotype elimination technique to identify those genotypes that need not be considered during the likelihood calculation. This technique was proposed to accelerate the computation of likelihood (Lange and Boehnke, 1983; Lange and Goradia, 1987), and it is fully efficient for pedigrees without loops, i.e. all extraneous genotypes in the pedigree can be excluded using this approach. An extension of this algorithm developed by O’Connell and Weeks (1999) is efficient and guaranteed to eliminate all superfluous genotypes in all types of pedigrees, including those with loops.

Theoretically, when the haplotypes in the pedigree are treated as alleles at a single locus, the genotype elimination algorithm can be used to reconstruct haplotypes directly (Lange and Weeks, 1989; O’Connell, 2000; Cox *et al.*, 2002), especially under our assumptions. O’Connell (2000) used such algorithms, together with a haplotype-recoding scheme and a divide-and-conquer strategy to reconstruct zero recombinant haplotypes. In the divide-and-conquer approach, all of the SNPs are broken down into units that contain only several SNPs and have one or two common SNPs with adjacent units. The haplotype-elimination algorithm is first employed within each unit. Then two adjacent partial haplotypes are combined using the haplotype-elimination algorithm again. This approach is efficient when many possible haplotype pairs can be removed in each unit, but will not reduce the complexity when some individuals have all possible haplotype pairs. Using our rule-based method proposed above, the complexity could be further reduced. First, we use a novel method to reconstruct the haplotypes in the pedigree, so the number of possible haplotypes in an individual can be lowered before performing the elimination procedure. Second, the rule-based method can eliminate incompatible haplotypes before and after each divide-and-conquer step. Third, our novel haplotype representations make it easy to

Table 5. The haplotypes identified on the basis of genotypes in Table 4 before and after using rules R10 and R11

People ID	Father ID	Mother ID	Haplotype before using rules R10 and R11					Haplotype after using rules R10 and R11						
			Locus					Locus						
			1	2	3	4	5	1	2	3	4	5		
101	0	0	H1	-1	1	-1	-1	1	H1	1	1	1	2	1
			H2	-1	1	-1	-1	1	H2	2	1	2	1	1
102	0	0	H1	-1	1	-1	-1	1	H1	1	1	1	2	1
			H2	-1	1	-1	-1	1	H2	2	1	2	1	1
201	101	102	H1	1	1	1	2	1	H1	1	1	1	2	1
			H2	2	1	2	1	1	H2	2	1	2	1	1
202	0	0	H1	1	2	1	2	2	H1	1	2	1	2	2
			H2	2	1	2	1	1	H2	2	1	2	1	1
301	201	202	F	1	1	1	2	1	F	1	1	1	2	1
			M	1	2	1	2	2	M	1	2	1	2	2

Table 6. Change of haplotype assignments in an individual using rules R12 and R13

Locus	Genotype					Haplotype before R12 and R13					Haplotype before R12 and R13						
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5		
Individual	11	22	12	22	12	F	1	2	-1	2	-1	→ H1	1	2	-1	2	1
						M	1	2	-1	2	-1	→ H2	1	2	-1	2	2

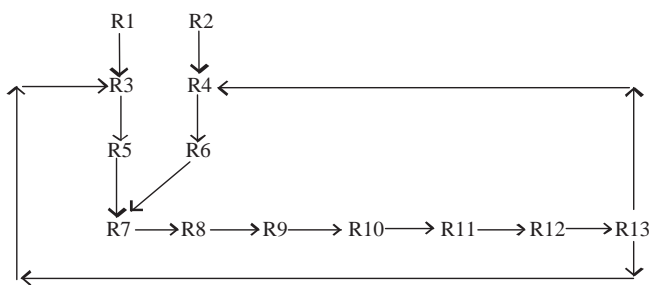


Fig. 1. Flow of logic rules used in the haplotype reconstruction. Rules R1–R13 are described in the text.

resolve all the possible haplotype pairs in the individual from its haplotypes and genotypes. It will play an important role as the number of marker loci and the number of marker alleles increase.

The PL–EM algorithm

The EM algorithm has been widely used to estimate haplotype frequencies based on genotype data from unrelated individuals as well as genotype data from general pedigrees

(Excoffier and Slatkin, 1995; Long *et al.*, 1995; O’Connell, 2000) due to its interpretability and stability. The accuracy assessment of the EM algorithm performed both in simulation studies (Fallin and Schork, 2000) and with molecular haplotyped data (Tishkoff *et al.*, 2000; Zhang *et al.*, 2001) indicated the good performance of the estimated frequencies of common haplotypes from unrelated individuals. In addition, genetic information from relatives in a general pedigree can help us resolve haplotype ambiguity. Even if ambiguities still exist with data from a very large pedigree, the reduction of haplotype ambiguity can help us improve the efficiency for estimating haplotype frequencies (Becker and Knapp, 2003; Rohde and Fuerst, 2001).

In this section, we first focus on deriving an EM algorithm for estimating haplotype frequencies from general pedigrees. In addition, we assume Hardy–Weinberg equilibrium (HWE) for haplotypes carried by unrelated individual. This is a fundamental assumption of the EM algorithm for haplotype frequency estimations (Excoffier and Slatkin, 1995; Hawley and Kidd, 1995). Note that recombination rates are not necessary in such inference. Under the assumption of no recombination, each haplotype can be recoded as an allele at a single

multiallelic locus. The multilocus genotypes can then be represented as single-locus genotypes using the recoded alleles, like those in the previous section and elsewhere (O'Connell, 2000).

We introduce the following notation in our computation. Suppose that there are a total of K haplotypes comprised SNPs s_i, \dots, s_j : $H = \{h_1, h_2, \dots, h_K\}$. Their frequencies in the population are represented as $\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$. Assume that we have a total of N families. We further assume that an individual has either both parents or has no parents at all. We define an individual in a pedigree as a non-founder if both parents are available and others are referred as founders. In a family f ($1 \leq f \leq N$), let n_f be the number of founders and m_f be the total number of individuals in the family. The n_f founders are indexed as $1, \dots, n_f$ and the $m_f - n_f$ non-founder individuals are indexed as $n_f + 1, \dots, m_f$. The genotype and two haplotypes of individual r in family f are denoted as G_{f_r} and $H_{f_r} = \{h_{f_r,1}, h_{f_r,2}\}$, respectively. It is likely that many haplotype pairs are compatible with the genotypes of families under the assumption of no recombination events. Let S_{f_r} denote all the compatible haplotype pairs for individual r in family f . It is worth noting that the unrelated individuals can be included into our model. Each individual forms a family and is the only founder member in this family. In this case, S contains all possible haplotype pairs compatible with the genotype of this individual. The objective is to estimate the haplotype frequencies based on the genotypes of the families. For a family f , the likelihood of the genotypes of the family is:

$$\begin{aligned} & L_f(G_{f_1}, G_{f_2}, \dots, G_{f_{m_f}} | \Theta) \\ &= \sum_{H_{f_1} \in S_{f_1}} \cdots \sum_{H_{f_{m_f}} \in S_{f_{m_f}}} \Pr(G_{f_1}, \dots, G_{f_{m_f}}, \\ & \quad \times H_{f_1}, \dots, H_{f_{m_f}} | \Theta) \\ &= \sum_{H_{f_1} \in S_{f_1}} \cdots \sum_{H_{f_{m_f}} \in S_{f_{m_f}}} \Pr(G_{f_1}, \dots, G_{f_{m_f}} \\ & \quad | H_{f_1}, \dots, H_{f_{m_f}}, \Theta) \Pr(H_{f_1}, \dots, H_{f_{m_f}} | \Theta) \\ &= \sum_{H_{f_1} \in S_{f_1}} \cdots \sum_{H_{f_{m_f}} \in S_{f_{m_f}}} \prod_{r=1}^{n_f} \Pr(H_{f_r} | \Theta) \\ & \quad \times \prod_{r'=n_f+1}^{m_f} \Pr(H_{f_{r'}} | H_{f_{r'}}^F, H_{f_{r'}}^M), \end{aligned}$$

where $\Pr(H_{f_r} | \Theta) = 2\theta_{f_r,1}\theta_{f_r,2}$ if $h_{f_r,1} \neq h_{f_r,2}$ and $\Pr(H_{f_r} | \Theta) = \theta_{f_r,1}^2 + \theta_{f_r,2}^2$ if $h_{f_r,1} = h_{f_r,2}$ for $r = 1, \dots, n_f$ under the assumption of HWE; and $\Pr(H_{f_{r'}} | H_{f_{r'}}^F, H_{f_{r'}}^M)$ ($r' = n_f + 1, \dots, m_f$) is the gamete transmission probabilities for unordered genotypes where $H_{f_{r'}}^F$ and $H_{f_{r'}}^M$ are the haplotype pairs for the father and mother, respectively (Elston and Stewart, 1971). The likelihood of all the data is obtained

by multiplying $L_f(G_{f_1}, \dots, G_{f_{m_f}} | \Theta)$ across all the families. We can estimate Θ using the maximum-likelihood estimation (MLE) approach. As it is difficult to directly obtain the MLE of Θ , the EM algorithm is frequently used to estimate Θ .

Suppose we know that $\Theta = \Theta^{(k)}$ and we want to estimate $\Theta^{(k+1)}$. In the E-step, we introduce the following notation. Let

$$\begin{aligned} \alpha_{l,f}(H_{f_1}, \dots, H_{f_{n_f}}) &= \{\#h_l \text{ in } H_{f_1}, \dots, H_{f_{n_f}}\} \\ &= \sum_{i=1}^{n_f} [I_{h_{f_i,1}}(h_l) + I_{h_{f_i,2}}(h_l)] \end{aligned}$$

which is the number of haplotype l presented in the founders of family f ($l = 1, 2, \dots, K$). Let

$$\begin{aligned} \beta_{l,f}^{(k)} &= \sum_{H_{f_1} \in S_{f_1}} \cdots \sum_{H_{f_{m_f}} \in S_{f_{m_f}}} \alpha_{l,f}(H_{f_1}, \dots, H_{f_{n_f}}) \\ & \quad \times \prod_{r=1}^{n_f} \Pr[H_{f_r} | \Theta^{(k)}] \prod_{r'=n_f+1}^{m_f} \Pr(H_{f_{r'}} | H_{f_{r'}}^F, H_{f_{r'}}^M) \quad (1) \end{aligned}$$

which is the weighted number of haplotype l appeared in the founders of family f . We define a normalizing constant $C_f^{(k)}$ satisfying $C_f^{(k)} \sum_{l=1}^K \beta_{l,f}^{(k)} = 2n_f$ to evade the calculation of Mendelian likelihood of the whole family. In the M-step, we can estimate $\Theta^{(k+1)}$ from all families as follows:

$$\theta_l^{(k+1)} = \frac{\left[\sum_{f=1}^N C_f^{(k)} \beta_{l,f}^{(k)} \right]}{(2 \sum_{f=1}^N n_f)} \quad l = 1, 2, \dots, K.$$

The EM-based method of estimating haplotype frequencies from general pedigrees can be very computationally intensive, since the likelihood for each configuration of founder haplotype pairs is computed by Equation (1) during each iteration. In HAPLORE, we employ the PL-EM technique to solve this problem (Qin *et al.*, 2002). In the PL-EM algorithm, all of the SNP loci are broken down into 'atomistic' units that only contain several SNPs (usually 5–8 SNPs) and have one or two common SNPs with adjacent units. The EM algorithm is first applied to each unit to infer haplotype frequencies within this unit. Then two adjacent partial haplotypes are 'ligated' using the EM algorithm again. Only the haplotypes with frequencies greater than a threshold (e.g. 10^{-6}) and a fixed number of haplotypes with frequencies less than this threshold are retained in each EM step. In general, the EM algorithm is time and memory efficient only for a small number of SNPs. Thus, this strategy could solve the speed and memory constraint generally existing in the EM algorithm and makes it suitable for large-scale recovery of haplotypes from genotype data. We use two additional strategies to reduce the computational complexity of the PL-EM algorithm in haplotype inference for general pedigrees: (1) all compatible haplotype configurations for each individuals are obtained using the rule-based algorithm and the haplotype-elimination algorithm described

in the previous sections. (2) For each family, a list is created to store all compatible haplotype configurations for founders that have non-zero transmission probability. This is based on the observation that the number of such configurations is much smaller than the product of the number of haplotype pairs for founder individuals. In other words, many items in Equation (1) are zero. As Equation (1) is calculated iteratively and the calculation of the transmission probabilities does not depend on Θ , the use of this list avoids computing the transmission probabilities repeatedly. We also suggest using a smaller number of SNPs in each unit, since even the number of compatible haplotype pairs for each founder is moderate, the total number of configurations could be too large to be calculated for large pedigrees. Furthermore, the haplotype-elimination algorithm itself can be very time consuming, even for a moderate number of SNPs in the presence of missing data.

RESULTS

The computer program and the test datasets

We have implemented the set of logic rules, the haplotype-elimination procedure and the PL-EM algorithm for general pedigrees in a computer program, named HAPLORE (HAPLOtype REconstruction), and have tested it on a Windows 2000 system having one Pentium IV 1400 MHz processor and 256 MB memory using simulated as well as a real data sets.

We use a data set simulated for the 12th Genetic Analysis Workshop (GAW12) to test HAPLORE (Wijsman *et al.*, 2001). This dataset provides two populations, a large general population and an isolated population founded approximately 20 generations ago by approximately 100 individuals from the general population. For each population, 50 replicates are generated. Different phenotypic and genotypic data are available for the same 23 extended families with a total of 1497 individuals (1000 living) for each replicate. The numbers of individuals for these families vary from 37 to 128. The sequence data for seven candidate genes, whose lengths vary from 13 (gene 2) to 20 kb (genes 1, 4 and 7), are only available for 1000 living individuals. The overall fraction of missing data is 33%. These multiple SNP markers are so tightly linked that they are ideal for testing our program.

We also apply HAPLORE to the Oxford ACE data, which was used to study the functional mutation in the angiotensin-I converting enzyme (ACE) gene due to a quantitative trait locus (Keavney *et al.*, 1998) and as a testing dataset before (O'Connell, 2001). The data contain 10 tightly linked biallelic markers in strong disequilibrium, spanning very small region (26 KB) within the ACE gene. There are a total of 666 individuals in 83 families. Most families have 2–3 generations with 5–18 individuals. After excluding a family with recombination events, a total of 82 families comprising 661

individuals are included in our final analysis. Among 239 individuals having missing data, 110 individuals are completely un-genotyped and 125 individuals have missing data at less than 3 loci. The overall fraction of missing data is 20.0%.

The results for the rule-based algorithm

We test the effectiveness and the efficiency of the rule-based algorithm using the GAW12 data and the Oxford ACE data. We consider the simulated GAW12 data from both the isolated population and the general population. The results from the analyses of 50 replications of the isolated population as well as of the general population are summarized in Table 7. As can be seen, although the number of completely haplotyped individuals varies for different genes, the haplotype pairs for a large number of individuals can be uniquely determined after rules R1–R13 are used. The algorithm can identify a certain number of individuals having haplotypes with no more than 10 unassigned loci. The overall fraction of missing data is reduced up to 50% (from 33.3 to $\sim 16.0\%$), indicating the effectiveness of the imputation by the logic rules. Note that most individuals in the first and the second generations in the pedigree are not genotyped at all. Thus it is difficult to completely haplotype these individuals. We can find that the use of pedigree information (R3–R13) increases the number of completely haplotyped individuals substantially and reduces the missing rate moderately. As an example, there are about 557 SNPs in gene 1 for the general populations. When only rules R1 and R2 are used, we can uniquely identify haplotypes in 284 individuals out of 1000 completely genotyped individuals. This number increases to 675 after R1–R13 are applied. The fraction of missing data is reduced from 20.5 to 15.6%. A more significant example is from gene 7. Only 89 individuals for the general population are completely haplotyped when only R1 and R2 are applied. This number increases to 622 after the other rules are applied. It is also worth noting that the results can be obtained in <10 s for the largest gene containing more than 500 SNPs.

We also apply the logic rules to reconstruct haplotypes of the Oxford ACE data. After using HAPLORE, there are 17 families in which haplotypes of every individual can be uniquely identified. A total of 356 individuals have a unique pair of haplotypes. The overall fraction of missing data is reduced from 20.0 to 10.3%. The number of individuals having missing data is decreased to 183. Again, the results here demonstrate that the rules R3–R13 play an important rule in this step. When R1 and R2 are used, only 233 individuals have been completely haplotyped and the overall fraction of missing data is 11.1%.

As a comparison, we apply PATCH (Wijsman, 1987) to the same two datasets. For the first replicate in the isolated population of the GAW 12 data, PATCH successfully identifies 781 individuals with unique haplotype pairs, while 682 individuals are completely haplotyped by HAPLORE. For the Oxford ACE data, both PATCH and HAPLORE can identify about

Table 7. Results of haplotype reconstruction using the rule-based algorithm for the GAW12 data

	Gene	Average number of loci ^a	After using R1–R2		Missing rate ^c (%)	After Using R1–R13		Missing rate(%)
			Average number of individuals $k = 0^b$	$1 \leq k \leq 10$		Average number of individuals $k = 0$	$1 \leq k \leq 10$	
The isolated population	1	290	326	312	20.5	684	129	16.6
	2	158	331	423	20.6	687	198	16.7
	3	159	354	401	20.5	693	139	16.6
	4	211	404	276	20.5	701	265	17.2
	5	44	538	423	20.7	703	265	18.4
	6	65	523	415	20.4	742	256	17.9
	7	233	123	426	20.8	630	137	15.8
The general population	1	557	284	368	20.5	675	142	15.6
	2	277	343	416	20.2	678	202	16.1
	3	249	361	440	20.2	697	203	16.0
	4	405	376	269	20.2	700	122	16.4
	5	53	553	392	20.6	777	236	18.2
	6	114	507	428	20.1	743	256	17.3
	7	279	89	466	20.6	622	154	15.1

^aThis is the number of markers at which at least one individual has one or two copies of the non-ancestral sequence variant.

^bThe number of unassigned loci in the haplotype.

^cThe overall fraction of missing data after the rules used.

360 individuals having a unique pair of haplotypes. However, some incompatibilities are found when using PATCH. This may cause PATCH to miss some haplotype assignments compatible with the data. As an example, the fourth family in the Oxford ACE data contains 18 individuals in 3 generations. The haplotypes of individuals 1 and 2 in the first generation can be obtained by the haplotypes from their children 3, 5 and 9 whose haplotypes are determined by their spouse and offspring. The haplotypes of individuals 1 and 2 given by PATCH are (111112111,0000021222) and (1111121222,222221222), respectively. First, the genotypes of these two individuals are completely missing. So the haplotypes for them should be symmetric. But the PATCH result is not symmetric. Second, even if we ignore the symmetry issue, some haplotypes compatible with the data are missed in the assignments using PATCH. For example, the haplotypes for individuals 1 and 2 can also be (111112111,1111121222) and (111112111, 2222221222), respectively, but they are not included in the PATCH assignments. On the other hand, the assignments using HAPLORE include all haplotypes compatible with the data.

To test the effectiveness and efficiency of the rule-based method in completely genotyped pedigrees, we apply HAPLORE to haplotype individuals from simulated datasets using the same pedigree structures but with every individual genotyped. In our simulations, the haplotypes of the founders are randomly sampled from 129 distinct haplotypes, with the number of SNPs equal to 296. These haplotypes are derived from the first replication for gene 1 using the isolated population. The haplotypes of the non-founders are randomly inherited from their parents under Mendelian inheritance. For this complete dataset, the haplotypes of all of the 1497

individuals can be uniquely determined in <10 s. When PATCH (Wijmsman, 1987) is used, it also uniquely determines the haplotypes of all 1497 individuals, but it takes more than 1 h.

The results for the haplotype-elimination algorithm and the PL–EM algorithm

Owing to the high missing rate and the large pedigrees, it is beyond the capability to perform the haplotype-elimination algorithm and the PL–EM algorithm for the GAW12 data. Therefore, we focus on the performance of HAPLORE based on the Oxford ACE data in the rest part of this section.

As recoding haplotypes and performing haplotype elimination are simple and straightforward, we only discuss the complexity for haplotype elimination with the rule-based algorithm and without the rule-based algorithm. Without the rule-based algorithm, the completely ungenotyped individuals have all possible $(1024 + 1) * 1024 / 2$ haplotype pairs at the beginning of the haplotype-elimination algorithm. The typed individuals with k ($1 \leq k \leq 10$) heterozygous SNP loci have 2^{k-1} possible haplotype pairs. After the rule-based algorithm, some individuals still have the same number of possible haplotype pairs. As can be expected, the haplotype-elimination algorithm is very time consuming for this dataset. Thus, we compare two quantities for haplotype elimination with the haplotypes constructed by the rule-based algorithm for a subset of SNPs. The two quantities are the number of possible haplotype pairs before haplotype elimination and that in each individual after haplotype elimination, respectively. Using the haplotypes obtained from the rule-based algorithm as the prior, the first quantity is substantially reduced in about 300 out of 661 individuals, but the second

Table 8. The results for the different length of unit and the overlap between units used in the PL–EM algorithm to infer haplotype frequencies by the Oxford ACE data

The length of unit (L)	The overlap between units (O)	The running time (s)	The number of retained haplotypes	The number of common haplotypes ^a	The fraction of common haplotype(%)
4	0	1601	27	5	92.14
4	1	281	24	5	92.14
4	2	200	24	5	92.14
5	0	8514	33	5	92.14
5	1	422	24	5	92.14
5	2	459	24	5	92.14
6	0	3135	24	5	92.14
6	1	3858	24	5	92.14
6	2	4094	24	5	92.14

^aThe common haplotypes are those with frequencies $>5\%$.

number is only reduced in about 10 individuals. It takes <2 min to perform haplotype elimination with or without the rule-based algorithm when the first six SNPs are used. However, the time increases to 25 min when the first seven SNPs are included in the analysis. In all these experiments, there is no difference for the running time between the data with and without the rule-based algorithm, indicating that the benefit from the rule-based algorithm is limited, especially when the divide and conquer technique is employed (O'Connell, 2000).

To perform the PL–EM algorithm and compare with the results obtained by ZAPLO (O'Connell, 2000), we set the convergence criterion as 10^{-5} and only retain the haplotypes that have frequencies greater than a threshold of 10^{-5} . The running time, the total number of retained haplotypes, the total number of haplotypes with frequencies $>5\%$ and their fraction with the varied length of units (L) and the overlap (O) between two adjacent units are shown in Table 8. In all settings, we can estimate five most common haplotypes and their frequencies, accounting 92.14% of the variation. The other haplotypes have frequencies $<1\%$. Among these haplotypes, 16 or 17 haplotypes, depending the different values of L and O , have frequencies $>0.01\%$ and account 7.62% of the variation. The overall difference of the 21 most common haplotype frequencies is $<10^{-6}$ for the different L and O . The six most common haplotypes and their frequencies are comparable with those estimated by O'Connell (2000) through a divide-and-conquer-with-pruning approach. The frequencies of the six most common haplotypes are 39.32, 30.69, 9.10, 7.88, 5.16 and 0.985%, respectively, when $L = 4$ and $O = 1$. The 7th most haplotype has frequency 0.974%, which is bigger than 0.767% estimated previously (O'Connell, 2000).

We note that there are substantial differences for the running time and the number of haplotypes retained between using different L and O . In most cases, only 24 haplotypes have frequencies higher than 10^{-6} , which is less than the 33 haplotypes estimated previously (O'Connell, 2000). This

maybe due to the fact that the PL–EM algorithm can be trapped in a local mode (Qin *et al.*, 2002). Some partial haplotypes with frequencies less than the threshold are discarded in a previous EM and ligation step, resulting in the removal of the haplotypes containing these partial haplotypes. To resolve this problem, we keep five additional haplotypes with highest frequencies among those haplotypes with frequencies less than the threshold in each EM step. A total of 37 haplotypes are identified with frequencies $>10^{-6}$ in more than 48 000 s for $L = 4$ and $O = 1$. However, five haplotypes with frequencies more than 5% still account for 92.14% of the variation. There are an additional 16 haplotypes having frequencies between 0.01 and 5%, accounting for 7.62% of the variation. In other words, the estimates of the frequency for the common haplotypes are not different from those obtained using different L and O . If our primary focus is to estimate the common haplotypes and their frequencies efficiently, we suggest using $L = 4$ or $L = 5$ and $O = 1$ or $O = 2$ to speed up this process.

DISCUSSION

Recent literature has suggested that haplotype-based methods may be more powerful than single marker-based approach in the identification of genes underlying complex diseases. As the current molecular methods for haplotype determination are very expensive and not feasible for practical use, it is necessary to develop efficient statistical and computational methods for haplotype inference from genotype data, both for unrelated individuals and for general pedigrees. In this paper, we describe a set of logic rules for haplotype inference in pedigrees. This method, along with the haplotype-elimination strategy and the PL–EM algorithm are integrated in a computer program, named HAPLORE. Although the haplotypes generated from the three aforementioned methods have different features, they all can be applied to various statistical methods for mapping disease genes (Toivonen *et al.*, 2000; Zhao *et al.*, 2000).

The set of logic rules can serve many purposes. First, it can be used to efficiently reconstruct the complete and partial haplotypes carried by each individual in the pedigrees. In fact, our studies suggest that haplotypes can be uniquely identified in all of the individuals in the pedigrees in a completely genotyped pedigree. At the same time, the missing data can be imputed and the genotype errors can be identified. Our results show that the fraction of missing data is reduced $\sim 50\%$ after this step. Second, the haplotypes constructed from the rule-based algorithm can be used in haplotype elimination and the PL-EM algorithm to reduce the computational complexity. Third, the haplotypes constructed from the rule-based algorithm can be very informative for finding the compatible haplotype configuration with a smallest set of haplotypes (Clark, 1990; Gusfield, 2001). Obviously, the haplotypes carried in completely haplotyped individuals using the rule-based algorithm must be enclosed. We apply this idea to the GAW12 data as well as the Oxford ACE data. For the GAW12 data, the identified haplotypes can provide the compatible haplotype configurations for all individuals. For the Oxford ACE data, 17 identified haplotypes are not sufficient for resolving this problem. However, 79 out of the total 82 families have compatible haplotype configurations with them. After a simple investigation, we identify that at most 20 haplotypes can generate compatible haplotype configurations for all families, which is less than the number of haplotypes having frequencies $>10^{-6}$ identified by the PL-EM and very close to the number of haplotypes with frequencies $>0.01\%$ (21 or 22 haplotypes).

After the rule-based algorithm, there may still be some inconsistency among individual haplotypes in an extended pedigree only using rules. The haplotype-elimination algorithm in HAPLORE is guaranteed to exclude all inconsistent haplotypes. The PL-EM algorithm can be applied further to estimate haplotype frequencies and identify compatible haplotype configurations with haplotypes having frequencies greater than a threshold (e.g. 10^{-6}). However, the number of compatible haplotype configurations is generally much less than that identified by the haplotype-elimination algorithm, since many haplotypes with frequencies less than a threshold (e.g. 10^{-6}) are discarded.

The complexity of haplotype algorithms varies substantially. The computational time and the computer memory of the rule-based algorithm only increase linearly rather than exponentially with pedigree size and the number of marker loci. Therefore, there is no limitation on pedigree structure, the number of loci and the number of alleles at each locus to apply the rule-based method. We test it on the GAW12 data and the Oxford ACE data. The simulated GAW12 data contains the genotypes of 1497 individuals in 23 large families expanding up to about 600 SNPs within a gene. While the Oxford ACE data comprises the genotypes of individuals in 82 small families spanning only 10 SNPs. The rule-based algorithm can completely or partially infer their haplotypes in 10 s

for both datasets. Both the haplotype-elimination algorithm and the PL-EM algorithm require the computational time that increases exponentially with the number of compatible haplotype configurations. Although the divide-conquer technique and the partition-ligation approach can speed up this process, they could not reduce the computational complexity essentially when there are many compatible haplotype configurations existing in pedigrees. Therefore, the large pedigrees in the presence of missing are generally beyond the reach of the haplotype-elimination algorithm and the PL-EM algorithm.

A crucial assumption in our analysis is that all marker loci are completely linked. That means no recombination occurs among these marker loci. Obviously, this will limit the application of our program. However, with the completion of the Human Genome Project, the availability of millions of SNPs and the technology to type these markers, many tightly linked markers will be considered in association studies, especially for candidate gene studies. It is likely that there are no recombination events or few recombination events within pedigrees for such tightly linked markers. In this context, our program will be relevant and most useful for the exploitation of linkage disequilibrium, the study of disease gene mapping, the construction of haplotype map and genetic linkage studies. We believe that there is a great need for the development of a new generation of space- and time-efficient algorithms for haplotype reconstruction with many tight linked markers in large pedigrees.

ACKNOWLEDGEMENTS

We would like to thank Shuanglin Zhang and Jinming Li for their helpful comments. We thank Dr MacCluer for providing us the simulated data from GAW12. We are grateful to the authors of reference (Keavney *et al.*, 1998) for the use of the Oxford ACE data, which is available on request from Martin Farrall (mfarrall@well.ox.ac.uk). We also thank three referees for their constructive comments. G.A.W. is supported by NIH grant GM31575 from NIGMS. This work was supported in part by grant NIH GM59507.

REFERENCES

- Akey, J., Jin, L. and Xiong, M. (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur. J. Hum. Genet.*, **9**, 291–300.
- Becker, T. and Knapp, M. (2003) Efficiency of haplotype frequency estimation when nuclear family information is included. *Hum. Hered.*, **54**, 45–53.
- Clark, A.G. (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.*, **7**, 111–112.
- Cox, R., Bouzekri, N., Martin, S., Southam, L., Hugill, A., Golamally, M., Cooper, R., Adeyemo, A., Soubrier, F., Ward, R. *et al.* (2002) Angiotensin-1-converting enzyme (ACE) plasma concentration is influenced by multiple

- ACE-linked quantitative trait nucleotides. *Hum. Mol. Genet.*, **11**, 2969–2977.
- Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. and Lander, E.S. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.*, **29**, 229–232.
- Douglas, J.A., Boehnke, M., Gillanders, E., Trent, J.M. and Gruber, S.B. (2001) Experimentally derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nat. Genet.*, **28**, 361–364.
- Du, F.X., Woodward, B.W. and Denise, S.K. (1998) Haplotype construction of sires with progeny genotypes based on an exact likelihood. *J. Dairy Sci.*, **81**, 1462–1468.
- Dudbridge, F., Koeleman, B.P.C., Todd, J.A. and Clayton, D.G. (2000) Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. *Am. J. Hum. Genet.*, **66**, 2009–2012.
- Elston, R.C. and Stewart, J. (1971) General model for genetic analysis of pedigree data. *Hum. Hered.*, **21**, 523–542.
- Excoffier, L. and Slatkin, M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, **12**, 921–927.
- Fallin, D. and Schork, N. (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation–maximization algorithm for unphased diploid genotype data. *Am. J. Hum. Genet.*, **67**, 947–959.
- Goldstein, D.B. (2001) Islands of linkage disequilibrium. *Nat. Genet.*, **29**, 109–211.
- Gusfield, D. (2001) Inference of haplotypes from samples of diploid populations: complexity and algorithms. *J. Comput. Biol.*, **8**, 305–323.
- Haines, J.L. (1992) Chromlook: an interactive program for error detection and mapping in reference linkage data. *Genomics*, **14**, 517–519.
- Hawley, M.E. and Kidd, K.K. (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J. Hered.*, **86**, 409–411.
- Hodge, S.E., Boehnke, M. and Spence, M.A. (1999) Loss of information due to ambiguous haplotyping of SNPs. *Nat. Genet.*, **21**, 360–361.
- Keavney, B., McKenzie, C.A., Connell, J.M.C., Julier, C., Ratcliffe, P.J., Sobel, E., Lathrop, M. and Farrall, M. (1998) Measured haplotype analysis of the angiotensin-I converting enzyme gene. *Hum. Mol. Genet.*, **7**, 1745–1751.
- Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.*, **22**, 139–144.
- Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. and Lander, E.S. (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.*, **58**, 1347–1363.
- Lander, E.S. and Green, P. (1987) Construction of multilocus genetic-linkage maps in humans. *Proc. Natl Acad. Sci. USA*, **84**, 2363–2367.
- Lange, K. and Boehnke, M. (1983) Extensions to pedigree analysis. V. Optimal calculation of Mendelian likelihood. *Hum. Hered.*, **33**, 291–301.
- Lange, K. and Goradia, T.M. (1987) An algorithm for automatic genotype elimination. *Am. J. Hum. Genet.*, **40**, 250–256.
- Lange, K. and Weeks, D.E. (1989) Efficient computation of LOD scores: genotype elimination, genotype redefinition, and hybrid maximum likelihood algorithms. *Ann. Hum. Genet.*, **53**, 67–83.
- Li, J. and Jiang, T. (2003) Efficient rule-based haplotyping algorithm for pedigree data. In Miller, W., Vingron, M., Istrail, S., Pevzner, P. and Waterman, M. (eds), *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB03)*. ACM, New York, pp. 197–206.
- Lin, S., Cutler, D.J., Zwick, M.E. and Chakravarti, A. (2002) Haplotype inference in random population samples. *Am. J. Hum. Genet.*, **71**, 1129–1137.
- Lin, S.L. and Speed, T.P. (1997) An algorithm for haplotype analysis. *J. Comput. Biol.*, **4**, 535–546.
- Long, J.C., Williams, R.C. and Urbanek, M. (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.*, **56**, 799–810.
- Michlataos-Beloin, S., Tishkoff, S.A., Bentley, K.L., Kidd, K.K. and Ruano, G. (1996) Molecular haplotyping of genetic markers 10 kb apart by allelic-specific long-range PCR. *Nucleic Acids Res.*, **24**, 4841–4843.
- Nejati-Javaremi, A. and Smith, C. (1996) Assigning linkage haplotypes from parent and progeny genotypes. *Genetics*, **142**, 1363–1367.
- Niu, T., Qin, Z., Xu, X. and Liu, J.S. (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **70**, 157–159.
- O’Connell, J.R. (2000) Zero-recombinant haplotyping: applications to fine mapping using SNPs. *Genet. Epidemiol.*, **19** (Suppl. 1), S64–S70.
- O’Connell, J.R. and Weeks, D.E. (1999) An optimal algorithm for automatic genotype elimination. *Am. J. Hum. Genet.*, **65**, 1733–1740.
- Patil, N., Berno, A.J., Hinds, D.A., Barrett, W.A., Doshi, J.M., Hacker, C.R., Kautzer, C.R., Lee, D.H., Marjoribanks, C., McDonough, D.P. et al. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.
- Qian, D. and Beckman, L. (2002) Minimum-recombinant haplotyping in pedigrees. *Am. J. Hum. Genet.*, **70**, 1434–1445.
- Qin, Z., Niu, T. and Liu, J. (2002) Partitioning-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **71**, 1242–1247.
- Rohde, K. and Fuerst, R. (2001) Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information. *Hum. Mutat.*, **17**, 289–295.
- Schaid, D.J. (2002) Relative efficiency of ambiguous vs. directly measured haplotype frequencies. *Genet. Epidemiol.*, **23**, 426–443.
- Sobel, E., Lange, K., O’Connell, J.R. and Weeks, D.E. (1995) Haplotype algorithms. In Speed, T.P. and Waterman, M.S. (eds) *Genetic Mapping and DNA Sequencing*. IMA Volumes in Mathematics and Its Applications. Springer, New York, pp. 89–110.
- Stephens, M., Smith, N.J. and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.
- Tapadar, P., Ghosh, S. and Majumder, P.P. (2000) Haplotyping in pedigrees via a genetic algorithm. *Hum. Hered.*, **50**, 43–56.
- Tishkoff, S.A., Pakstis, A.J., Ruano, G. and Kidd, K.K. (2000) The accuracy of statistical methods for estimation of haplotype

- frequencies: an example from the CD4 locus. *Am. J. Hum. Genet.*, **67**, 518–22.
- Toivonen,H.T.T., Onkamo,P., Vasko,K., Ollikainen,V., Sevon,P., Mannila, H., Herr,M. and Kere,J. (2000) Data mining applied to linkage disequilibrium mapping. *Am. J. Hum. Genet.*, **67**, 133–145.
- Wang,N., Akey,J.M., Zhang,K., Chakraborty,K. and Jin,L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.*, **71**, 1227–1234.
- Weeks,D.E., Sobel,E., O’Connell,J.R. and Lange,K. (1995) Computer programs for multilocus haplotyping of general pedigrees. *Am. J. Hum. Genet.*, **56**, 1506–1507.
- Wijsman,E.M. (1987) A deductive method of haplotype analysis in pedigrees. *Am. J. Hum. Genet.*, **41**, 356–373.
- Wijsman,E.M., Almasy,L., Amos,C.I., Borecki,I., Falk,C.T., King,T.M., Martinez,M.M., Meyers,D., Neuman,R., Olson,J.M. *et al.* (2001) Genetic analysis workshop 12: analysis of complex genetic traits: applications to asthma and simulated data. *Genet. Epidemiol.*, **21** (Suppl. 1), S1–S853.
- Zhang,S., Pakstis,A.J., Kidd,K.K. and Zhao,H. (2001) Comparisons of two methods for haplotype reconstruction and haplotype frequency estimates from population data. *Am. J. Hum. Genet.*, **69**, 906–912.
- Zhang,S., Zhang,K., Li,J. and Zhao,H. (2002) On a family-based haplotype pattern mining method for linkage disequilibrium mapping. *Pac. Symp. Biocomput.*, 100–111.
- Zhao,H., Zhang,S., Merikangas,K.R., Trixler,M., Wildenauer,D.B., Sun,F.Z. and Kidd,K.K. (2000) Transmission/disequilibrium tests using multiple tightly linked markers. *Am. J. Hum. Genet.*, **67**, 936–946.