European Journal of Human Genetics (2004) 12, 805–812
© 2004 Nature Publishing Group All rights reserved 1018-4813/04 $30.00
www.nature.com/ejhg

ARTICLE

Handling missing values in population data: consequences for maximum likelihood estimation of haplotype frequencies

Pierre-Antoine Gourraud*[,1], Emmanuelle Génin[2], Anne Cambon-Thomsen[1]

[1]*Unité INSERM 558-Faculté de médecine, 37 allées Jules Guesde, F-31073 Toulouse, France;*
[2]*Unité INSERM 535-Hôpital Paul Brousse, Bâtiment Leriche, BP1000, 94817 Villejuif Cedex, France*

Haplotype frequency estimation in population data is an important problem in genetics and different methods including expectation maximisation (EM) methods have been proposed. The statistical properties of EM methods have been extensively assessed for data sets with no missing values. When numerous markers and/or individuals are tested, however, it is likely that some genotypes will be missing. Thus, it is of interest to investigate the behaviour of the method in the presence of incomplete genotype observations. We propose an extension of the EM method to handle missing genotypes, and we compare it with commonly used methods (such as ignoring individuals with incomplete genotype information or treating a missing allele as any other allele). Simulations were performed, starting from data sets of haematopoietic stem cell donors genotyped at three HLA loci. We deleted some data to create incomplete genotype observations in various proportions. We then compared the haplotype frequencies obtained on these incomplete data sets using the different methods to those obtained on the complete data. We found that the method proposed here provides better estimations, both qualitatively and quantitatively, but increases the computation time required. We discuss the influence of missing values on the algorithm's efficiency and the advantages and disadvantages of deleting incomplete genotypes. We propose guidelines for missing data handling in routine analysis.

European Journal of Human Genetics (2004) **12,** 805–812. doi:10.1038/sj.ejhg.5201233
Published online 14 July 2004

Keywords: EM algorithm; HLA; haplotype; missing values; bioinformatics; linkage disequilibrium

Introduction

The numerous polymorphic genetic markers throughout the genome, the recent improvements in molecular techniques and the new possibilities of automation[1] allow the development of large genetic studies in populations. HLA population genetics data were one of the first applications of maximum likelihood (ML) estimation of haplotypes 30 years ago.[2–4] The genetic structure of the HLA region (6p21.3) is of particular interest, since it has numerous contiguous loci and a high number of alleles at many loci, generating a theoretical number of phenotypes and haplotypes greater than the usual sample size (for example, HLA-DRB1 $N = 330$ alleles).[5] Further, there is a high number of low-frequency haplotypes. The occurrence of incomplete genotypes has been reduced by the continuing improvements in HLA typing techniques. Nevertheless, when analysing large data sets such as volunteer potential haematopoietic stem cell donor Registries, the influence of missing values in haplotype frequency estimation must be addressed. Thus, haplotype estimation in large data

*Correspondence: P-A Gourraud, Unité INSERM 558-Faculté de médecine, 37 allées Jules Guesde, F-31073 Toulouse, France. Tel: +33(0)561145959; Fax: +33(0)562264240; E-mail: gourraud@cict.fr

Received 27 June 2003; revised 29 April 2004; accepted 5 May 2004

sets of contiguous loci becomes an important issue in population-based molecular genetics.

To overcome the lack of phase information provided by the techniques, likelihood-based calculations in the general framework of the Expectation Maximisation algorithm have been formalised and further developed by Dempster.[6] Many of the properties of the EM algorithm for ML estimation method have already been discussed.[7,8] These include the accuracy of the estimation of haplotype frequency, departure from Hardy–Weinberg equilibrium, the number of alleles, the number of loci, the type of markers, linkage disequilibrium measure, the influence of collapsing over a locus, computational properties and genotyping error.[9–12]

The EM algorithm is primarily set to handle the missing phase information, and it can be adapted to deal with complete and incomplete genotypes at the same time (i.e. missing phase information and missing values within a genotype).[6] We were interested in the influence of missing values on haplotype frequency estimation. In practice, missing values are usually handled in one of two ways: individuals with incomplete data are ignored (as in the EH software;[13] further referred as MVDEL), or missing values are coded as an additional allele (in the ARLEQUIN software).[14] This last 'method' is an acknowledged bug in the ARLEQUIN implementation for the estimation of haplotype frequencies. Several types of stand-alone software[15–21] propose to take into account incomplete genotypes in their analyses, but they are suitable only for specific kinds of data (biallelic markers), for specific kinds of missing data (for example recessive data) or within the framework of familial data.

Here we explore several possible solutions to this problem of missing values and look at the consequences on the estimation of haplotype frequencies by maximum likelihood methods. We implemented them in software named LOGINSERM_ESTIHAPLO.

Population and methods

Population data

The data were obtained from the French Registry of volunteer unrelated potential haematopoietic stem cell donors.[22,23] In all, 30 independent data sets of 1000 individuals were obtained by randomly drawing individuals without replacement from the main database of 85 933 individuals typed for HLA-A, -B and -DR. For each of these 30 data sets without missing values (referred to as 'initial data sets'), HLA-A, B, DR haplotype frequencies were estimated by EM and used as references to study the impact of missing data on haplotype frequency estimation.

Missing values definition

We considered a genotype to present a 'missing value' when one or zero alleles is reported at a particular locus. We assumed that the missing values were independent from the nature of the other reported polymorphisms.

Missing values simulations

Simulations were used to generate missing values ranging from 5 to 25%. In order that the missing values were randomly distributed in the data, a uniform random number was drawn for each allele in each data set. If this number was smaller than the required percentage of missing values, then the allele was deleted. Thus, one or two alleles could be missing at each locus.

Missing values handling

Two methods were compared:

1. The MVDEL method (for Missing Values Deleted) ignores any individual with missing data. If data are missing at any locus, all information is deleted for that individual. This is the method implemented in the program EH.[13]
2. The MVSAS method (for Missing Values Statistically Assessed) allows for the missing value to be any allele, which is consistent with the incomplete genotype and the haplotypes already observed in the sample. This second method was inspired by Excoffier and Dempster.[6,24] However, not all the alleles at a locus were possible. Only those already found associated with the observed alleles at the other loci in the data set were considered to substitute missing values. Indeed, the contribution of incomplete observations to the haplotype estimation is weighted by the probability of possible haplotypes in the same data set. All complete or incomplete observations are used to identify the allelic association resulting in the possible haplotype diversity. For instance, consider an observed individual with genotype (1,1) at one biallelic locus and genotype (1,?) at a second biallelic locus. In the Dempster or Excoffier approach, '?' will be replaced by 1 or 2. In our method, it will depend on the possible haplotypes that have been deduced from individuals without missing data. If haplotype 1,2 is never estimated to exist, then '?' could only be replaced by 1. This procedure is extended to all possible pattern of missing values.

EM algorithm

The estimation of haplotype frequencies by maximum likelihood within the EM algorithm has been performed as described elsewhere.[7] The Expectation Step (E-step) generally computes the likelihood of the sample using haplotype estimations of the previous iteration, or the initial values at first step (which are chosen at random; no multiple starting conditions are used). The counting procedure is extended to the presence of missing values.

The criteria to stop iteration are modified in order to compare different models.

1. *Maximisation step (M-step)*: In the M-step, haplotype frequency estimation is inspired from a gene-counting procedure.[25,26] For each genotype, the presence of a haplotype is counted through the probability of its resulting phase. We extend this procedure to incomplete genotypes in the example below, using notation for three loci, indexed by *i*, *j* and *k*. This notation generalises to any number of loci by extending the number of indices. The implementation in our software works with up to seven loci:

$$h_{ijk}^{(t+1)} = \frac{1}{2N} \times \sum_{n=1}^{N} \left[P_1^{(t)}(h_{ijk}/n) + P_2^{(t)}(h_{ijk}/n) \right]$$

where h_{ijk} is the estimation of haplotype i–j–k at iteration $t+1$; N is the number of genotypes observed; $P_1(h_{ijk}/n)$ and $P_2(h_{ijk}/n)$ are the probabilities of observing the haplotype i–j–k as first and second, respectively; $P_1(h_{ijk}/n)$ and $P_2(h_{ijk}/n)$ could be calculated as functions of haplotype estimations at iteration (t):

$$P_1(h_{ijk}/n) = \frac{\sum\limits_{h_c^{(t)}/(h_{ijk};n)} \left(2 \times h_{ijk}^{(t)} \times h_c^{(t)} - \delta_{h_{ijk};h_c} \times h_{ijk}^{(t)} \times h_c^{(t)}\right)}{\sum\limits_{(h_1,h_2)/n} \left(2 \times h_1^{(t)} \times h_2^{(t)} - \delta_{h_1;h_2} \times h_1^{(t)} \times h_2^{(t)}\right)}$$

where $P_1(h_{ijk}/n)$ is the probability of observing the i–j–k haplotype for a given genotype n; $h_c^{(t)}$ is the complementary haplotype or possible haplotypes to observe the i–j–k haplotype in genotype n; $h_1^{(t)}$ and $h_2^{(t)}$ are the pseudo-haplotype frequency estimations at iteration (t) for genotype n; (h_1,h_2) notation refers then to all possible pairs of haplotypes which may result in the observation of the genotype n, whereas $h_{ijk}^{(t)}$ is the i–j–k haplotype frequency estimated at iteration (t); $\delta_{h1,h2}$ is the Kronecker delta defined by: $\delta_{h_1;h_2} = \begin{cases} 0 \textit{ for} & h_1 \neq h_2 \\ 1 \textit{ for} & h_1 = h_2 \end{cases}$ $h_1 = h_2 \Leftrightarrow$ homozygous genotype. Such M-step is performed for all h_{ijk}; that is, all haplotype estimations are computed at each iteration. This probability is the ratio of the probability of the haplotype combination and the probability of observing such a genotype n. Although notation h_{ijk} describes the set of three locus (*i, j, k*) haplotypes as parameters in the three-locus HLA data used here, generalisation of indexes '*i, j, k*' to any number of loci is possible; h_{ijk} would refer then to the appropriate set of pseudo-haplotypes (set of haplotype compatible with given genotype). Asymptotic properties of the EM algorithm are not modified.[6]

2. *Iterations of EM*: The method usually assumes that if likelihood does not vary from more than a given very small value (say for instance 10e−4) between two iterations, estimations of haplotype frequencies are stable. Here, since we compare estimations obtained under different likelihood models, we use a direct measure of the whole stability of estimations by considering the sum of absolute errors (SAE) that is defined as follows:

$$SAE = \sum_{i=1}^{H} \left| h_i^{(t+1)} - h_i^{(t)} \right|$$

where H is the total number of haplotypes estimated and $h_i^{(t+1)}$ and $h_i^{(t)}$ are the estimations of haplotype number i at iterations t and $t+1$, respectively. No multiple starting conditions were used routinely, but convergence was assessed separately. Iterations of the algorithm were stopped when SAE reached 10e-4.

The modified EM algorithm described above for the integration of missing genotype data has been implemented in a C-written program called 'LOGINSERM_ESTIHAPLO' (available on request).

Method for comparison

A comparison of the accuracy of haplotype estimations was made using the 'IH' measure (for Identification of Haplotypes). IH takes the value of 1 if the set of estimated haplotypes is identical to the reference set of haplotypes.[7] It can be applied for all parameters estimated (all haplotype frequencies estimated) or for estimation of frequencies, which are estimated above 1/2N, where N is the sample size. 1/2N is the threshold of estimated existence of a haplotype in the sample.

$$IH = \frac{2 \times (K_{\text{ref}} - K_{\text{missed}})}{K_{\text{ref}} + K_{\text{est}}}$$

where K_{ref} is the number of parameters (frequency estimations) in the reference, K_{missed} is the number of parameters that are absent in the estimated frequencies and K_{est} is the number of parameters estimated in the haplotype estimations compared to the reference.

Haplotype inference methods on complete data can generate errors as compared to population ('true') frequencies, due to sampling errors.[16] To compare the haplotype estimation obtained on the complete data to those obtained in the presence of missing values, and to evaluate specifically the impact on the missing values on haplotype inference, we computed the difference between haplotype estimations using three classical indexes:

1. The Mean Square Error (MSE) defined as:

$$MSE = \frac{1}{H} \times \sum_{i=1}^{H} (h_i^{(\text{ref})} - h_i^{(\text{est})})^2$$

The Mean Absolute Error (MAE) defined as:

$$MAE = \frac{1}{H} \times \sum_{i=1}^{H} \left| h_i^{(\text{ref})} - h_i^{(\text{est})} \right|$$

where H is the number of parameters shared by the reference and the compared estimations. $h_i^{(\text{ref})}$ and $h_i^{(\text{est})}$ are the estimations of haplotype i frequency in the initial data set and in data with missing values, respectively.

2. The similarity index 'If'[7,12] used to measure the estimation accuracy:

$$If = \sum_{i=1}^{K_{\text{shared}}} \min\left(h_i^{(\text{ref})}; h_i^{(\text{est})}\right)$$

$$\Leftrightarrow \begin{cases} If = 1 - \dfrac{1}{2} \times \displaystyle\sum_{i=1}^{K_{\text{shared}}} \left|h_i^{(\text{ref})} - h_i^{(\text{est})}\right|; \; K_{\text{shared}} > 0 \\ If = 0; \; K_{\text{shared}} = 0 \end{cases}$$

We also introduced another measure of the accuracy through a normalised similarity index 'Ifn'. 'Ifn' considers both the number of shared haplotype estimations and the absolute error on frequencies:

$$Ifn = \left(1 - \frac{1}{2} \times \sum_{i=1}^{K_{\text{shared}}} \left|h_i^{(\text{ref})} - h_i^{(\text{est})}\right|\right) \times \frac{K_{\text{shared}}}{K_{\text{true}}}$$

where K_{shared} is the number of parameters (haplotype frequency estimations) shared by the initial data set; and K_{true} is the number of parameters (haplotype frequency estimated) in the initial data set.

Results

The different ways of handling missing values may affect the estimation of haplotype frequencies at different levels: qualitative (identification of possible haplotypes) and quantitative (their frequencies). These two levels can be considered for all possible haplotypes, or for those expected to be present in the sample. The results presented correspond to the haplotypes present in the sample. (All comparisons are available on request.)

Identification of haplotypes expected to be present in the sample according to the reference estimation

Tables 1 and 2 are built on the comparison of frequency estimates above $1/2N$,where N is the analysed sample size. The 'MVDEL' method reduces the number of haplotypes shared with the reference, in the presence of missing values (Table 1; column 'Kept' and 'Lost', rows 'MVDEL'). Deleting incomplete observations results in a decrease in the haplotypic diversity in the population, with some haplotypes being lost. The number of lost haplotypes is greater with MVDEL (Table 1; column 'Lost', rows 'MVDEL') than with MVSAS (Table 1; lumn 'Lost'). Interestingly, in the two methods, while analysing estimation above $1/2N$ frequency threshold only, no haplotype estimations were added compared to the reference (not shown). In this case, added haplotype estimation number is not forced to zero. The value of IH that summarises the conservation of haplotype estimations *vs* the reference without missing values is greater for the algorithm developed here (Table 1; Figure 1). MVSAS is therefore qualitatively better than the MVDEL method. Following the qualitative analysis of the nature of haplotypes generated through the algorithms, it is necessary to analyse the influence of handling missing values on the frequency estimation.

Frequency estimation of haplotypes

The haplotype frequencies estimated by the different methods using incomplete data are similar to those obtained on the initial sample with no missing data. Several global measures of the accuracy of these methods are presented in Table 2. This shows that the accuracy of the method developed (MVSAS) is at least as good as that of the MVDEL method. The question of the global

Table 1 Comparison of the average number of different haplotypes with frequency estimation above 1/2000, obtained according to two different ways of handling missing values

MV handling	*MV %*	*Kept*	*Lost*	*IH*
MVSAS	5	495.4[489.6; 501.1]	3.5[2.7; 4.3]	0.996[0.996; 0.997]
MVSAS	10	476.1[471.8; 480.4]	6.9[6.2; 7.7]	0.993[0.992; 0.994]
MVSAS	15	462.9[457.8; 468.1]	11.3[9.6; 13.1]	0.988[0.986; 0.990]
MVSAS	20	450.5[444.6; 456.4]	16.2[14.8; 17.6]	0.982[0.981; 0.984]
MVSAS	25	440.4[434.5; 446.4]	19.6[17.7; 21.5]	0.978[0.976; 0.980]
MVDEL	5	493.0[487.8; 498.2]	6.2[5.1; 7.3]	0.994[0.993; 0.995]
MVDEL	10	475.4[470.3; 480.5]	12.3[11.2; 13.3]	0.987[0.986; 0.988]
MVDEL	15	460.7[455.9; 465.4]	19.1[16.9; 21.4]	0.980[0.977; 0.982]
MVDEL	20	450.0[445.0; 454.9]	27.3[25.2; 29.3]	0.971[0.968; 0.973]
MVDEL	25	439.9[434.0; 445.9]	34.3[31.9; 36.7]	0.962[0.960; 0.965]

The figures represent the average number of different haplotypes for frequency estimations above 1/2000 and over 30 simulations. For each simulation, the reference is the number of haplotypes obtained by maximum likelihood estimation from a sample of 1000 genotypes randomly generated with no missing value (initial data sets). The average number of haplotypes estimated in the references depends on the percentage of missing values. 'MV%' is the percentage of missing values simulated in the data; 'Kept' is the number of haplotypes shared with the reference; 'Lost' is the number of haplotypes lost due to the presence of missing values. 'IH' is the identification index of possible haplotypes in presence of missing values in the data as regards reference (as defined in the 'Method' part); 95% Confidence intervals are given between brackets. 'MVDEL' (Missing Values DELeted) stands for the method where incomplete genotypes are deleted. 'MVSAS' (Missing Values Statistically ASsessed) stands for the method developed to handle statistically phase information in the presence of missing values through the EM algorithm.

Table 2 Comparison of the accuracy of haplotype frequency estimations in the presence of missing values in the data set, restricted to estimations above 1/2000

MV handling	*MV %*	*MSE*	*MAE*	*Ifn*
MVSAS	5	2.62E−07 [2.33E−07; 2.91E−07]	2.50E−04 [2.38E−04; 2.63E−04]	0.744 [0.734; 0.755]
MVSAS	10	3.65E−07 [3.21E−07; 4.09E-07]	3.05E−04 [2.91E−04; 3.19E−04]	0.707 [0.699; 0.715]
MVSAS	15	4.49E−07 [4.06E−07; 4.92E−07]	3.49E−04 [3.36E−04; 3.62E−04]	0.682 [0.672; 0.691]
MVSAS	20	4.90E−07 [4.57E−07; 5.23E−07]	3.85E−04 [3.74E−04; 3.97E−04]	0.659 [0.650; 0.667]
MVSAS	25	5.83E−07 [5.16E−07; 6.50E−07]	4.22E−04 [4.01E−04; 4.43E−04]	0.640 [0.630; 0.651]
MVDEL	5	2.71E−07 [2.34E−07; 3.08E−07]	2.60E−04 [2.45E−04; 2.75E−04]	0.739 [0.729; 0.750]
MVDEL	10	3.78E−07 [3.38E−07; 4.18E−07]	3.21E−04 [3.08E−04; 3.34E−04]	0.703 [0.695; 0.712]
MVDEL	15	4.31E−07 [3.95E−07; 4.67E−07]	3.60E−04 [3.46E−04; 3.75E−04]	0.677 [0.667; 0.686]
MVDEL	20	5.31E−07 [4.91E−07; 5.71E−07]	4.06E−04 [3.94E−04; 4.19E−04]	0.655 [0.647; 0.663]
MVDEL	25	5.86E−07 [5.24E−07; 6.48E−07]	4.42E−04 [4.24E−04; 4.59E−04]	0.636 [0.626; 0.646]

The figures represent several average measures of the accuracy of the haplotype frequencies estimated above 1/2000 (1/2*N*) in the presence of missing values independently simulated over 30 data sets. Estimations are compared to reference estimations obtained using the data set with no missing values and estimated above 1/2000. The threshold 1/2000 is the minimum frequency to reach for a haplotype to be truly present in the data set of 1000 genotypes. 'MV%' is the percentage of missing values simulated in the data. 'MSE' is the Mean Square Error; 'MAE' is the Mean Absolute Error, 'Ifn' is the normalised accuracy index of haplotype estimations compared to the reference as defined in the 'Population and methods' part; 95% confidence intervals are given between brackets. 'MVDEL' (Missing Values DELeted) stands for the method where incomplete genotypes are deleted. 'MVSAS' (Missing Values Statistically ASsessed) stands for the method developed to handle statistically phase information in the presence of missing values through the EM algorithm.

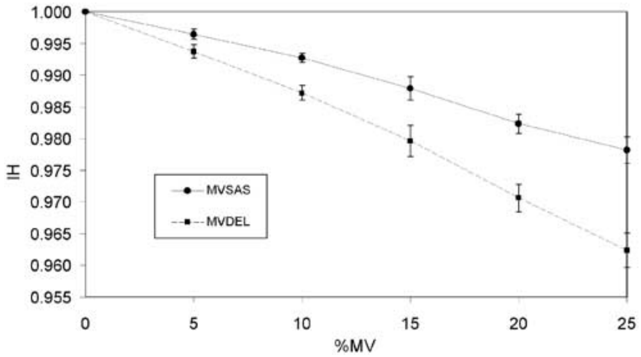

Figure 1 Comparison of the average number of possible different haplotypes through haplotype identification index (IH) according to two different ways of handling missing values and restricted to haplotypes, with estimations above 1/2*N*. The figures represent the average IH calculated with the haplotypes with frequency above 1/2000 (1/2*N*) obtained over 30 independent simulations of missing values. 'IH' is the identification index of possible haplotypes where there are missing values in the reference (as defined in 'Population and methods'). For each simulation, the reference for comparison is the haplotype estimations above 1/2000 obtained with the data set without missing values. The threshold 1/2000 is the minimum frequency to reach for a haplotype to be truly present in the data set of 1000 genotypes. The percentage of missing values (%MV) ranges from 0 to 25%; 95% confidence intervals are given by a vertical bar. 'MVDEL' (Missing Values DELeted, filled square) stands for the option where incomplete genotypes are deleted. 'MVSAS' (Missing Values Statistically Assessed, filled circle) stands for the method developed for the statistical handling of phase information where there are missing values, through the EM algorithm.

accuracy of haplotype estimation was addressed in several ways. Using squared errors, there was no apparent difference in error range in the two methods (Table 2; column MSE). Similarly, using absolute error as an evaluation of the differences between estimations, no significant global modifications were seen (Table 2; column MAE). The frequencies obtained with MVSAS seem to be closer to the reference estimations than those obtained with MVDEL. Analysis of vectorial error (not shown) shows a tendency to overestimate the haplotype frequencies, which may be due to several reasons. For example, for MVSAS, the weight of lost haplotypes is distributed over other possible haplotypes. For MVDEL, the decrease in the sample size due to deletion of observations leads to an overestimation of the frequency of the remaining ones. The Ifn global measure of the accuracy of estimations is consistent with the computed MAE (Table 2; column 'Ifn') and confirms the slight improvement of the estimations provided.

Calculation time

Obviously, the calculation cost of the MVSAS method is greater than that required when there are no missing values. This depends both on the percentage and the distribution of the missing values. From the simulations performed using Quadri Xeon 700 Mhz (cache 1 Mo; random access memory 4 Go; Operating system : Linux Red hat 7.3), the additional computation time costs approximately 1 min for 1% of missing values. On these data, the observed relationship is linear.

Discussion

Even though other methods are available (Parsimony,[27] Pseudo Bayesian[28] and Partition-Ligation Bayesian),[29] EM remains the most widely used algorithm for the estimation of haplotype frequencies. Thus, we focus only on the modification of the ML estimations provided by the EM method in the presence of missing values. We do not discuss the general properties of the estimations provided by this method as these have been discussed previously in the literature.

Although the ideal situation is to have no missing values, this is rarely the case. The use of unrelated individuals does not allow deduction of the missing values or genotyping errors. Missing values are sometimes nonrandom as they can be related to typing difficulties or to particular combinations of alleles. Such cases are addressed at the technical level as part of the quality control procedure. In the statistical handling, the assumption is made that the missing values are independent from the identity of the missing allele at the locus being considered and independent from the alleles at the other loci. This is the case, for example, for a nongenotyped locus. The importance of data validation for large data sets has been underlined,[30] as along with the consequences of genotyping error.[10]

Consequences related to the presence of missing values

The incidence of missing values in the data set modifies the information deduced on the phase information, for at least three reasons. First, computational algorithms cannot replace experimental data, thus missing information is handled in the framework of the theoretical model but remains unsolved. Secondly, it modifies the likelihood model because the parameters (ie the number of haplotypes) are different, and because the sample itself is modified. If one ignores the actual implementation of missing values handling by the software, the influence of the incomplete observations cannot be anticipated. Having incomplete observations influences the distribution of an observation over its possible phases. In ARLEQUIN,[14] missing values are considered as an additional allele at each locus. Consequently, the algorithm creates artificial haplotypes. This results in a systematic bias surrounding haplotype frequencies. In the MVDEL method, lost haplotypes may arise for two reasons: either from the missing values themselves (MVSAS method; Table 1; Column 'lost'), or from the initial decision to delete all the information about individuals with missing values. Lost (or added) estimates are expected to influence the accuracy of the estimations.

Criteria of choice for handling incomplete genotypes

The adaptation of ML estimation of haplotype frequencies to incorporate missing values slightly increases the accuracy of the estimations obtained. The MVSAS method is particularly relevant when the main interest of the study is focused on rare haplotypes. We have shown that the two methods presented here for handling missing genotypes (MVDEL and MVSAS) have different consequences on the haplotype estimates. Depending on the aim of the study and on whether one is interested in the most frequent haplotypes, a rare haplotype (disease or candidate haplotype) or the whole set of estimations for global population analysis and gametic disequilibrium measurement, one or the other methods may be best. If one is interested in common haplotypes then MVDEL may be used, since, even though haplotypes may be lost with this method, this will mainly concern rare haplotypes. If the sample size is sufficiently large, therefore, haplotype diversity is usually not affected. If, however, one is interested in rare haplotypes, then MVSAS should be used. Estimation of rare haplotype distribution remains a difficult issue. In fact, MVSAS can be adapted to any situation and it works well even if missing values are concentrated over data at a given locus. The price to pay, however, is the computation time. Indeed, the calculation cost may be prohibitive when the number of missing values, the sample size and the number of loci increase.

The main difference between MVDEL and MVSAS is attributable to missing values distribution over the sample. As underlined by Fallin and Schork,[9] the ML estimations are sensitive to sampling error. This is particularly true for the missing values sampling. Using the MVDEL method, the decrease of the sample size makes sampling errors more frequent than in the other methods and therefore results in less accurate estimations.

Ambiguities, nomenclature in the data set

Techniques sometimes give results as 'ambiguities', and from the molecular observation some of the known alleles can be discarded (for example, when the results provide a list of possible alleles and a list of absent alleles). These are not missing values but partial information, and could easily be handled using the same statistics as those presented for missing values. Depending on the complexity of ambiguities in nomenclatures (see Marsh[5] for HLA), it turns out that this theoretically simple process becomes complicated to implement. Such ambiguities might be taken into account to set the initial nature of the haplotypes (preliminary step of EM).

The methods presented here for HLA are of general relevance and can be applied to microsatellite and SNP haplotype estimation. Regarding the general properties of the method, it is all the more efficient, as the gametic disequilibrium is strong in the region.[11] Thus, the genetic structure of the region influences the statistical reassessment of missing values. It means that the gametic disequilibrium allows the deduction of a polymorphism, based on knowledge of the contiguous one in the MVSAS method. In the MVDEL method, it suggests that enough global information for phase information reconstruction remains after deleting some observations. Similarly, for polymorphic markers, missing values are expected to affect low-frequency haplotypes qualitatively, whereas high-frequency ones are affected quantitatively. In this sense, it is consistent with the general differential confidence inherent to the method on rare *vs* frequent haplotype frequencies estimation (i.e. the more frequent the haplotype, the more reliable its estimation). If missing values affect bi-allelic markers, the estimation of haplotype frequencies may essentially be quantitative, with a higher impact on low frequency and low gametic disequilibrium haplotypes. In such cases, the PLEM strategy may be an alternative method for dealing with missing values; keeping multiple outputting of possible haplotypes should be recommended, as reported.[15]

Convergence velocity and stopping criteria

We did not choose the classical likelihood stability criteria to stop iteration. Indeed, the likelihood not only depends on the values of the parameters estimated (haplotype frequencies), but also on the number of the parameters, the likelihood model and the data set retained for the analysis, which are different because of the way missing values are handled. Similar considerations were made in SNPHAP software (available at David Clayton's web site): the skimming procedure used to speed up computations modifies the likelihood model while iterating. Thus, the stability of estimators was measured directly, using the estimations by the sum of absolute variation (SAE) on the estimations from one iteration to the next.

Another choice we made here may differ from the classical ones: we only used haplotypes that have been estimated to exist in complete observations, thereby reducing the number of parameters. The alternative – inclusion of all the possible alleles – does not change the result, but increases the running time.

The M-step is the limiting one, together with the number of estimation of haplotype frequencies. Other calculations may solve the problem or may allow multi-point haplotype frequency to be computed. Although trimming procedures[31,32] have been proposed to reduce the number of parameters while iterating, the final likelihood cannot be used in log likelihood-based tests.

The evolution of the possibilities in large-scale genotyping requires the statistical treatment of the data and motivated our investigation on handling missing values for ML estimation. The statistical handling of missing values increases the quality of the haplotype frequencies provided. Deleting the incomplete observations is acceptable when using large data sets or when the estimation is computer intensive. These conclusions contribute to the enhancement of the use of haplotype estimation and allow better analysis of the data. The structure of the data influences the effectiveness of the method and puts the methodological consideration on this haplotype estimation into perspective. Indeed, the kind of polymorphism, the number of loci, the sample size, or the population may require different computational implementations.

Acknowledgements

We thank the three reviewers for their comments on this manuscript, Pr Jean-Pierre Florens for his advice and comments on this work, and acknowledge the help of France Greffe de Moelle and Dr C Raffoux for the use of HLA Registry data. This work was supported by the EU grant 'MADO': FP5-QLG7-CT-2001-00065, and Etablissement Français des Greffes Grant. The Ecole Normale Supérieure de Lyon, France, has supported PAG in his studies.

References

1 Gut IG: Automation in genotyping of single nucleotide polymorphisms. *Hum Mutat* 2001; **17**: 475–492.

2 Morton NE, Simpson SP, Lew R, Yee S: Estimation of haplotype frequencies. *Tissue Antigens* 1983; **22**: 257–262.

3 Piazza A: Haplotypes and linkage disequilibrium from three-locus phenotypes. *Histocompat Test Munksgaard* 1975; 923–927.

4 Yasuda N: Estimation of haplotype frequency and linkage disequilibrium parameter in the HLA system. *Tissue Antigens* 1978; **12**: 315–322.

5 Marsh SG: Nomenclature for factors of the HLA system, update February 2003. *Hum Immunol* 2003; **64**: 656–657.

6 Dempster AP: Maximum likelihood from incomplete data from incomplete. *J Roy Statist Soc* 1977; **39**: 921–927.

7 Excoffier L, Slatkin M: Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995; **12**: 921–927.

8 Schipper RF, D'Amaro J, Bakker JT, Bakker J, van Rood JJ, Oudshoorn M: HLA gene haplotype frequencies in bone marrow donors worldwide registries. *Hum Immunol* 1997; **52**: 54–71.

9 Fallin D, Schork NJ: Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am J Hum Genet* 2000; **67**: 947–959.

10 Kirk KM, Cardon LR: The impact of genotyping error on haplotype reconstruction and frequency estimation. *Eur J Hum Genet Oct* 2002; **10**: 616–622.

11 Xu CF, Lewis K, Cantone KL *et al*: Effectiveness of computational methods in haplotype prediction. *Hum Genet* 2002; **110**: 148–156.

12 Single RM, Meyer D, Hollenbach JA *et al*: Haplotype frequency estimation in patient populations: the effect of departures from Hardy–Weinberg proportions and collapsing over a locus in the HLA region. *Genet Epidemiol* 2002; **22**: 186–195.

13 Xie X, Ott J: Testing linkage disequilibrium between a disease gene and marker loci. *Am J Hum Genet* 1993; **53** (Suppl): 1107.

14 ARLEQUIN a program for population genetic analysis [computer program]. Version;, 1996–2002.

15 Qin ZS, Niu T, Liu JS: Partition-Ligation–Expectation-Maximization Algorithm for Haplotype Inference with Single-Nucleotide Polymorphisms. *Am J Hum Genet* 2002; **71**: 1242–1267.

16 Hawley ME, Kidd KK: HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered* 1995; **86**: 409–411.

17 Long JC, Williams RC, Urbanek M: An E–M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet* 1995; **56**: 799–810.

18 Mander AP: Haplotype analysis in population based study. *Stata J* 2001; **1**: 58–75.

19 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002; **30**: 97–101.

20 Clayton D, Jones H: Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet* 1999; **65**: 1161–1169.

21 Zhao JH, Sham PC: Faster haplotype frequency estimation using unrelated subjects. *Hum Hered* 2002; **53**: 36–41.

22 Raffoux C, Baouz A, Cozic F, Marry E: *France Greffe de Moelle: Rapport d'activité 2001*. Paris: France Greffe de Moelle, December 2001.

23 Lonjou C, Clayton J, Cambon-Thomsen A, Raffoux C: HLA -A, -B, -DR haplotype frequencies in France – implications for recruitment of potential bone marrow donors. *Transplantation* 1995; **60**: 375–383.

24 Excoffier L: Arlequin Bugs; Available at: http://lgb.unige.ch/arlequin/software/2.000/doc/buglist/buglist.html.

25 Smith CAB: Counting methods in genetical statistics. *Ann Hum Genet* 1957; **21**: 254–276.

26 Cepellini R: The estimation of gene frequencies in random mating population. *Ann Hum Genet* 1955; **20**: 97–115.

27 Clark AG: Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol* 1990; **7**: 111–122.

28 Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68**: 978–989.

29 Niu T, Qin ZS, Xu X, Liu JS: Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet* 2002; **70**: 157–169.

30 Schipper RF, Oudshoorn M, D'Amaro J *et al*: Validation of large data sets, an essential prerequisite for data analysis: an analytical survey of the Bone Marrow Donors Worldwide. *Tissue Antigens* 1996; **47**: 169–178.

31 SNPHAP [computer program]. Clayton DG, http://www-gene.cimr.cam.ac.uk/clayton/software/.

32 Thomas A: GCHap: fast MLEs for haplotype frequencies by gene counting. *Bioinformatics* 2003; **19**: 2002–2003.