THE EM ALGORITHM AND ITS IMPLEMENTATION FOR THE ESTIMATION OF FREQUENCIES OF SNP-HAPLOTYPES

JOANNA POLAŃSKA*

* Institute of Automatic Control
Silesian University of Technology
ul. Akademicka 16, 44–100 Gliwice, Poland
e-mail: `jkp@stat.rice.edu`

A haplotype analysis is becoming increasingly important in studying complex genetic diseases. Various algorithms and specialized computer software have been developed to statistically estimate haplotype frequencies from marker phenotypes in unrelated individuals. However, currently there are very few empirical reports on the performance of the methods for the recovery of haplotype frequencies. One of the most widely used methods of haplotype reconstruction is the Maximum Likelihood method, employing the Expectation-Maximization (EM) algorithm. The aim of this study is to explore the variability of the EM estimates of the haplotype frequency for real data. We analyzed haplotypes at the BLM, WRN, RECQL and ATM genes with 8–14 biallelic markers per gene in 300 individuals. We also re-analyzed the data presented by Mano *et al.* (2002). We studied the convergence speed, the shape of the loglikelihood hypersurface, and the existence of local maxima, as well as their relations with heterozygosity, the linkage disequilibrium and departures from the Hardy-Weinberg equilibrium. Our study contributes to determining practical values for algorithm sensitivities.

Keywords: algorithms, haplotypes, likelihood functions, gene frequency

1. Introduction

Much of recent research in clinical genetics relies on resolving the genetic structure of complex diseases (traits). Complex genetic traits are linked with DNA loci located in multiple regions in the genome. Eventually, the studies will allow associating risks for complex diseases with sets of specified haplotypes. Problems to be solved to achieve this aim stem from (a) the necessity to carry out large population-based studies and to collect large amounts of data, and (b) the necessity of developing robust and efficient numerical algorithms for haplotype reconstruction from unphased genotypes.

In this paper we are concerned with the latter problem, i.e., haplotype reconstruction from unphased genotype data. The first practical approach to solve this problem was a parsimony-type method developed by Clark (1990). The necessity for a better use of the information contained in the collected samples led to the increased interest in maximum likelihood estimates of the haplotype structure. However, the likelihood function associated with the samples of unphased genotypes with the underlying haplotype structure is complicated and cannot be maximized by the standard techniques. A breakthrough was the application of the Expectation Maximization (EM) method (Dempster *et al.*, 1977) to maximize the likelihood of the observed genotype data (Excoffier and Slatkin, 1995; Hawley and Kidd, 1995; Long *et al.*, 1995). Since the publication of the EM algorithm for haplotype discovery, further studies have appeared discussing this method and its properties. Fallin and Schork (2000) presented the results of their research on the accuracy of haplotype frequency estimation as a function of a number of factors, including the sample size, the number of loci studied, allele frequencies, and locus-specific allelic departures from Hardy-Weinberg and linkage equilibria. Via extensive simulation studies, they demonstrated that the haplotype frequency estimation for biallelic diploid genotype samples by using the EM algorithm performs very well under a wide range of population and data-set scenarios. They concluded that much of the overall error is due to sampling, rather than to algorithmic and estimation problems or inaccuracies. Their conclusion that the accuracy of the frequency estimation of rare haplotypes mainly depends on the proper sampling procedure was confirmed by Clark *et al.* (2001), who applied the

EM algorithm to the inference of the CCR2-CCR5 haplotypes in the CEPH families. The accuracy of the haplotype frequency estimation performed using the EM algorithm was also studied by Tishkoff *et al.* (2000), who compared the results of the EM algorithm with the algorithm based on the counting of the phase-known gametes. They found that only the frequencies of rare haplotypes might be wrongly estimated and suggested that applying the molecular haplotyping method when obtaining highly accurate estimates of the frequencies of rare alleles is necessary. McKeigue (2000) reported that for the estimation of two-locus haplotype frequencies, the above-mentioned two strategies did not differ significantly in terms of the information obtained for a given genotyping workload, if the assumption of the Hardy-Weinberg equilibrium holds true. The problem of the sensitivity to the departures from the Hardy-Wienberg equilibrium was discussed by Rohde and Fuerst (2001) and Single *et al.* (2002) where the authors tested the properties of the EM algorithm on the loci which were closely linked and significantly deviating from the Hardy-Weinberg equilibrium. Their conclusions did not differ much from those formulated by Fallin and Schork (2000).

Since the first publication (Excoffier and Slatkin, 1995), a number of articles proposing some improvements in the EM algorithm have appeared (Long *et al.*, 1995; Hawley and Kidd, 1995; Chiano and Clayton, 1998; Rohde and Fuerst, 2001). Some authors focused on other variants of the EM algorithm to solve more specific tasks (Slatkin and Excoffier, 1996; Ghosh and Majumber, 2000; Kalinowski and Hedrick, 2001). Extensive studies on the framework for haplotype inference other than the parsimony and EM resulted in a new idea proposed by Stephens *et al.* (2001a) known as the PHASE algorithm. Their Bayesian method exploits ideas from population genetics and coalescent theory to make predictions about the patterns of haplotypes to be expected in natural populations. Several comparative studies concerning the EM algorithm and the PHASE algorithm of Stephens *et al.* (2001a) were published (Stephens *et al.*, 2001a; Xu *et al.*, 2001; Zhang *et al.*, 2001; Stephens *et al.*, 2001b). All authors agree that generally both algorithms give the same level of the accuracy of the haplotype frequency estimates. According to the authors of the PHASE algorithm, it overperforms other existing methods (i.e. parsimony and EM) when there is “clustering” in the true haplotype configuration. The observation of haplotype blocks in sequenced DNA (Patil *et al.*, 2001) and in coalescent simulations (Wang *et al.*, 2002) led to further modifications of both algorithms. Niu *et al.* (2002) proposed a partition-ligation strategy to investigate the haplotype block structure. The resulting methodology, named EM-PL, constitutes an improvement in both accuracy and capacity in comparison with the standard, previously existing algorithms (Qin *et al.*, 2002). The partition-ligation idea was also used in the PHASE algorithm (Lin *et al.*, 2002) improving its performance for the data with a large number of loci.

Due to the interest in haplotype blocks, coming from the abundance of SNP data, the problem of the accuracy and reliability of haplotype reconstruction methods gained great importance. Despite many studies on the properties of the EM algorithm, several problems related to its application are still unsolved. Among the most important ones are: determining the speed of convergence, the sensitivity to the stopping criterion and the existence of multiple local maxima.

In this paper we explain the basic structure of the EM algorithm and discuss the above-mentioned properties with the focus on practical applications. We developed a Matlab-based implementation of the EM method. Using our program we study and illustrate several aspects of the EM application: the speed of convergence, the reliability of estimates, and the existence of multiple solutions. We analyze the data set, recently presented by Mano *et al.* (2002), leading to a nonunique solution, and we show the existence of a local maximum, not found by Mano *et al.* (2002). We also use a real data set of the observed genotypes at the BLM, WRN, RecQL and ATM genomic regions with 8–14 biallelic markers per gene in individuals from four ethnic groups (Bonnen *et al.*, 2000; Trikka *et al.*, 2002) to estimate practical convergence rates, the computational complexity and the sensitivity of estimates to parameters. We demonstrate that convergence and complexity depend on the number of loci (data size) and observed variations in phenotypes (data structure).

2. Data

In our study we used the following three data sets:

- *Data set #1*: ATM region.

The data set consists of 14 biallelic neutral-sequence variants that span 142 kbp of the ATM region. These ATM intronic single-nucleotide polymorphisms (SNPs) were genotyped in 183 DNA samples from individuals of four different ethnic origins: African American, Asian American, white European American, and Hispanic. These samples were part of the collection analyzed in (Bonnen *et al.*, 2000). The detailed information on PCR and sequencing primers, PCR amplification of genomic DNA, DNA sequencing, and allele-specific oligonucleotide (ASO) hybridizations can be found in the original paper.

- *Data set #2*: BLM, WRN, RecQL regions.

The second data set consists of 8 SNPs identified within 154 kbp of the BLM, 12 SNPs within 186 kbp of WRN and 11 SNPs within the 180 kbp of RecQL region. These noncoding SNPs were genotyped in 300, 309, and 310 DNA samples for BLM, RecQL, and WRN regions, respectively. The general collection included samples from four ethnic groups: African American, Asian, Caucasian, and Hispanic. That data set was used for cancer association studies and is described in detail in (Trikka *et al.*, 2002).

- *Data set #3*.

The third data set was presented in (Mano *et al.*, 2002). It consists of 16 three-locus DNA samples and is shown in Table 1.

Table 1. Data set proposed in (Mano *et al.*, 2002).

Genotype	Number of cases
TTT TGT	4
TGT GCT	6
TGT CCT	2
TGT CCG	2
TCT CCG	2

3. Problem Formulation

With the discovery of the polymerase chain reaction (PCR) going from genomic DNA to sequence data has highly accelerated. The direct sequencing of the PCR product for heterozygous diploids results in the amplification of both alleles and does not allow us to resolve the amplification products, which produces a vast number of possible haplotypes. The more of such “ambiguous” sites in an individual, the more haplotypes possible. The data set is comprised of a number of polymorphic loci observed in a sample of individuals. Let us call a multilocus genotype whose haplotypic phase is unknown the phenotype (unphased genotype). A multilocus genotype defined as a particular combination of two multilocus haplotypes will be called a genotype hereafter. The number of genotypes (c_j) leading to the j-th phenotype is a function of the number of heterozygous loci s_j:

$$c_j = \begin{cases} 2^{s_j-1} & \text{if} \quad s_j > 0, \\ 1 & \text{if} \quad s_j = 0. \end{cases} \tag{1}$$

Example 1. Assume the following unphased genotype data ($s_j = 3$):

$$TGTC^{GCA}_{TGC}G.$$

The list of all possible phased genotypes is shown below ($c_j = 2^2 = 4$):

$$\begin{array}{c|c} TGTCGCAG & TGTCGGAG \\ TGTCTGCG & TGTCTCCG \\ \hline TGTCTCAG & TGTCTGAG \\ TGTCGGCG & TGTCGCCG \end{array}$$

♦

The answer to the question ‘Which of those c_j genotypes is the proper one?’ cannot be found without additional studies. It is possible to solve this problem by using genealogical information in families, but then some members of the families should be omitted in further studies. This is by no means satisfying because of the additional cost, especially when large samples are analyzed. The goal is to find the best (in some sense) estimates of the haplotype frequencies in the population using only limited information included in the unphased genotype sample data.

4. Maximum Likelihood Estimates

Clark (1990) introduced an algorithm to infer the haplotypes from such population samples. The principle is to start by examining complete homozygotes and single-site heterozygotes. Then other individuals are screened for a possible occurrence of previously recognized haplotypes. For each positive identification, the complementary haplotype is added to the list of the recognized haplotypes, and so forth. The weak points of the above algorithm are as follows: (a) homozygotes are not always present; for example, it happens very often in a small sample study; (b) the final result depends on the order in which the individuals were listed; (c) the information in the sample is not fully used.

Example 2. For the *Data set #3* there exist 145 possible algorithm paths leading to 50 different configurations. Choosing the path $hap1 \rightarrow hap2 \rightarrow hap3 \rightarrow hap5 \rightarrow hap4$ one can reach the solution called ‘Configuration 1’ in Table 2, while the path $hap1 \rightarrow hap2 \rightarrow hap4 \rightarrow hap3 \rightarrow hap5$ leads to ‘Configuration 2’. ♦

Haplotypes can be inferred and their frequencies can be estimated via a maximum likelihood approach. Under the assumption of the Hardy-Weinberg equilibrium and random mating, the probability P_j of the j-th phenotype is given by the sum of the probabilities of each of the possible genotypes:

$$P_j = \sum_{i=1}^{c_j} P(genotype\ \ i) = \sum_{i=1}^{c_j} P(h_k h_l), \tag{2}$$

Table 2. Numerical complexity of the problem of inferring the haplotype frequencies.

Gene name	No. of SNP loci	Sample size	No. of observed phenotypes	No. of feasible haplotypes	No. of feasible genotypes
BLM	8	300	112	256	1124
WRN	12	310	116	498	954
RecQL	11	309	96	1323	4015
ATM	14	183	45	4885	6701
Mano's data	3	16	5	11	11

where $P(h_k h_l)$ is the probability that the i-th genotype is composed of haplotypes k and l:

$$P(h_k h_l) = \begin{cases} p_k^2 & \text{if} \quad k = l, \\ 2p_k p_l & \text{if} \quad k \neq l, \end{cases} \tag{3}$$

and p_i denotes the frequency of the i-th haplotype.

The probability of a sample of n individuals, conditional on the phenotype frequencies $P_1, P_2, \dots, P_m$, is given by the multinomial expression

$$P = \frac{n!}{n_1! n_2! \cdots n_m!} \times P_1^{n_1} \times P_2^{n_2} \times \cdots \times P_m^{n_m}, \tag{4}$$

where m denotes the total number of phenotypes and n_j is the number of individuals carrying the j-th phenotype:

$$\sum_{j=1}^{m} n_j = n.$$

Therefore the likelihood of the haplotype frequencies given phenotypic counts is

$$L(p_1, p_2, \dots, p_h) = a_1 \prod_{j=1}^{m} \left(\sum_{i=1}^{c_j} P(h_{ik} h_{il}) \right)^{n_j}. \tag{5}$$

The maximum likelihood estimates of haplotype frequencies could, in principle, be found analytically or numerically by solving a set of equations resulting from the $h-1$ partial derivatives equated to 0:

$$\frac{\partial \log L}{\partial p_t} = \sum_{j=1}^{m} \frac{n_j}{P_j} \frac{\partial P_j}{\partial p_t}, \quad t = 1, 2, \dots, h-1. \tag{6}$$

However, the nonlinearity of (6) and a large number of equations to be solved when practical data are analyzed make this approach prohibitive.

5. EM Algorithm

The number of haplotypes is most often unknown, so that the analytical solution cannot be found by using (4)–(6). Even if it is known, the problem becomes numerically intractable for large h, as has been mentioned before.

A method of overcoming these difficulties is the application of the expectation maximization (EM) algorithm. The EM algorithm was formalized by Dempster *et al.* (1977), but its application to the problem of inferring haplotype frequencies was formulated almost simultaneously by several authors (Excoffier and Slatkin, 1995; Long *et al.*, 1995; Hawley and Kidd, 1995). For a detailed description of the algorithm we refer the reader to their articles. The following presents an outline of the algorithm.

The EM algorithm is an iterative method of computing sets of haplotype frequencies $p_1, p_2, \dots, p_h$ starting with arbitrary initial values $p_1^{(0)}, p_2^{(0)}, \dots, p_h^{(0)}$. These initial values are used to estimate genotype frequencies $\widetilde{P}(h_k h_l)$ as if they were the unknown true frequencies (the expectation step). These expected genotype frequencies $\widetilde{P}(h_k h_l)$ are standardized and used, in turn, to estimate haplotype frequencies $\widehat{p}$ at the next iteration (the maximization step). In the iteration of the algorithm, we have

- The expectation step

$$\widetilde{P}(h_k h_l)^{(g)} = \begin{cases} p_k^{(g)2} & \text{if} \quad k = l, \\ 2p_k^{(g)} p_l^{(g)} & \text{if} \quad k \neq l. \end{cases} \tag{7}$$

- The maximization step

$$P_j^{(g)} = \sum_{i=1}^{c_j} P(genotype \;\; i)^{(g)} = \sum_{i=1}^{c_j} \widetilde{P}(h_k h_l)^{(g)}, \tag{8}$$

$$P(h_k h_l)^{(g)} = \frac{n_j}{n} \frac{\widetilde{P}(h_k h_l)^{(g)}}{P_j^{(g)}}, \tag{9}$$

$$\widehat{p}_t^{(g+1)} = \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{c_j} \delta_{it} P_j (h_k h_l)^{(g)}, \tag{10}$$

where δ_{it} is an indicator variable equal to the number of times haplotype t is present in genotype i (0, 1, or 2). The flow chart depicting the implementation of the steps (7)–(10) is shown in Fig. 1. The stopping (convergence) criterion is defined as the absolute value of the difference between the consecutive ML function values being less than an arbitrary parameter $\varepsilon > 0$.

Data Preprocessing. The probabilities appearing in (7)–(10) are indexed by both haplotype and genotype numbers. Therefore, for a practical implementation of the algorithm, a data preprocessing step is necessary. In this

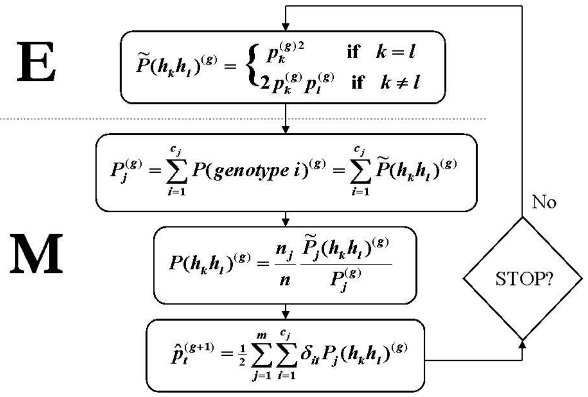
E
$\widetilde{P}(h_k h_l)^{(g)} = \begin{cases} p_k^{(g)2} & \text{if } k=l \\ 2p_k^{(g)} p_l^{(g)} & \text{if } k \neq l \end{cases}$
$P_j^{(g)} = \sum_{i=1}^{c_j} P(genotype\ i)^{(g)} = \sum_{i=1}^{c_j} \widetilde{P}(h_k h_l)^{(g)}$
M
$P(h_k h_l)^{(g)} = \frac{n_j}{n} \frac{\widetilde{P}_j(h_k h_l)^{(g)}}{P_j^{(g)}}$
$\hat{p}_t^{(g+1)} = \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{c_j} \delta_{it} P_j(h_k h_l)^{(g)}$
No
STOP?

Fig. 1. Flow chart of the EM algorithm.

step, given the observed phenotypes, all feasible genotypes and haplotypes are constructed and indices are assigned to all feasible genotypes and haplotypes.

The initial conditions for the EM algorithm. The sensitivity to the initial conditions is a known property of nonlinear algorithms. Generally, there are several possibilities of initializing the haplotype frequencies $p_1^{(0)}, p_2^{(0)}, \ldots, p_h^{(0)}$. They can be summarized as follows:

- All haplotypes are equally likely,

$$p_t^{(0)} = \frac{1}{n_h}, \quad t = 1, 2, \ldots, n_h. \tag{11}$$

- All possible genotypes for each phenotype are equally likely,

$$P_j(h_k h_l)^{(0)} = \frac{1}{c_j}, \quad j = 1, 2, \ldots, m. \tag{12}$$

- Initial haplotype frequencies are chosen at random.
- All initial haplotype frequencies are equal to the product of the corresponding single-locus allele frequencies (i.e., a complete linkage equilibrium).
- The input data influence the initial haplotype frequencies.

Since in practical applications several iterations of the EM algorithm should be performed to avoid reaching a local maximum, randomized initial haplotype frequencies are recommended, although the idea of composing the deterministic values depending on the number of possible haplotypes or genotypes with a randomized additive perturbation might work very well too.

6. Practical Application of the EM Method

We implemented the EM algorithm for haplotype reconstruction, as has been presented in the previous section, using the Matlab programming environment. Our implementation allows us to rerun the program automatically with different random or deterministic initial values. During the iterations all the values of the haplotype frequencies and the likelihood function are stored. An extra script allows us to summarize the results of all simulations and to compare the values of the maximum likelihood functions and the obtained solutions to the problem. Partial or summary results can be presented graphically. To test the reliability of our implementation, we compared the results of our calculations with the haplotype frequencies found by Arlequin 2.0 (Schneider *et al.*, 2000) and RIGHT 1.0 (Mano *et al.*, 2002) for all three data sets. The estimated haplotype frequencies did not differ by more than 0.1% for the stopping criterion $\varepsilon = 10^{-8}$ and the maximum number of iterations for the Arlequin program equal to 5000. Below, we show the results of our extensive simulation studies regarding the EM algorithm convergence, numerical complexity, and the existence of local maxima. We report several facts which are important for practical applications of EM to haplotype reconstruction. We show an example of genotypic data, which lead to multiple maxima of the likelihood surface.

6.1. Properties of the Data Sets Used for the Simulation Studies

Two of our three data sets used in the studies were real DNA samples collected for disease association studies. Testing the deviation from the EM algorithm assumptions is necessary for a proper evaluation of the final results of our work. The basic assumptions, which are of great importance for the haplotype reconstruction method, are the lack of departures from the Hardy-Weinberg equilibrium and occurrences of free recombination events.

The Hardy-Weinberg Equilibrium and Loci Heterozygosity. Tests for deviations from the Hardy-Weinberg equilibrium (HWE) were conducted for each locus of SNP genotypes of all data sets (Schneider *et al.*, 2000). Within ATM and RecQL, none of the SNPs deviated significantly from HWE. Discrepancies were observed at two SNPs on BLM: B18.1 ($p = 0.029$) and B22 (with significance $p = 0.032$). Some departures were also noticed within WRN: W1 ($p = 0.030$) and W18.2 ($p = 0.036$). A detailed analysis of the departures from HWE within BLM, RecQL, and WRN made for each locus-population combination is presented in (Trikka *et al.*, 2002). As will be demonstrated further on (cf. the subsection on numerical complexity), the deficiency or excess heterozygosity with respect to the HWE is related to the numerical complexity of the EM algorithm and the structure of the likelihood hypersurface. Among the 14 SNPs within the ATM locus, the smallest heterozygosity was observed for the 6-th consequent SNP named IVS46-257a→c, and was equal to 0.130. The maximum observed value was equal to 0.494 for two SNPs numbered 7 and 13 (coded IVS55+186c→t and IVS62-973a→c, respectively). The average heterozygosity for all SNPs was 0.374. Within the BLM region the heterozygosity pattern looked very similar to that of the ATM region. We did not notice extreme differences in heterozygosity among the eight SNPs analyzed. The smallest value of H was 0.196 (B4.2 coded as IVS1-20290g→a), the highest was 0.451 (B22 coded as IVS22+9303c→t), and the average was 0.352. The other regions showed two clusters of SNPs differing significantly in heterozygosity. The WRN region, consisting of ten SNPs had four of them (W1 IVS1-8213g→a; W23 IVS35+11737g→c; W26.1 IVS53+30673c→t; and W26.2 IVS35+30764c→a) almost homozygous (heterozygosity 0.065, 0.021, 0.014, and 0.018, respectively), while the other six demonstrated the heterozygosity varying from 0.229 to 0.520. The average heterozygosity for that genomic region was 0.238, but for the interior of the region it was highly heterozygous and equal to 0.376. The highest heterozygosity was observed at RecQL, where the heterozygosity no higher than 0.1 was observed only at two flanking sites. All the other SNPs were highly heterozygous and the average was equal to 0.4007. This differentiation in the heterozygosity structure influences the shape of the likelihood hypersurface and the existence of saddle-like or local maximum points. It was pointed out by other authors (Fallin and Schork, 2000) that an excess of homozygosity could, in effect, decrease the amount of ambiguous phase information in a data set and, as such, improve the estimation accuracy. These results allow assuming that the observed discrepancies in HWE result from a hidden substructure of the data set and can be disregarded.

One of the loci in *Data set #3* deviated significantly from HWE (locus 2, $p = 0.014$). We observed an excess of heterozygosity for the deviating locus, but there is no pattern of excessive heterozygosity within the pooled data. The comparative studies by other authors (Fallin and Schork, 2000; Niu *et al.*, 2002; Single *et al.*, 2002) showed the robustness of the haplotype frequency estimation toward the HWE deviations, so it was assumed in further studies that such deviations will not result in a significant differentiation in the haplotype frequency estimation.

Recombination Events and a Linkage Disequilibrium. The EM algorithm was developed assuming free recombination events within the genomic region. A detailed investigation of the ATM region is presented by Trikka *et al.* (2002). The small number of haplotypes seen suggested the possibility that recombination is reduced at the ATM locus. The four-gamete test (Hudson and Kaplan, 1985) allowed the authors to conclude that the ATM locus exhibits a reduced recombination in all four ethnic groups. To better understand the background of this observation, we measured the disequilibrium by using the likelihood ratio test proposed by Slatkin and Excoffier (1996), and implemented in the Arlequin package. The majority of pairs are in a high disequilibrium. The lack of recombination or the reduced number of recombination events can be partly explained by an extremely high disequilibrium observed within those loci. The analyses made by Bonnen *et al.* (2000) demonstrate that the recombination events seem to have occurred throughout the BLM and WRN genomic regions. The disequilibrium tests performed by us confirmed the results obtained by Bonnen *et al.* (2000). For RecQL, a linkage disequilibrium was showed throughout the gene, with the exception of the outermost markers. Within the BLM gene no such significant disequilibrium was observed for all loci, but some of them (loci 5 – B20, 6 – B21.1, and 7 – B22) were in a high disequilibrium with all other loci. A significant disequilibrium was also noticed between neighboring loci throughout the BLM gene. We observed a high disequilibrium within the interior of the WRN gene (between W12.1 and W22 loci), which is generally coherent with the results of Bonnen *et al.* (2000). The test performed by Bonnen *et al.* (2000) showed that the recombination is absent within the middle part of RecQL and is consistent with the increased linkage disequilibrium observed in the same area. They noticed the occurrence of the recombination events within the WRN gene. They conclude that these are probably of a recent origin. Taking under consideration all the above, we presume that the low recombination rate could be explained by an extensive linkage disequilibrium, which does not impact highly on the common haplotype frequency estimates (Fallin and Schork, 2000).

Numerical Complexity. As was mentioned above, it is necessary to prepare lists of feasible haplotypes and genotypes before launching the EM algorithm. The final numbers depend not only on the number of loci within the genomic region, but also on the level of their heterozygosity. Given the ATM gene with its 14 loci and a relatively small number of the observed different phenotypes (45), the number of feasible haplotypes increases to 4885, which gives the number of feasible genotypes equal to 6701. At the 8-th locus BLM gene, demonstrating the average heterozygosity equal to 0.352, we noticed 112 different phenotypes leading to 256 possible haplotypes and 1124 possible genotypes. The RecQL and WRN genes, having almost the same numbers of loci (11 versus 12), demonstrate different levels of numerical complexity. The number of the observed phenotypes inside the WRN sample was 116 with 498 feasible haplotypes and 954 feasible genotypes. But for the RecQL gene those numbers were 96, 1323, and 4015, respectively. Such a large difference can be partly explained by varying heterozygosity within those regions. The WRN demonstrates significantly lower heterozygosity (average 0.238, with four loci being almost homologous, $-H < 0.065$) than RecQL does (average 0.401 for all loci, and 0.474 without the outermost, homozygous loci). It might influence both the speed of the convergence and the monotonicity of the likelihood surface. Table 2 presents a summary of the above analyses.

6.2. Convergence Speed the EM Algorithm

Since the introduction of the EM algorithm by Dempster *et al.* (1977) a lot of works have been published regarding its numerical properties. Their authors mainly focus on the convergence of the algorithm and modifications, the purpose of which is to speed it up. The majority of these studies are carried up by mathematicians and the best reference source is the review by Meng and van Dyke (1997) or the book by McLachlan and Thriyambakam (1997). A large number of articles cited there are devoted to the general EM algorithm and do not concern the specific form introduced by Slatkin and Excoffier (1995). The basic study on the convergence of the EM algorithm for the estimation of haplotype frequencies was published by Fallin and Schork (2000). Using numerical simulations they noticed that the algorithm converges relatively close to the maximum in less than 50 iterations. Their simulations showed that the convergence criterion influences the estimated haplotype frequencies to some limits. Increasing the accuracy beyond 10^{-8} did not significantly change the estimates of the haplotype frequencies. These simulations were performed using data simulated by other software. Our goal was to check the speed of the algorithm convergence applied to “real” data, which exhibit some deviations from the model. We launched 100 runs of our implementation of the EM algorithm for each gene data set, tracing the variability of the loglikelihood function and the estimates of the haplotype frequencies. We observed that on average in less than 20 iterations the algorithm reached its convergence point with the accuracy defined by the criterion that the increase in loglikelihood function be less than 0.01. The distance of the estimates of the haplotype frequencies from their convergence point (in the sense of the Euclidean norm) was sufficiently small in only few steps of the algorithm. Figure 2 presents a typical profile of the loglikelihood function changes during one run of the algorithm, while Fig. 3 demonstrates the distance of the frequency estimates from the convergence point for consecutive algorithm iterations. Increasing the accuracy by setting the stopping criterion equal to 10^{-5} results in an increase in the average number of iterations to 80. The results of our studies performed on genetic data confirm the relatively high convergence speed of the EM algorithm when applied to the problem of the estimating haplotype frequencies.

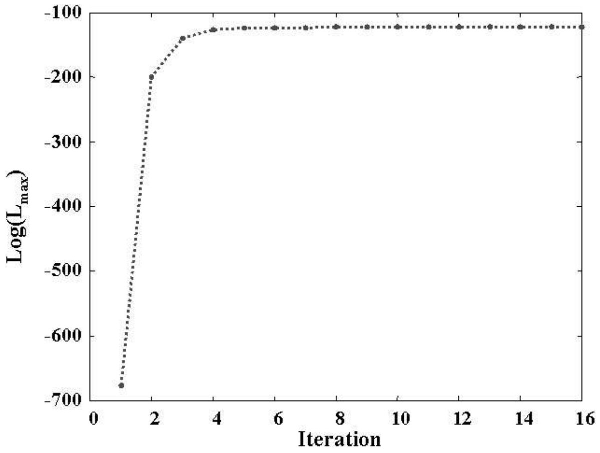

Fig. 2. Convergence of the EM algorithm.

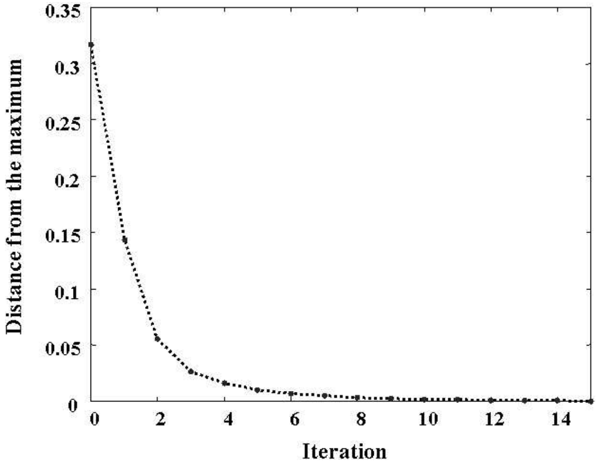

Fig. 3. Euclidean distance from the maximum for the subsequent iterations of the EM algorithm.

6.3. Multiple Local Maxima

Iterations of the EM algorithm, as defined by Dempster *et al.* (1977), always lead to nondecreasing values of the likelihood. However, there is no proof of the uniqueness of a likelihood function maximum. Likelihood hypersurfaces may have multiple local maxima. The problem of the existence of local maxima has been noticed by several authors (see *Discussion part* in Dempster *et al.*, 1977; Wu, 1983), but there are no studies concerning the details of that issue.

To evaluate the global maximum values of the loglikelihood function for each gene data set, we carried out

250,000 runs of the EM algorithm for haplotype reconstruction, using the Arlequin software (Schneider *et al.*, 2000), with randomized initial conditions. These values and the estimates of the haplotype frequencies obtained in the simulations (data not presented here) served as a reference in further studies. In the next step of our experiment, we compared these reference maxima with the results of 100 runs of our algorithm implementation with the default value of the stopping criterion equal to 0.01. We considered the accuracy of reaching the global maximum satisfactory, if the distance from the global maximum (in the sense of the Euclidean norm) to the convergence point reached by the algorithm was less than or equal to 0.005. A global maximum is characterized by the value of the loglikelihood function, up to an additive constant. In the case of the BLM gene, this value is equal to -122.08. Four out of 100 runs of the algorithm for the BLM gene reached their convergence points outside the defined neighborhood of the global maximum. These values were equal to -125.99, -124.09, -123.11, and -122.42. Despite noticeable differences in the loglikelihood function values, there were no significant variations in the estimates of the haplotype frequencies. The frequency estimates of the most common haplotypes did not differ by more than 3.5% from the values for the global maximum. The estimated frequencies for the common haplotypes were equal to 0.2114, 0.1691, 0.1079, 0.0839, and 0.0728, while their values for the global maximum were respectively equal to 0.2146, 0.1684, 0.1116, 0.0849, and 0.0745. Decreasing the stopping condition to 10^{-5} resulted neither in significant changes in the value of the loglikelihood function, nor in the estimates of the haplotype frequencies. After 67 additional iterations, the value of the loglikelihood function for the convergence point with the lowest likelihood increased from -125.99 to -125.91 without a significant improvement in the accuracy of the estimates. The new estimates were 0.2128, 0.1685, 0.1080, 0.0841, and 0.0728, respectively. The distance from the global maximum decreased from 0.0207 after 15 iterations to 0.0202 after 82 iterations.

We performed similar experiments for the WRN and ATM genes. All runs of the EM algorithm converged to the close neighborhood of the global maximum in about 15 steps. We did not observe irregularities around the global maximum of the type observed for the BLM. The hypersurface seems to be close to a negative definite quadratic form in the vicinity of the global maximum.

To explore the above phenomena, we launched the EM algorithm implemented in the Arlequin package several times with various numbers of randomized initial conditions. Table 3 presents a summary of these simulations. The stopping criterion is the difference in the sum of the haplotypic frequency change between two successive iterations being less than an arbitrary value ε. The high linkage disequilibrium within the ATM genomic region together with relatively low heterozygosity may explain the shape of the loglikelihood hypersurface and a small number of the observed irregularities. A lower linkage disequilibrium and higher heterozygosity only inside the WRN and RecQL genes may result in a less regular vicinity of the global maximum of the loglikelihood hypersurface.

Table 3. Number of different convergence points found in n runs of the EM algorithm with randomized initial values of the haplotype frequencies.

Gene	Number of algorithm runs					
name	50	100	500	1,000	5,000	25,000
Stopping criterion $\varepsilon = 10^{-5}$						
BLM	43	84	250	343	588	765
WRN	46	87	298	425	876	1441
RecQL	28	44	80	100	182	351
ATM	47	79	228	286	482	775
Stopping criterion $\varepsilon = 10^{-8}$						
BLM	14	19	25	29	30	35
WRN	17	21	29	33	39	65
RecQL	4	5	8	10	22	37
ATM	5	6	9	9	10	8
Stopping criterion $\varepsilon = 10^{-12}$						
BLM	12	14	20	17	26	27
WRN	4	4	4	4	4	4
RecQL	2	2	3	3	7	11
ATM	1	1	1	1	1	1

Another aspect of the problem of the hypersurface structure, focusing on multiple global maxima, was discussed by Mano *et al.* (2002). Below we recall the data set proposed in (Mano *et al.*, 2002) and we show the results of our analysis:

Example 3. The data set consists of unphased 3-locus genotypes of 16 individuals (shown in Table 1). As regards the number of loci and their variability, one notices that there are 11 possible haplotypes. ♦

In (Mano *et al.*, 2002), it was found that the likelihood surface had two separate global maxima, corresponding to different configurations of haplotypes, with the value of the logarithm of the likelihood function equal to -19.418 for each of them. The estimates of the haplotype frequencies for each configuration are presented in Table 4. In our experiments we found one more local maximum, with the logarithm of the likelihood function value equal to -23.975, to which the algorithm may converge for some initial conditions. The local maximum configuration is shown in Table 5. Table 6 shows the examples of initial values $p_1^{(0)}, p_2^{(0)}, \dots, p_h^{(0)}$, for which the algorithm reaches all these maxima. To evaluate the probability of reaching a particular local or a global maximum, we launched 20,000 simulations with randomly chosen initial conditions, and the stopping criterion at the level of 10^{-6}. Fifty-four percent of the simulations reached the global maximum called Configuration 2 (Table 4), 24% reached the second global maximum (Configuration 1), but 22% stopped at the local maximum presented in Table 5.

Table 4. EM haplotype frequency estimates for both global maxima configurations.

Haplotype sequence	EM frequency estimate	
	Configuration 1	Configuration 2
TGT	0.4375	0.4375
GCT	0.1875	0.1875
TTT	0.1250	0.1250
CCT	0.1250	0.0625
TCG	0.0625	0
CCG	0.0625	0.1250
TCT	0	0.0625
GGT	0	0
CGT	0	0
TGG	0	0
CGG	0	0

Table 5. EM haplotype frequency estimates for a local maximum configuration.

Haplotype sequence	EM frequency estimate
	Local maximum configuration
TGT	0.1875
GCT	0
TTT	0.1250
CCT	0
TCG	0
CCG	0.1250
TCT	0.3125
GGT	0.1875
CGT	0.0625
TGG	0
CGG	0

Table 6. Initial conditions leading to each of the global and local maxima.

Haplotype sequence	Initial values of the haplotype frequency		
	Global maximum		Local maximum
	Configuration1	Configuration 2	
TGT	0.0603	0.1421	0.0654
GCT	0.1662	0.0371	0.1048
TTT	0.1365	0.1150	0.1477
CCT	0.0647	0.0202	0.0327
TCG	0.0990	0.1440	0.0366
CCG	0.0161	0.0619	0.0498
TCT	0.0371	0.1460	0.1536
GGT	0.0960	0.0336	0.1303
CGT	0.1384	0.1005	0.0401
TGG	0.0212	0.1347	0.0975
CGG	0.1645	0.0649	0.1415

7. Conclusions

Using simulated and real data examples, we substantiated the observation made by others (Dempster *et al.*, 1977; Excoffier and Slatkin, 1995; Long *et al.*, 1995) that using only one initial condition for the EM algorithm may lead to incorrect results, associated with the existence of local maxima of the likelihood function. Based on multiple simulation studies, we determined that the most frequent haplotypes remain the same, but their frequency estimates could differ. It was already mentioned by the algorithm developers (Dempster *et al.*, 1977; Excoffier and Slatkin, 1995; Long *et al.*, 1995) that the EM algorithm should be started from several initial conditions, but the resulting sensitivity of the final estimates was not fully recognized.

Our study contributes to the understanding of the properties of the EM algorithm when applied to real genomic data with loci deviating from HWE and exhibiting a high linkage disequilibrium. We demonstrated that high average heterozygosity within the analyzed genomic region resulted in a higher numerical complexity of the maximization problem. We evaluated the convergence speed of the algorithm for these data. We observed that on average after less than 20 iterations the algorithm reached its convergence point with the accuracy defined by the criterion that the increase in the loglikelihood function be less than 0.01. The distance of the estimates of the haplotype frequencies from their convergence point (in the sense of the Euclidean norm) was sufficiently small after only few steps of the algorithm. Increasing the accuracy by setting the stopping criterion equal to 10^{-5} results in an increase in the average iteration number to 80.

We looked for relations among the locus heterozygosities, the linkage disequilibrium pattern within the genomic neighborhood and the shape of the loglikelihood function in the vicinity of the global maximum. The very high linkage disequilibrium within the ATM genomic region together with relatively low heterozygosity may explain the smooth shape of the loglikelihood hypersurface and a small number of the observed irregularities. A lower linkage disequilibrium and higher heterozygosity inside the BLM, WRN and RecQL genes may result in a less regular vicinity of the global maximum of the loglikelihood hypersurface and the existence of multiple local maxima on the loglikelihood hypersurface.

Our approach is more systematic than that of Mano *et al.* (2002), who only presented the existence of two global maxima of the loglikelihood hypersurface of the simulated data set. We showed the existence of one local maximum omitted in (Mano *et al.*, 2002), and by using stochastic Monte Carlo simulations, we estimated the probabilities of reaching each of the maxima.

8. Software

There exist several software packages utilizing algorithms to infer haplotype frequencies that are available on the Internet:

- S. Schneider, D. Roessli, L. Excoffier. Arlequin ver. 2.001: Software for population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

  http://anthro.unige.ch/arlequin

- T. Niu, Z.S. Qin, X. Xu, J.S. Liu (Niu *et al.*, 2002). HAPLOTYPER and EM-DeCODER

  http:/www.people.fas.harvard.edu/~junliu/em/em.html

- S. Mano, N. Yasuda, G. Tamiya, H. Inoko, T. Gojobori, T. Imanishi (Mano *et al.*, 2002). RIGHT ver. 1.0: A reasonable indicator of global maxima for haplotype frequencies at the time.

  http://www.jbirc.aist.go.jp/gendiv/RIGHT/

- D. Clayton. SNPHAP ver. 1.0: A program for estimating frequencies of large haplotypes of SNPs.

  http://www-gene.cimr.cam.ac.uk/clayton/software/

- M. Stephens, N.J. Smith, P. Donnely (Stephens *et al.*, 2001). PHASE ver. 1.0: A program for reconstructing haplotypes from population data.

  http://www.stats.ox.ac.uk/mathgen/software.html

- J.S. Liu, S. Qin and T. Niu (Qin *et al.*, 2002). PL-EM. An algorithm for haplotype construction of the Single Nucleotide Polymorphism.

  http://www.people.fas.harvard.edu/~junliu/plem

  Some authors offer their programs on request. Among them are the following ones:

- A.G. Clark (1990): program INFERX,
- J.C. Long *et al.* (1995): program MLOCUS,
- M.E. Hawley and K.K. Kidd (1995): program HAPLO.

Acknowledgments

The author is grateful to Professor Marek Kimmel for his inspiration, many helpful suggestions and his constant encouragement during the preparation of this paper. Most of the work was done during the author's visit at the Department of Statistics, Rice University, Houston TX USA, which was supported by NIH grants GM58545 and CA75432. This work was also partly supported by the grant of the State Committee for Scientific Research (KBN) in Poland 4T11F 01824 in 2003 and the NATO Collaborative Linkage Grant LST.CLG.977845.

References

Bonnen P.E., Story M.D., Ashorn C.L., Buchholz T.A., Weil M.M. and Nelson D.L. (2000): *Haplotypes at ATM identify coding-sequence variation and indicate a region of extensive linkage disequilibrium*. — Am. J. Hum. Genet., Vol. 67, No. 6, pp. 1437–1451.

Chiano M.N. and Clayton D.G. (1998): *Fine genetic mapping using haplotype analysis and the missing data problem*. — Ann. Hum. Genet., Vol. 62, Pt. 1, pp. 55–60.

Clark A.G. (1990): *Inference of haplotypes from PCR-amplified samples of diploid populations*. — Mol. Biol. Evol., Vol. 7, No. 2, pp. 111–122.

Clark V.J., Metheny N., Dean M. and Peterson R.J. (2001): *Statistical estimation and pedigree analysis of CCR2-CCR5 haplotypes*. — Hum. Genet., Vol. 108, No. 6, pp. 484–493.

Dempster A.P., Laird N.M. and Rubin D.B. (1977): *Maximum likelihood from incomplete data via the EM algorithm*. — J. R. Stat. Soc., Vol. 39, No. 1, pp. 1–38.

Excoffier L. and Slatkin M (1995): *Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population*. — Mol. Biol. Evol., Vol. 12, No. 5, pp. 921–927.

Fallin D. and Schork N.J. (2000): *Accuracy of haplotype frequency estimation for biallelic loci, via the Expectation-Maximization algorithm for unphased diploid genotype data*. — Am. J. Hum. Genet., Vol. 67, No. 4, pp. 947–959.

Ghosh S. and Majumder P.P. (2000): *Mapping a quantitative trait locus via the EM algorithm and Bayesian classification*. — Genet. Epidemiol., Vol. 19, No. 2, pp. 97–126.

Hawley M.E. and Kidd K.K. (1995): *HAPLO: A program using the EM algorithm to estimate the frequencies of multi-site haplotypes*. — J. Heredity, Vol. 86, No. 5, pp. 409–411.

Hudson R.R. and Kaplan N.L. (1985): *Statistical properties of the number of recombination events in the history of a sample of DNA sequence*. — Genetics, Vol. 111, No. 1, pp. 147–164.

Kalinowski S.T. and Hedrick P.W. (2001): *Estimation of linkage disequilibrium for loci with multiple alleles: Basic approach and an application using data from boghorn sheep*. — Heredity, Vol. 87, Pt. 6, pp. 698–708.

Lin S., Cutler D.J., Zwick M.E. and Chakravarti A. (2002): *Haplotype inference in random population samples*. — Am. J. Hum. Genet., Vol. 71, No. 5, pp. 1129–1137.

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., Vol. 56, No. 3, pp. 799–810.

Mano S., Yasuda N., Tamiya G., Inoko H., Gojobori T. and Imanishi T. (2002): *Phase space structure if haplotype frequency estimation by the EM algorithm*. — Proc. Waterfront Symp. *Human Genome Science WASH 2002*, Tokyo, Japan.

McKeigue P.M. (2000): *Efficiency of estimation of haplotype frequencies: Use of marker phenotypes of unrelated individuals versus counting of phase-known gametes*. — Am. J. Hum. Genet., Vol. 67, No. 6, pp. 1626–1627.

McLachlan G.J. and Thriyambakam K. (1997): *The EM algorithm and extensions*. — New York: Wiley.

Meng X. and van Dyke D. (1977): *The EM algorithm — An old folk-song sung to a fast new tune*. — J. R. Statist. Soc. B, Vol. 59, No. 3, pp. 511–567.

Niu T., Qin Z.S., Xu X. and Liu J.S. (2002): *Bayesian haplotype inference for multiple linked Single-Nucleotide Polymorphisms*. — Am. J. Hum. Genet., Vol. 70, No. 1, pp. 157–169.

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 halplotype diversity revealed by high-resolution scanning of human chromosome 21*. — Science, Vol. 294, No. 5547, pp. 1719–1723.

Qin Z.S., Niu T. and Liu J.S. (2002):*Partition-Ligation-Expectation-Maximization algorithm for haplotype inference with Single-Nucleotide Polymorphism*. — Am. J. Hum. Genet., Vol. 71, No. 5, pp. 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., Vol. 17, No. 4, pp. 289–295.

Schneider S., Roessli D. and Excoffier L. (2000): *Arlequin 2.001: A software for population genetics data analysis*. — Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Single R.M., Meyer D., Hollenbach J.A., Nelson M.P., Noble J.A., Erlich H.A. and Thomson G. (2002): *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., Vol. 22, No. 2, pp. 186–195.

Slatkin M. and Excoffier L. (1996): *Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm*. — Heredity, Vol. 76, Pt. 4, pp. 377–383.

Stephens M., Smith N.J. and Donnelly P. (2001a): *A new statistical method for haplotype reconstruction from population data*. — Am. J. Hum. Genet., Vol. 68, No. 4, pp. 978–989.

Stephens M., Smith N.J. and Donnelly P. (2001b): *Reply to Zhang et al*. — Am. J. Hum. Genet., Vol. 69, No. 4, pp. 912–914.

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., Vol. 67, No. 2, pp. 518–522.

Trikka D., Fang Z., Renwick A., Jones S.H., Chakraborty R., Kimmel M. and Nelson D.L. (2002): *Complex SNP-based haplotypes in three human helicases: implication for cancer association studies*. — Genome Res., Vol. 12, No. 4, pp. 627–639.

Wang N., Akey J.M., Zhang K., Chakraborty R. 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., Vol. 71, No. 5, pp. 1227–1234.

Wu C.F.J. (1983): *On the convergence properties of the EM algorithm*. — Ann. Stat., Vol. 11, No. 1, pp. 95–103.

Xu C.F., Lewis K., Cantone K.L., Khan P., Donnelly C., White N., Crocker N., Boyd P.R., Zaykin D.V. and Purvis I.J. (2002): *Effectivness of computational methods in haplotype prediction*. — Hum. Genet., Vol. 110, No. 2, pp. 148–156.

Zhang S., Pakstis A.J., Kidd K.K. and Zhao H. (2001): *Comparision of two methods for haplotype reconstruction and haplotype frequency estimation from population data*. — Am. J. Hum. Genet., Vol. 69, No. 4, pp. 906–912.