Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population

Laurent Excoffier and Montgomery Slatkin†*

*Departments of Anthropology and Ecology, University of Geneva and †Department of Integrative Biology, University of California, Berkeley

Molecular techniques allow the survey of a large number of linked polymorphic loci in random samples from diploid populations. However, the gametic phase of haplotypes is usually unknown when diploid individuals are heterozygous at more than one locus. To overcome this difficulty, we implement an expectation-maximization (EM) algorithm leading to maximum-likelihood estimates of molecular haplotype frequencies under the assumption of Hardy-Weinberg proportions. The performance of the algorithm is evaluated for simulated data representing both DNA sequences and highly polymorphic loci with different levels of recombination. As expected, the EM algorithm is found to perform best for large samples, regardless of recombination rates among loci. To ensure finding the global maximum likelihood estimate, the EM algorithm should be started from several initial conditions. The present approach appears to be useful for the analysis of nuclear DNA sequences or highly variable loci. Although the algorithm, in principle, can accommodate an arbitrary number of loci, there are practical limitations because the computing time grows exponentially with the number of polymorphic loci.

Introduction

With the advent of molecular techniques, the survey of polymorphism at several loci or nucleotide sites on the same chromosome has become common. Restriction enzyme or DNA sequence studies now provide data on dozens or hundreds of contiguous nucleotide sites, many of which are expected to be polymorphic. This polymorphism is often desired in population studies, allowing the determination of genetic affinities among populations or groups of populations (Avise 1994), but it is often a problem for interpreting individual genotypes because the gametic phase of multiple-site heterozygous diploids cannot be determined. Different strategies for inferring haplotypes may be used to partially overcome this difficulty. One possibility is that the multiple-site heterozygotes can be eliminated from the analysis, keeping only the homozygotes and the single-site heterozygote individuals, but this approach might lead to a possible bias in the sample composition and the underestimation of low-frequency haplotypes. Another possibility is that single chromosomes can be studied independently, for example, by asymmetric PCR amplification (Newton et al. 1989; Wu et al. 1989) or by isolation of single chromosome by limit dilution followed by PCR amplification (Ruano et al. 1990). Or multiple-site haplotypes can sometimes be inferred using genealogical information in families (Perlin et al. 1994), but then some members of the families must be ignored to get rid of redundant information. These approaches are thus not entirely satisfying either because of their technical complexity, the additional cost they entail, their lack of generalization at a large scale, or the possible biases they introduce.

To overcome these difficulties, Clark (1990) introduced an algorithm based on Hardy-Weinberg equilibrium to infer the phase of PCR-amplified DNA genotypes. The principle is to start filling a preliminary list of haplotypes present in the sample by examining unambiguous individuals, that is, the complete homozygotes and single-site heterozygotes. Then other individuals in the sample are screened for the possible occurrence of previously recognized haplotypes. For each positive identification, the complementary haplotype was added to the list of recognized haplotypes, until the phase information for all individuals is either resolved or identified as unresolved. As discussed by Clark (1990), several problems can arise with this procedure, including the possibility of never being able to start the iterative algorithm because of the absence of any unambiguous individuals in the sample. Although the probability of such an event was assumed to be small for realistic sam-

Key words: maximum likelihood, EM algorithm, multilocus genotype, gene frequency estimation.

Address for correspondence and reprints: Laurent Excoffier, Genetics and Biometry Laboratory, Department of Anthropology, CP 511, 1211 Geneva 24, Switzerland. E-mail: excoffie@sc2a.unige.ch.

Mol. Biol. Evol. 12(5):921–927. 1995.

0737-4038/95/1205-0021$02.00

ples of DNA sequences (Clark 1990), this problem remains. Moreover, Clark's (1990) method assigns a single genotype to each multiheterozygous individual, whereas several genotypes are possible when there are more than one heterozygous site. The assigned genotype may then depend on the order in which individuals are listed in the sample.

Here we take a different approach and place the problem of estimating haplotype frequencies in the general framework of the EM algorithm, formalized by Dempster et al. (1977). Hill (1974) suggested using the equivalent of the EM algorithm with two-locus, two-allele systems, and Weir (1990) noted that the EM algorithm could be adapted to the problem of inferring phase relationships in haplotypes (see also Elandt-Johnson 1971; Piazza 1975; Yasuda 1978; Imanishi et al. 1991). We will first describe the EM algorithm as it applies to the problem of inferring haplotype frequencies under that assumption of Hardy-Weinberg equilibrium and then test its performance using simulated data.

Deriving the Likelihood of Haplotype Frequencies

Here, we will call a *phenotype* a multilocus genotype whose haplotypic phase is unknown a priori. A multilocus genotype defined as a particular combination of two multilocus haplotypes will be called a *genotype* hereafter. The probability of a sample of n individuals conditioned by the phenotype frequencies $P_1, P_2, \ldots, P_m$ (i.e., the likelihood of the data given the parameters) is given by the multinomial probability,

$$P(\text{sample} \mid P_1, P_2, \ldots, P_m) = \frac{n!}{n_1! n_2! \ldots n_m!} \times P_1^{n_1} \times P_2^{n_2} \times \ldots P_m^{n_m}, \quad (1)$$

where m different phenotypes are observed with counts $n_1, n_2, \ldots, n_m$.

The number of genotypes (c_j) leading to the jth phenotype is a function of the number of heterozygous loci s_j,

$$\begin{aligned} c_j &= 2^{s_j - 1}, & s_j &> 0. \\ c_j &= 1, & s_j &= 0. \end{aligned} \quad (2)$$

Under the assumption of random mating, the probability P_j of the jth phenotype is given by the sum of the probabilities of each of the possible c_j genotypes,

$$P_j = \sum_{i=1}^{c_j} P(\text{genotype } i) = \sum_{i=1}^{c_j} P(h_k h_l), \quad (3)$$

where $P(h_k h_l)$ is the probability of the ith genotype made up of haplotypes k and l, $P(h_k h_l) = p_k^2$ if $k = l$ and $P(h_k h_l) = 2p_k p_l$ if $k \neq l$, where p_k and p_l are the population frequencies of the kth and the lth haplotypes. Substituting equation (3) in equation (1), we obtain the probability of the sample as a function of the unknown haplotype frequencies. Therefore, the likelihood of the haplotype frequencies given phenotypic counts is

$$L(p_1, p_2, \ldots p_h) = a_1 \prod_{j=1}^{m} \left(\sum_{i=1}^{c_j} P(h_{ik} h_{il}) \right)^{n_j}, \quad (4)$$

where $p_h = 1 - p_1 - p_2 - \ldots - p_{h-1}$, and a_1 is a constant incorporating the multinomial coefficient.

In principle, the maximum likelihood (ML) estimates of haplotype frequencies could be found analytically by solving a set of $h - 1$ equations involving first partial derivatives of the logarithm of the likelihood, generally called scores. If U_t represents the score for the tth haplotype,

$$U_t = \frac{\partial \log L}{\partial p_t} = \sum_{j=1}^{m} \frac{n_j}{P_j} \frac{\partial P_j}{\partial p_t}, \quad (5)$$

then setting the scores for the $h - 1$ functionally independent haplotypes to zero and solving the resulting set of equations would lead to the ML estimates, but this procedure is tedious when h is large, and the number h is often unknown a priori. Alternative procedures involving numerical iterations have been developed. Among those, when the data are incomplete in the sense that there are more data categories (genotypes) than can be distinguished (phenotypes), Dempster et al. (1977) have formalized the use of an EM algorithm to estimate the haplotype frequencies that maximize the sample probability. In this article, we show how the EM algorithm can be extended to an arbitrary number of loci with an arbitrary number of alleles, allowing the treatment of DNA sequence and highly variable loci data.

The EM Algorithm

The EM algorithm is an iterative method to compute successive sets of haplotype frequencies $p_1, p_2, \ldots, p_h$, starting with initial arbitrary values $p_1^{(0)}, p_2^{(0)}, \ldots, p_h^{(0)}$. These initial values are used as if they were the unknown true frequencies to estimate genotype frequencies $P(h_k h_l)^{(0)}$ (the expectation step). These expected genotype frequencies are used in turn to estimate haplotype frequencies at the next iteration $p_1^{(1)}, p_2^{(1)}, \ldots, p_h^{(1)}$ (the maximization step), and so on, until convergence is reached (i.e., when the changes in haplotype frequency in consecutive iterations are less than some small value).

In the implementation of the EM algorithm, several sets of initial values for haplotype frequencies can be envisioned. For instance, one could assume that, for each phenotype, all possible genotypes are equally likely, so

$$P\,(\text{genotype } h_k h_l \text{ in phenotype } j)^{(0)} = P_j(h_k h_l)^{(0)} = c_j^{-1},\ \forall j. \tag{6}$$

Other possible initial conditions are that all haplotypes are equally frequent in the sample, that haplotype frequencies are equal to the product of single-locus allele frequencies (i.e., complete linkage equilibrium), or that initial haplotype frequencies are chosen at random. In the following, we used equation (6) to find the initial frequencies in most cases, but we also considered other choices as discussed below.

The expectation step at the gth iteration consists of using the haplotype frequencies in the previous iteration to calculate the probability of resolving each phenotype into the different possible genotypes:

$$P(h_k h_l)^{(g)} = \frac{n_j}{n}\,\frac{P_j(h_k h_l)^{(g)}}{P_j^{(g)}}. \tag{7}$$

Haplotype frequencies are then computed for each maximization step using a procedure equivalent to the gene-counting method (Ceppellini et al. 1955; Smith 1957):

$$\hat{p}_t^{(g+1)} = \frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{c_j} \delta_{it} P_j(h_k h_l)^{(g)}, \tag{8}$$

where δ_{it} is an indicator variable equal to the number of times haplotype t is present in genotype i (0, 1, or 2).

A potential problem with this method is that the variance of the estimates of haplotype frequencies is not known. Approximate estimates of the variance-covariance matrix may however be derived for large samples by inverting the estimated $(h-1)\times(h-1)$ information matrix **I**, whose elements are defined as

$$I_{kl} = -E\left(\frac{\partial^2 \log L}{\partial^2 p_k \partial p_l}\right) = n\sum_{j=1}^{m}\frac{1}{P_j}\frac{\partial P_j}{\partial p_k}\frac{\partial P_j}{\partial p_l}; \qquad I_{kk} = n\sum_{j=1}^{m}\frac{1}{P_j}\left(\frac{\partial P_j}{\partial p_k}\right)^2, \tag{9}$$

(see Elandt-Johnson, 1971, p. 332), where these equations are evaluated at $p_k = \hat{p}_k$, the expected haplotype frequencies. In practice, this approach may not lead to the desired results, because the information matrix may prove difficult or impossible to invert for one of the following reasons: the number of possible haplotypes in the sample may be extremely large; some haplotypes may have ML frequencies equal or close to zero, leading to singular or nearly singular information matrices; or a particular estimated information matrix may be singular even when all haplotypes have nonzero frequencies. In our implementation of the algorithm, the estimated information matrix is given as an output.

EM Algorithm Efficiency

Monte-Carlo Simulations

In order to test the effectiveness of the EM algorithm in inferring haplotype frequencies, we generated samples of multilocus haplotypes using the simulation program described by Slatkin (1994). Samples of 25 and 100 diploid individuals were then formed by drawing random pairs of haplotypes. DNA polymorphisms were simulated allowing for four possible alleles at each locus, while highly polymorphic loci, such as microsatellite loci, were simulated by assuming 20 possible alleles. For pseudo-DNA sequence data, the genealogy of haplotypes made up of 2, 5, 10, 20, 50, and 100 distinct sites were simulated either with (R = $4Nc$ = 0.1, where c is the recombination rate per chromosome per generation) or without (R = 0) recombination. For the highly variable loci, we assumed two to eight loci with (R = 1) and without recombination (R = 0). For each locus, the mutation parameter theta (θ) was set to 0.1 for pseudo-DNA data and to 10 for highly variable loci. One hundred independent random samples were generated for each type of data.

Although the present EM algorithm can theoretically handle an arbitrary number of polymorphic sites in a sample, it is limited in practice by the number of possible genotypes, which grows exponentially with the number of polymorphic sites. In our case we have considered samples only when all individuals were heterozygous for fewer than 16 loci and when the total number of possible haplotypes in the sample did not exceed 16,384.

Algorithm Performance Indices

The present EM algorithm may be used for many different goals, such as finding the list of all haplotypes in a sample, finding the list of the most frequent haplotypes, estimating haplotype population frequencies, inferring which gametes are most likely associated to form genotypes in all sampled individuals, or finding the best estimates of coefficients of linkage disequilibrium among loci. There is no single measure of performance of the EM algorithm, because there are many possible uses for it, and the choice of a measure depends on the intended purpose. Here, we propose two measures of performance by comparing the actual values for the simulated data (denoted by the index 0) to ML estimates.

Haplotype frequency estimations.—To examine how close the estimated frequencies are to the actual frequencies, we use the similarity index I_F of Renkonen (1938), defined as the proportion of haplotype frequencies in common between estimated and true frequencies,

$$I_F = \sum_{i=1}^{h} \min(\hat{p}_i, p_{0i}) = 1 - \frac{1}{2}\sum_{i=1}^{h} |\hat{p}_i - p_{0i}|, \qquad (10)$$

where the $\widehat{p_i}$s are the estimated frequencies and the p_{0i}s are the true simulated frequencies. Note that I_{F} may also be considered as one minus half the sum of absolute differences between estimated and true frequencies. It varies between zero, when true haplotypes have estimated frequencies tending to zero, and one, when observed and estimated frequencies are identical. This index gives more weight to the high-frequency haplotypes.

Haplotype identification.—Because the algorithm begins by identifying all possible combinations of haplotypes, the set of true haplotypes will necessarily be included in the set of estimated haplotypes. We will consider that a given haplotype is identified as being present in the true sample if its estimated frequency is above the threshold value of $1/(2n)$. We can then define an index of performance in terms of haplotype identification:

$$I_H = \frac{2(k_{\mathrm{true}} - k_{\mathrm{missed}})}{k_{\mathrm{true}} + k_{\mathrm{est}}}, \qquad (11)$$

where k_{true} is the number of haplotypes in the true sample, k_{est} is the number of estimated haplotypes with frequency above the threshold, and k_{missed} is the number of true haplotypes not identified in the sample. The value of I_H can vary between one, when the identified haplotypes are exactly those present in the true sample, to zero when none of the true haplotypes has been identified.

EM Algorithm Convergence

Although an EM algorithm will always climb the multidimensional likelihood surface, there is no guarantee that the surface is convex or that it is not nearly flat with a narrow isolated peak. In other words, depending on the initial conditions, the EM algorithm may not find the true ML solution because it can lead to an optimum that may not be the global optimum, or the iterative process may stop before reaching the optimum.

We have investigated for the possible occurrence of multiple peaks in the likelihood surface by applying the EM algorithm to our simulated data sets starting from different initial conditions and examining whether convergence to the same solution occurs. As computing time would have been prohibitive if this strategy had been used for all data sets examined here, we have applied it to a few arbitrarily chosen test cases for the highly variable loci. For each test case shown in table 1, 100 runs of the EM algorithm were performed. Each run was done with initial conditions obtained by assigning random but nonzero frequencies to all possible haplotypes in the sample. For each run, the final log likelihood was recorded as well as the performance index I_F. To compare these results to those obtained with initial conditions using equation (6), each data set was also analyzed assuming equal initial haplotype frequencies.

Results

The two algorithm performance indices are plotted against the number of polymorphic sites found in samples of different sizes in figure 1. The general trends appear similar for both DNA and highly variable data. The EM algorithm seems almost insensitive to the recombination rate but performs much better for the largest sample sizes. For DNA data, although the performance is comparable when there are fewer than five polymorphic sites in small and large samples, the decay in both haplotype identification and frequency estimation is rapid in smaller samples. Note that more than 90% of haplotype frequencies are correctly estimated with the larger sample size. For highly variable loci, the average performance of the EM algorithm is consistently worse for samples of 25 than for samples of 100 individuals, even when only two polymorphic loci are considered. The ability of the EM algorithm to infer the correct haplotype frequencies in large samples is high (above 90%) for all cases. This is quite remarkable because usually more than 95% of the individuals had different phenotypes when eight loci were considered.

The results presented in table 1 suggest either that the likelihood surface has multiple peaks or that it has large flat areas causing the EM algorithm to stop at different points, even though the same restrictive stopping criterion (epsilon = 10^{-7}) was used in all instances. This effect appears more pronounced for small sample sizes as the coefficient of variations of final log likelihood reached from initial random frequencies are much smaller for large sample sizes. Considering equal haplotype frequencies as a random initial condition, one sees that the resulting log likelihoods can be considerably smaller than the observed maximum when only 25 individuals are considered, whereas they are always larger than the mean and not far from the maximum when data for 100 individuals are available. This suggests first that a great many initial conditions should be tried for small sample sizes and second that the difference in EM algorithm efficiency observed in figure 1 between small

Table 1
Comparison of Estimated Haplotype Frequencies Obtained from Different Initial Conditions

Sample Size	No. of Polymorphic Loci	Mean log L[a]	Maximum log L[b]	Equal Haplotype log L[c]	I_F Associated with Maximum log L	Maximum I_F Value[d]	I_F Associated with Equal Haplotypes[e]
25	2	−59.1620402 (0.0011716)	−59.1620400	−59.85518732	1.00000	1.00000	0.98000
25	3	−96.5177648 (0.0230308)	−93.8465002	−95.23279525	0.92000	0.96000	0.85000
25	4	−95.0487869 (0.0264914)	−91.9369577	−97.48213574	0.88000	0.88000	0.85000
25	5	−90.7776002 (0.0203197)	−88.6546718	−105.29020472	0.68000	0.68000	0.61250
100	2	−253.8528369 (0.0007217)	−253.7976769	−253.79768073	0.96198	0.96198	0.96198
100	3	−303.5804950 (0.0023830)	−302.8596190	−302.85962398	0.99000	0.99000	0.98000
100	4	−327.7313520 (0.0031837)	−326.8317555	−326.83175863	0.98000	0.98000	0.96000
100	5	−400.9734231 (0.0048590)	−398.3952436	−400.82169383	0.89571	0.90873	0.88214

NOTE.—The haplotypes were generated as described in the text for highly variable loci data with 20 possible alleles per locus, no recombination, and $\theta = 10$ per locus.

[a] Mean final log likelihoods of haplotype frequencies obtained from 100 replicates with random initial haplotype frequencies. The coefficient of variation is shown within parentheses.

[b] Maximum value among the 100 replicates.

[c] Final log likelihood obtained by assuming identical initial frequencies for all haplotypes, as implied from eq. (6).

[d] Among the 100 replicates with random initial haplotype frequencies.

[e] I_F efficiency index obtained for the case where all haplotypes have identical initial frequencies.

and large samples could certainly be attenuated by using several initial conditions and picking up the global ML solution. The last three columns in table 1 show that higher log likelihoods generally lead to better estimation of true sample frequencies, although the best estimates are not always obtained for the highest likelihoods. This may be due to the fact that Hardy-Weinberg equilibrium may not be met in some samples.

Discussion

EM Algorithm Predicted Performances

The present algorithm is based on the assumption of random union of gametes (or haplotypes), and any departure from Hardy-Weinberg (HW) equilibrium may lead to biased estimates of haplotype frequencies and indices based on these frequencies. Besides the assumption of HW equilibrium, information redundancy in the form of multiple copies of the same haplotype in the data set is required for the algorithm to work properly. Thus, the EM algorithm is expected to perform better in large samples than in small ones, both because HW proportions will be more closely achieved and because the number of new haplotypes is not expected to increase linearly with sample size.

Note that no assumption is made about linkage equilibrium among sites in the EM algorithm. However, the present approach is most useful in the presence of linkage disequilibrium, because if there is complete equilibrium alleles would be randomly assigned to possible haplotypes and haplotype frequencies could be obtained from allelic frequencies. For closely linked sites, mutation rates and population demographic history will determine how much linkage disequilibrium one would expect. For example, in a sample from a rapidly expanding population, the correlation between haplotypes is expected to be quite low (Slatkin and Hudson 1991; Slatkin 1994).

Fit of the ML Solutions to the Actual Data

The likelihood is guaranteed to increase under EM algorithm, but the result may be a local and not a global maximum in some cases (see table 1; Weir 1990, p. 64). To avoid such situations, one should try several sets of initial haplotype frequencies. In the sample cases shown in figure 1, we have initially assumed that all genotypes within an individual were equally frequent. Results shown in table 1 suggest that larger likelihoods and better performance indices would have been obtained under

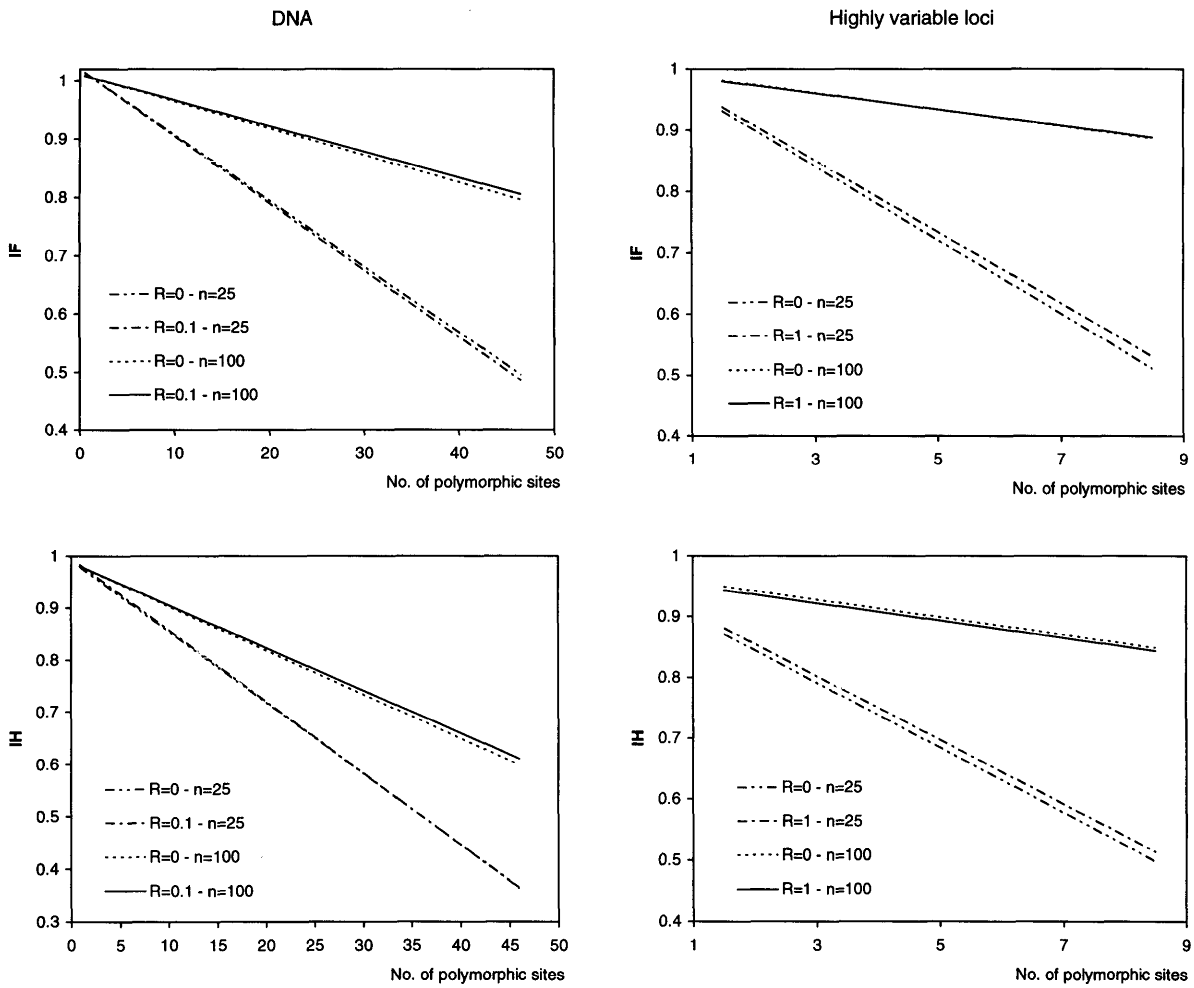

FIG. 1.—Linear regressions of EM algorithm performance indices I_F and I_H against the number of polymorphic sites in the samples for simulated DNA and highly variable loci. Samples of $n = 25$ and $n = 100$ individuals were considered; $R = 4Nr$ is the recombination parameter between adjacent sites or loci. For highly variable loci, each locus was polymorphic in all samples, and the regression lines are build on 100 replicates for each given number of loci considered. For DNA data, only a fraction of the sites were polymorphic in each sample, and the number of polymorphic sites could vary between replicates of the same parameter's conditions.

different initial conditions, especially for small sample sizes. With this enhancement, the conclusions drawn from figure 1, such as the insensitivity of the algorithm to moderate recombination rates and the better performance of the algorithm in large samples, should remain true. We recommend using several initial conditions to ensure the best estimate possible of haplotype frequencies.

Improvement over Other Estimation Procedures

As stated above, the ML treatment presented here elaborates on Clark's (1990) approach for inferring DNA haplotypes. Instead of finding only the list of possible haplotypes, our approach has the advantage of estimating haplotype frequencies, even when there are no homozygotes or single-site heterozygotes in the sample or when all individuals have different phenotypes. Although Clark's algorithm is much faster in situations when there are a large number of polymorphic loci in the sample, it assigns a single genotype to each individual, an assignment that may depend on the order in which individuals are listed in the sample. The EM algorithm is not sensitive to the ordering in the data.

The EM algorithm extends conventional frequency estimation or gene-counting procedures to a large number of loci, whereas procedures based on analytical solutions (Elandt-Johnson 1971; Hill 1974; Imanishi et al. 1991) are limited to a few loci. In practice, the number of loci that can be handled is limited by the amount of polymorphism per locus and by computing time, because the number of possible genotypes grows exponentially with the number of polymorphic sites. Roughly a billion genotypes would have to be examined for an individual heterozygous at 31 sites, and thus the calculations would reach the limits of computing capacity of midsized work stations. But with current estimates of DNA variability in mammals where two homologous nucleotides differ on the average every 200 bp (Nei and Hughes 1991), coding sequences of about 6,000 nucleotides could be accommodated by the present algorithm, although our current implementation can only handle nuclear DNA sequences up to a length of 1,000 nucleotides. The good results obtained for highly variable loci suggest that the present algorithm could be successfully applied to microsatellite data and used to identify the haplotypes formed by the combination of microsatellite patterns at several linked loci. The EM algorithm can in principle also be applied to any other multilocus haplotypes, like RFLP haplotypes, molecular haplotypes derived from oligotyping techniques, or serologically derived haplotypes.

Acknowledgments

We thank Rosalind Harding for initially raising the question, Peter Smouse for his suggestions on the ML approach, and André Langaney for his help throughout this study. L.E. was supported by Swiss National Science Foundation grants no. 31-39847.93 and 32.37821.93, and M.S. was supported in part by National Institutes of Health grant GM40282. A computer program (MLHAPFRE) to compute multilocus haplotype frequencies using the present approach is available from L.E. on request.

LITERATURE CITED

AVISE, J. C. 1994. Molecular markers, natural history and evolution. Chapman & Hall, New York.

CEPPELLINI, R., M. SINISCALCO, and C. A. B. SMITH. 1955. The estimation of gene frequencies in a random mating population. Ann. Hum. Genet. **20**:97–115.

CLARK, A. G. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol. **7**: 111–122.

DEMPSTER, A. P., N. M. LAIRD, and D. B. RUBIN. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. **39**:1–38.

ELANDT-JOHNSON, R. C. 1971. Probability models and statistical methods in genetics. Wiley, New York.

HILL, W. G. 1974. Estimation of linkage disequilibrium in randomly mating populations. Heredity **33**:229–239.

IMANISHI, T., T. AKAZA, A. KIMURA, K. TOKUNAGA, and T. GOJOBORI. 1991. Estimation of allele and haplotype frequencies for HLA and complement loci. Pp. 76–79 *in* K. TSUJI, M. AIZAWA, and T. SASAZUKI, eds. HLA 1991: proceedings of the Eleventh International Histocompatibility Workshop and Conference. Vol. **1**. Oxford University Press, New York.

NEI, M., and A. L. HUGHES. 1991. Polymorphism and evolution in the major histocompatibility complex loci in mammals. Pp. 222–247 *in* R. K. SELANDER, A. G. CLARK, and T. S. WHITTAM, eds. Evolution at the molecular level. Sinauer, Sunderland, Mass.

NEWTON, C. R., A. GRAHAM, L. E. HEPTINSTALL, S. J. POWELL, C. SUMMERS, N. KALSHEKER, J. C. SMITH, and A. F. MARKHAM. 1989. Analysis of any point mutation in DNA: the amplification refractory mutation system (ARMS). Nucleic Acids Res. **17**:2503–2516.

PERLIN, M. W., M. B. BURKS, R. C. HOOP, and E. C. HOFFMAN. 1994. Toward fully automated genotyping: allele assignment, pedigree construction, phase determination, and recombination detection in Duchenne muscular distrophy. Am. J. Hum. Genet. **55**:777–787.

PIAZZA, A. 1975. Haplotypes and linkage disequilibria from three-locus phenotypes. Pp. 923–927 *in* F. KISSMEYER-NIELSEN, ed. Histocompatibility testing 1975. Munskgaard, Copenhagen.

RENKONEN, O. 1938. Statisch-ökologische Untersuchungen über die terrestiche Kaferwelt der finnishen Bruchmoore. Ann. Zool. Soc. Bot. Fenn. Vanamo. **6**:1–231.

RUANO, G., K. K. KIDD, and J. C. STEPHENS. 1990. Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. Proc. Natl. Acad. Sci. USA **87**:6296–6300.

SLATKIN, M. 1994. Linkage disequilibrium in growing and stable populations. Genetics **137**:331–336.

SLATKIN, M., and R. R. HUDSON. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129**:555–562.

SMITH, C. A. B. 1957. Counting methods in genetical statistics. Ann. Hum. Genet. **21**:254–276.

WEIR, B. S. 1990. Genetic data analysis. Methods for discrete population genetic data. Sinauer, Sunderland, Mass.

WU, D. Y., L. UGOZZOLI, B. K. PAL, and R. B. WALLACE. 1989. Allele-specific amplification of β-globin genomic DNA for diagnosis of sickle-cell anemia. Proc. Natl. Acad. Sci. USA **86**:2757.

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

JULIAN P. ADAMS, reviewing editor

Received October 13, 1994

Accepted March 22, 1995