

HAPLO: A Program Using the EM Algorithm to Estimate the Frequencies of Multi-site Haplotypes

M. E. Hawley and K. K. Kidd

A DNA haplotype system combines information from two or more distinct polymorphic systems located within a small length of DNA in which there is little recombination. Each possible combination of alleles, one from each component system, constitutes a distinct haplotype that is treated as an allele of the haplotype system. Combining the information from several polymorphic systems (such as RFLPs) into such a haplotype system defines alleles that arise from recombination among as well as mutation at the individual systems. Haplotype systems generally have been found to show a great deal of variability within and among populations, and thus can be more informative for genetic comparisons between populations and for linkage mapping purposes than the component systems treated independently.

Even when all component systems are codominant, a characteristic of haplotype systems is ambiguity, by which we mean that some phenotypes may correspond to several genotypes. When an individual is heterozygous for no more than one of the component systems, the genotype is uniquely specified; there is no ambiguity. However, even the simplest haplotype system of two loci, each with two alleles, can present an ambiguous phenotype: an individual who is heterozygous at both sites (typing Aa and Bb) may have either genotype AB/ab or Ab/aB. Sometimes these ambiguities can be resolved when data are available for related individuals; they are usually not directly resolvable for an "isolated" individual, but see Stephens et al. (1990) and Ruano et al. (1990).

Another characteristic of real datasets is missing data on one (or

more) of the component loci. Because it is now common for there to be five or more polymorphic loci close enough together to be haplotyped, it is often very time-consuming and expensive to go back to complete typing for every system on all individuals. As the number of polymorphic sites (loci) in a haplotype increases and the number of alleles at one or more sites increases above two, the proportion of individuals with ambiguity can increase dramatically; so can the difficulty in having complete typing data on all individuals.

Existing computer programs have limitations. Those of Weir (1990) are limited to analysis of two loci with two codominant alleles each and to data sets with complete typing information, and do not allow any known phase information on individuals to be incorporated. Clark (1990) describes an algorithm for determining a minimally sufficient set of haplotypes to explain an observed set of phenotypes; it does not estimate frequencies. Long et al. (1995) describe a program that uses the EM algorithm of Dempster et al. (1977) to simultaneously estimate allele frequencies and the necessary lower and higher order disequilibrium coefficients to determine haplotype frequencies. Though the program can handle one recessive allele at each site (the rest being codominant), it cannot use known phase information or include incomplete data.

We have written a FORTRAN program, HAPLO, that implements the EM algorithm to estimate haplotype frequencies from phenotype data on samples of unrelated individuals. The EM algorithm is a generalized iterative maximum likelihood approach to estimation that is useful when data are ambiguous and/or incomplete. Our implementation is for autosomal loci in Hardy-Weinberg proportions; we are working to extend it to X-linked systems. In addition to our desire that it should analyze large haplotype systems, HAPLO was specifically designed to deal with the two limitations in the existing programs that were most relevant to our studies: (1) incomplete data on some individuals due to failure of typing for one (or more) of the component loci, and (2) the availability of data on relatives that allows complete or partial resolution of the genotype for some individuals with otherwise ambiguous phenotypes. Both situations were common for much of the DNA marker data being collected in our lab. We have used this program to estimate frequencies of haplotypes at several different loci in different human

populations (Kidd et al. 1993; Lu et al., in press). With the growing use of DNA technology, studies of natural populations of many species can involve haplotype systems, making this program more broadly useful. We also note that, although motivated by haplotype studies, HAPLO treats each haplotype as an allele in a multi-allelic system and does not use the underlying nature of the data to obtain frequency estimates. In this sense, it is no different from frequency estimation for any multi-allelic system except that it allows the more complicated genotype-phenotype correspondences resulting from ambiguity and missing data.

The EM algorithm is an iterative process; each iteration gives a set of frequency estimates that converge to stable maximum likelihood estimates. The iterations start with all haplotypes (alleles) at equal frequency. It is easy to show that the frequency estimates for haplotypes that are not definitely "observed," i.e., required to explain a phenotype, will go to zero. This is what is expected for a maximum likelihood estimate, as discussed by Clark (1990) in his description of an algorithm for determining the minimum set of haplotypes required to explain a sample of phenotypes. This would appear to indicate that, if the set of phenotypes in the dataset is sufficiently clear to show from simple inspection that a possible haplotype is not required to explain the data, it can be omitted from the genetic model without altering the ultimate frequency estimates. This is usually true, but one interesting counterexample we observed in one dataset shows that "required" has a probabilistic aspect. We observed one individual with a multiply heterozygous phenotype that could be explained by several genotypes. Only two of those possibilities involved haplotypes otherwise definitely present in the sample. In these two possibilities, however, a different common haplotype was heterozygous with a different haplotype not otherwise seen. Neither of these "new" haplotypes was absolutely required, but one or the other had to be present. Their maximum likelihood frequency estimates were fractions of $1/2N$, proportional to the frequency estimates of the two common haplotypes. Thus, although prior elimination of haplotypes can be valid, care must be used.

The HAPLO program optionally estimates standard errors in two ways. First a jackknife procedure is used. Estimates of all haplotype frequencies

are recalculated with each individual in turn removed from the data set. For each haplotype the standard deviation of those frequency estimates is an estimate of the standard error of the original frequency estimate. This estimate of the standard error takes into account the added uncertainty in the data because of the ambiguity and missing data as well as the sampling error. The second estimate of the standard error of each haplotype frequency applies the formula for the binomial standard error assuming all haplotypes were directly observed and counted. This assumption can be correct or a very close approximation in some instances and in any case the binomial standard error estimate serves as a lower bound check on the jackknife estimates, which can become inaccurate when the number of observations is small. Neither of these approaches is particularly satisfying, because the jackknife can be inaccurate and the binomial is biased, but they are both calculated to provide some sense of the accuracy of the frequency estimates for each specific dataset.

An option in the program allows the user to define the relationships between the haplotypes and specific alleles at multiple, multi-allelic sites (currently up to a maximum of 20 alleles at each of 10 sites). If this is done, allele frequencies at the component sites are calculated and these used to estimate "expected" haplotype frequencies assuming random association of alleles across sites. A likelihood ratio χ^2 is calculated to test whether the maximum likelihood haplotype frequency estimates are significantly different from those "expected." This is a test of whether linkage disequilibrium exists in the overall system of sites but does not indicate which sites are involved and does not calculate any disequilibrium coefficients – those are beyond the scope of this program. The program prints out the degrees of freedom for the likelihood ratio χ^2 of the full system, but the actual degrees of freedom may be different in different populations and not always obvious. We explicitly leave it to the user to decide the appropriate degrees of freedom for each specific instance.

Input for the program consists of a single ASCII file, prepared using any available editor. Our DEC VMS implementation prompts the user interactively for the names of the input and output files. The input file contains the definition of basic parameters, such as the number of alleles in the haplotype

system, a list defining the correspondence between the observed phenotypes and the possible genotypes, and the number of observations of each phenotype in each population. The current version of HAPLO allows up to 114 haplotypes (alleles), 114 observed phenotypes, and 500 genotypes. Because only observed phenotypes and the genotypes necessary to explain them are required, these limits have not been an impediment to analyses of quite large systems, but these arbitrary limits can easily be changed. A single input file can include observations for up to 30 populations for the given haplotype system; independent analyses are done for each population in sequence.

The use of a genotype-phenotype correspondence list gives users great flexibility in defining the haplotype system. Because the user defines the correspondence, it is not necessary to supply any explicit information about the component polymorphic systems. Definitions need to be entered only for the observed phenotypes and whatever genotypes may be implied; it is not necessary to define all the possible combinations of alleles. It is possible to define phenotypes that make use of any available phase information or that cover cases where data on the component systems are incomplete. Individuals with an ambiguous phenotype that is partially or fully resolved are included in the analysis by simply defining a new phenotype that corresponds to only the single or few remaining genotype(s). Individuals with missing data are included by defining a phenotype that corresponds to the set of genotypes that includes all phenotypes at the missing site but only those possible for the typed sites. Recessive alleles can be incorporated as can even more complex genotype-to-phenotype correspondence systems such as linear dominance (Cotterman 1953).

The basic structure of the program consists of a subroutine that executes the iterative algorithm, and a main routine that reads in the data file and executes calls to the subroutine. For each population, this routine is called once to estimate the haplotype frequencies; if the option to calculate jackknife errors is selected, the routine is called once more for each phenotype observed in that population with the appropriately altered "raw" data. The resulting set of frequencies is weighted according to the number of observations of that phenotype to calculate the jackknife error.

At each iteration, the likelihood of the phenotypes is calculated from the current frequency estimates. This likelihood is compared to that of the previous iteration. In the current version of the program, the algorithm is considered to have converged when the difference in the log likelihoods from two successive iterations is <0.001 . In practice, this usually means convergence to at least the sixth significant figure for the log likelihood and at least 10^{-4} for the haplotype frequencies. As a precaution, the calculation will halt after 100 iterations, because such a high number would probably be due to an error in the system definition.

Both our approach and that of Long et al. (1995) handle multiple loci and multiple alleles at each locus. Both calculate MLE estimates for alleles and haplotypes, and both use a likelihood ratio test for overall disequilibrium. However, the programs are different in significant ways. An advantage of their program is that higher order disequilibrium coefficients can be calculated and a variety of specific hypotheses tested. Also, they obtain confidence limits by a bootstrap approach. The advantage of ours is that it can incorporate incomplete data and known phase data. For such data, a jackknife estimate of error is more appropriate. Also, because the genotype-phenotype relations are explicitly defined, our program can handle more complicated dominance relations, such as linear dominance. Indeed, HAPLO has been demonstrated to accurately estimate frequencies for several many-to-one complex phenograms (Cotterman 1953); it cannot handle many-to-many phenograms.

This article describes a revised version of the program, previously called EM-HAPLO (Hawley et al. 1994); significant changes include (1) defining the phenotype-genotype associations in a list format that is more user-friendly than the previous, rather unwieldy matrix format, (2) testing for convergence of the iterations on the basis of likelihood, rather than the specific frequency values, and (3) the addition of the option for a likelihood ratio test of overall disequilibrium. HAPLO version 2.0 is written in VMS FORTRAN, version 5.9; we are distributing the program as well-documented source code, which we have tried to make readily adaptable to other systems. We also plan to specifically prepare a Microsoft FORTRAN version for use on IBM-

compatible PC's. The program runs quickly: as an example, on a VAX 7610, rated at 48 Specmarks, a very complex model based on 3 RFLPs, two with 2 alleles and one with 6, which includes 24 haplotypes, 114 observed phenotypes (some with missing data, some with known or partially known phase), and 210 required genotypes, was analyzed for 26 populations with 20-68 individuals each. The full analysis with error calculations needed 7.5 min of CPU time; estimating only the MLE haplotype frequencies, without the error calculations, took <31 CPU-seconds. For 24 of the populations, all of the subroutine calls converged in 6-27 iterations; the other two populations required up to 51 iterations per subroutine call. Data and results for nine of those 26 populations are in Lu et al. (in press).

FORTTRAN source code, example data files, and user's documentation are available via anonymous ftp from [paella.med.yale.edu](ftp://paella.med.yale.edu), in the directory `pub/haplo`. News of improvements and extensions to the program will be posted to this site.

Note added in proof. After this article was accepted, we learned of a similar program developed independently for Windows platforms (Excoffier and Slatkin, in press).

From the Department of Genetics, Yale University School of Medicine, New Haven, CT 06510-8005. We would like to thank Dr. Kathryn Roeder (Department of Statistics, Carnegie-Mellon University) for her advice and two anonymous reviewers for their suggestions. The research was funded in part by NIH grants AA09379, HG00365, and MH38929 and NSF grant DBS9208917.

The Journal of Heredity 1995:86(5)

References

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

Cotterman CW, 1953. Regular two-allele and three-allele phenotype systems. *Am J Hum Genet* 5:193-235.

Dempster AP, Laird NM, and Rubin DB, 1977. Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc Ser B* 39:1-22.

Excoffier L and Slatkin M, in press. Maximum-likelihood estimates of molecular haplotype frequencies in diploid populations. *Mol Biol Evol*.

Hawley ME, Pakstis AJ, and Kidd KK, 1994. A computer program implementing the EM algorithm for haplotype frequency estimation. *Am J Phys Anthropol* 18:104.

Kidd JR, Pakstis AJ, and Kidd KK, 1993. The genetic relationship of new and old world populations based on haplotypes. *Am J Hum Genet* 53(suppl.):818. (Abstract.)

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

Lu RB, Ko HC, Chang CM, Castiglione G, Schoolfield G, Pakstis AJ, Kidd JR, and Kidd KK, in press. No association between alcoholism and multiple polymorphisms at the dopamine D2 receptor gene (DRD2) in three distinct Taiwanese populations. *Biol Psychol*.

Ruano G, Kidd KK, and Stephens JC, 1990. Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. *Proc Natl Acad Sci USA* 87:6296-6300.

Stephens JC, Rogers J, and Ruano G, 1990. Theoretical underpinnings of the single-molecule-division (SMD) method of direct haplotype resolution. *Am J Hum Genet* 46:1149-1155.

Weir BS, 1990 *Genetic data analysis*. Sunderland, Massachusetts: Sinauer.

Received May 25, 1994
Accepted May 2, 1995

Corresponding Editor: Stephen O'Brien