Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 11684–11688, November 1994
Genetics

Sequential imputation for multilocus linkage analysis

MARK IRWIN[†], NANCY COX[‡], AND AUGUSTINE KONG[§]

[†]Department of Statistics, Ohio State University, Columbus, OH 43210-1247; [‡]Department of Medicine, The University of Chicago, Chicago, IL 60637; and [§]Department of Statistics, The University of Chicago, Chicago, IL 60637

Communicated by David Botstein, August 5, 1994

ABSTRACT A Monte Carlo method called sequential imputation is proposed for multilocus likelihood computations. This method is most useful in mapping situations where the data consist of large pedigrees with substantial missing information and it is desirable to perform linkage analysis utilizing data from many polymorphic markers simultaneously. A pedigree example with 155 individuals, 9 loci, and 155,520 haplotypes is used for illustration.

In mapping disease loci or genetic markers, often many linked loci have to be handled simultaneously. Efficient algorithms for calculating likelihoods are available for large pedigrees with a small number of loci (1–5) and for small pedigrees with a large number of loci (6). However, for large pedigrees with a large number of loci, especially those that have substantial missing data, exact evaluation of a single likelihood value can be prohibitive because of memory requirements and computing time. This difficulty was noted explicitly by investigators studying disorders having late age at onset (7, 8). A Monte Carlo method called *sequential imputation* (9, 10) is proposed here to handle problems of this type. Loci are processed one, or a few, at a time to reduce the computational burden. Instead of evaluating likelihood values individually, the whole likelihood surface can sometimes be obtained by using results from a single simulation run. Also, unlike some other methods (6), sequential imputation can incorporate genetic interference with no extra difficulties.

Sequential Imputation

In multilocus linkage problems, if, for each person and each locus, it is known exactly what allele is inherited from the father and what allele is inherited from the mother, the likelihood function is often trivial to evaluate. We refer to this information, which is desirable but often not available in its entirety, as *missing data* and denote it by $\mathbf{z}$. The observed data, denoted by $\mathbf{y}$, usually include genotypes of each individual marker for some members of the pedigree. An individual may be typed for some, but not all, of the marker loci. In disease mapping, $\mathbf{y}$ will also include available disease phenotypes of the members. The combination $(\mathbf{y}, \mathbf{z})$ is referred to as the *complete data.*

Let θ be the unknown parameter vector so that the likelihood function is $L(\theta) = p_\theta(\mathbf{y})$. In disease mapping, θ is often a scalar that denotes the location of the disease gene relative to a set of markers whose locations are assumed to be known. In more complex situations, θ may also incorporate other parameters such as marker allele frequencies and parameters relating the disease genotype and phenotype. In linkage mapping of markers, θ is a vector that denotes the relative locations among a collection of markers.

Let $\{y_1, \ldots, y_n\}$ and $\{z_1, \ldots, z_n\}$ be some decomposition of $\mathbf{y}$ and $\mathbf{z}$. At this time, assume that there are n loci so that for $t = 1, \ldots, n$, y_t and z_t are, respectively, the observed and missing data on locus t. Other decompositions will be considered later. Note that the labels t $(t = 1, \ldots, n)$ do not necessarily correspond to the *physical* ordering, assumed or real, of the loci. Given a certain value of θ, sequential imputation is a method that allows us to obtain an unbiased estimate of $L(\theta)$ and generate weighted samples of $\mathbf{z} = \{z_1, \ldots, z_n\}$ from the conditional distribution $p_\theta(\mathbf{z}|\mathbf{y})$. These weighted samples can then be used to estimate likelihoods of other parameter values. The method involves first drawing z_1^* from $p_\theta(z_1|y_1)$ and computing $w_1 = p_\theta(y_1)$. Then the following two steps are applied for $t = 2, \ldots, n$, in increasing order of t:

(*i*) Draw z_t^* from the conditional distribution $p_\theta(z_t|y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*, y_t)$. Notice that the z_t^* values have to be drawn sequentially since each z_t^* is drawn conditioned on the previously imputed missing parts $z_1^*, \ldots, z_{t-1}^*$.

(*ii*) Sequentially compute the predictive probabilities $p_\theta(y_t|y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*)$ and $w_t = w_{t-1}p_\theta(y_t|y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*)$. Let $w = w_n$ so that $w = p_\theta(y_1)\Pi_{t=2}^n p_\theta(y_t|y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*)$.

Given the decompositions described above, for each t, steps *i* and *ii* are done simultaneously and involve a single locus computation (11–13). This type of computation is commonly referred to as *peeling*. Steps *i* and *ii* are done independently m times. The choice of m, the number of imputations, is discussed later. Denote the results by $\mathbf{z}^*(1), \mathbf{z}^*(2), \ldots, \mathbf{z}^*(m)$ and $w(1), \ldots, w(m)$, where $\mathbf{z}^*(j) = (z_1^*(j), \ldots, z_n^*(j))$ for $j = 1, \ldots, m$. Note that $\mathbf{z}^*(j)$ is sampled from a distribution, denoted by $p_\theta^*(\mathbf{z}|\mathbf{y})$, which is different from the actual conditional distribution $p_\theta(\mathbf{z}|\mathbf{y})$. It can be demonstrated that

$$w(j) = \frac{p_\theta(\mathbf{y}, \mathbf{z}(j))}{p_\theta^*(\mathbf{z}(j)|\mathbf{y})} = \frac{p_\theta(\mathbf{z}(j)|\mathbf{y})}{p_\theta^*(\mathbf{z}(j)|\mathbf{y})} p_\theta(\mathbf{y}),$$

and as a consequence $E_{p^*}[w(j)] = p_\theta(\mathbf{y})$ (see *Appendix I*). It follows that an unbiased estimate of $L(\theta) = p_\theta(\mathbf{y})$ is

$$\hat{L}(\theta) = \bar{w} = \frac{1}{m}\sum_{j=1}^{m} w(j).$$

Furthermore, if the simulations are performed based on a parameter value θ_0, then, for any other parameter value θ_1,

$$\hat{L}(\theta_1) = \frac{1}{m}\sum_{j=1}^{m} \frac{p_{\theta_1}(\mathbf{y}, \mathbf{z}(j))}{p_{\theta_0}(\mathbf{y}, \mathbf{z}(j))} w(j) \qquad [1]$$

is an unbiased estimate of $L(\theta_1) = p_{\theta_1}(\mathbf{y})$ (see *Appendix II*). Since both $p_{\theta_0}(\mathbf{y}, \mathbf{z}(j))$ and $p_{\theta_1}(\mathbf{y}, \mathbf{z}(j))$ are complete data likelihoods, they can be easily computed. It is noted that $\hat{L}(\theta_1)$ is only expected to be a good estimate of the true likelihood

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: IBD, identity by descent; MODY, maturity onset diabetes of the young; CEPH, Centre D'Étude du Polymorphisme Humain; EM, expectation-maximization.

if θ_1 and θ_0 are not too far apart; i.e., they do not correspond to two very different positions for the disease gene.

Efficiency of the Method

The coefficient of variation of $\overline{w}$, $C[\overline{w}]$, measures the *relative* standard error of $\overline{w}$ as an estimate of $L(\theta)$. Based on the delta method, $C[\overline{w}]$ can be shown to be approximately the standard deviation of $\log_e(\overline{w})$. Changing from the natural logarithm to the logarithm to the base 10, the standard error of $\log_{10}(\overline{w})$ as an estimate of the log-likelihood $\log_{10}L(\theta)$ is approximately $\log_{10} e \times C[\overline{w}] \approx 0.43 \times C[\overline{w}]$. We have

$$C[\overline{w}] = \frac{1}{\sqrt{m}} C[w(j)] = \frac{1}{\sqrt{m}} \frac{\sqrt{\mathrm{Var}_{p^*}[w(j)]}}{E_{p^*}[w(j)]}$$

$$= \frac{1}{\sqrt{m}} \frac{\sqrt{\mathrm{Var}_{p^*}[w(j)]}}{p_\theta(\mathbf{y})}.$$

Its sample estimate is

$$\hat{C}[\overline{w}] = \frac{1}{\sqrt{m}} \hat{C}[w(j)] = \frac{1}{\sqrt{m}} \frac{s_w}{\overline{w}},$$

where s_w denotes the sample standard deviation of the $w(j)$ values. For $C[\overline{w}]$ to be some desirable value δ, m, the number of imputations, needs to be $\delta^{-2} \times (C[w(j)])^2$. In other words, the efficiency of the method is inversely proportional to $(C[w(j)])^2$. For example, suppose we want $C[\overline{w}]$ to be around 0.2. Basing our decision on the simulated samples, this implies that m, the number of imputations, needs to be about $25 \times (s_w^2/\overline{w}^2)$.

Sequential imputation is a form of importance sampling. The distribution $p^*_\theta(\cdot|\mathbf{y})$ from which the $\mathbf{z}^*$ values are drawn is called the *trial distribution*, the ratio $p_\theta(\mathbf{z}^*(j)|\mathbf{y})/p^*_\theta(\mathbf{z}^*(j)|\mathbf{y})$ is the importance sampling weight. So $w(j)$ is the importance sampling weight multiplied by the unknown constant $p_\theta(\mathbf{y})$. It follows that $(C[w(j)])^2 = \mathrm{Var}_{p^*}[p_\theta(\mathbf{z}^*(j)|\,\mathbf{y})/p^*_\theta(\mathbf{z}^*(j)|\,\mathbf{y})]$ and is a measure of *distance* between $p_\theta(\mathbf{z}|\mathbf{y})$ and $p^*_\theta(\mathbf{z}^*|\mathbf{y})$. To keep this distance small, it is desirable to have $p^*_\theta(\cdot|\mathbf{y})$ as close to $p_\theta(\cdot|\mathbf{y})$ as possible by choosing an appropriate decomposition of $\mathbf{y}$ and $\mathbf{z}$. In general, choosing an optimal decomposition requires making compromises between the ease of performing steps *i* and *ii*, and keeping $C[w(j)]$ small. In the previous section (*Sequential Imputation*), for simplicity, we considered a special decomposition of the observed data $\mathbf{y}$ and the missing data $\mathbf{z}$; i.e., y_t and z_t denote, respectively, the observed and missing data of a single locus t. We now present a few modifications of the basic procedure that can reduce the variation of $w(j)$ considerably without necessarily increasing difficulties in computation.

Note that $p_\theta(\mathbf{z}|\mathbf{y})$ can be written as $p_\theta(z_1|\mathbf{y})\Pi_{t=2}^n p_\theta(z_t|\mathbf{y}, z_1, \ldots, z_{t-1})$. So drawing z_1^* from $p_\theta(z_1|\mathbf{y})$ is obviously preferable to drawing z_1^* from $p_\theta(z_1|\,y_1)$ *if* the former is feasible. While this is not the case, it suggests that when drawing z_1^*, we should try to condition on more information if possible. For each locus and each parent–offspring pair, define an identity by descent (IBD) variable as the indicator of whether the allele inherited by the offspring came from the grandfather or the grandmother. Often some of the IBD variables can be deduced from the observed data $\mathbf{y}$. Here we redefine y_1 to include the observed data on the first locus processed plus the IBD variables of other loci, which can be deduced from the observed data. Conditioning on these IBD variables has virtually no effect on the amount of computations needed to perform steps *i* and *ii*.

Apart from the deducible IBD variables, it had so far been assumed that y_1 consists of observed data on a single locus. This is not necessary and, indeed, it is usually preferable to incorporate more than one locus into y_1. Note that the first step of sequential imputation involves computing $p_\theta(y_1)$ and drawing $z_1^*(j)$, $j = 1, \ldots, m$ from $p_\theta(z_1|\,y_1)$. This requires peeling the loci incorporated in y_1 jointly, but the key is that a single peel is needed for all m imputations. As long as this first peel can be practically performed, y_1 should incorporate as many loci as possible. This will decrease the coefficient of variation of the weights and as a consequence can reduce the overall computing time.

In the previous section (*Sequential Imputation*), $\mathbf{z}$ is defined to include every locus and every member in the pedigree. In some cases, some members of the pedigree are typed for some, but not all, of the loci. For a particular person and locus, we call the missing allele data *ignorable* if neither the person nor any of his/her descendants is typed for that locus. Imputing ignorable data will only add noise and inflate the variation of the weights. Hence, for each t, $t = 1, \ldots, n$, we redefine z_t to include only data that are not ignorable.

Order of Imputation and Location Scores

The order in which the loci are processed also affects the trial distribution and the variance of $w(j)$. The best order is one that minimizes the weight variance. A simple rule for choosing a good processing order is to start with loci that have the least amount of missing information among the nonignorable data. Thus, marker loci with more untyped individuals who are not ignorable should be processed late. For two marker loci that are typed in the same individuals, the more informative one, usually the one with more alleles, should be processed first. These are however only guidelines, and sometimes experimentation with different orders is necessary.

Location scores for a disease gene, the differences between the $\log_{10}$ likelihoods of specific gene locations and the $\log_{10}$ likelihood of a position unlinked to the markers, can sometimes be estimated by a simple strategy. Set y_n to be the observed disease data and process the markers first based on the above criteria. The average of the weights before processing the disease data, $\overline{w_{n-1}} = m^{-1} \Sigma_{j=1}^m w_{n-1}(j)$, is an unbiased estimate of $p(y_1, \ldots, y_{n-1})$. Hence $\overline{w_{n-1}} \times p(y_n)$ is an unbiased estimate of the likelihood for a locus unlinked to the markers. Then process the disease locus by placing it at various locations linked to the marker loci. This approach has the advantage that one set of marker imputations can be used to estimate likelihoods of all locations (14). Moreover, since the likelihood estimates at different locations result from a single simulation run, they tend to be positively correlated. In consequence, standard errors of estimates of likelihood ratios among different locations are lowered. Because of these advantages, this strategy of processing the disease locus last is applied to the example presented below. However, recent experience with other data suggests that it is sometimes necessary to process the disease locus first, maybe jointly with one or two markers. This alternative strategy, even though it requires multiple simulation runs for the different gene locations, is preferred when the disease status is available for many individuals in the upper generations while marker genotypes of the same individuals are missing. This can occur, for example, when a highly penetrant disease can be diagnosed in three or more generations at the top of the pedigree for whom marker data are unavailable. Also, if the disease allele is very rare in the population, disease genotypes of many individuals in the upper generations, even if not available directly, can often be deduced with little uncertainty.

Whether the disease locus is processed first or last, it is usually enough to apply sequential imputation to a single location, probably in the middle, within each interval spanned by two physically adjacent markers. Eq. 1 can then be applied to approximate the likelihoods for other locations in the interval.

An Example

The RW pedigree (Fig. 1) segregating for maturity onset diabetes of the young (MODY) is used to illustrate different properties of sequential imputation. The form of MODY segregating in this pedigree has been linked to markers on 20q (15). Note that the diagnostic information summarized in Fig. 1 is derived from both the clinical diagnosis of MODY and from biochemical studies. Thus, some individuals who do not have clinical disease are considered affected in these analyses. It is understood that results of the analyses are dependent on the diagnostic assumptions made. The analyses here are not presented to justify any particular diagnostic criteria or localization of the *MODY* locus within this region but strictly as an illustration of the use of sequential imputation. An unaffected branch of the pedigree is included because of the additional marker information it provides for untyped members of the upper generations.

Of interest is the location of the *MODY* gene relative to eight markers, *ADA1* (5 alleles), *ADA2* (2 alleles), *L127* (6 alleles), *S22* (3 alleles), *S4* (2 alleles), *RM292* (12 alleles), *GPR* (6 alleles), and *GSA* (3 alleles), all on the long arm of chromosome 20. Over half the people in the pedigree are not typed for at least one of the eight markers, and most members of the top two generations have all the marker data missing. The recombination probabilities between the eight markers are assumed to be 0 between *ADA1* and *ADA2*, 0.034 between *ADA* and *L127*, 0.050 between *L127* and *S22*, 0.121 between *S22* and *S4*, 0.011 between *S4* and *RM292*, 0.111 between *RM292* and *GPR*, and 0.132 between *GPR* and *GSA*. The locations of *ADA*, *L127*, *S22*, *S4*, *GPR*, and *GSA* are based on information from Centre D'Étude du Polymorphisme Humain (CEPH) families (16). The position of *RM292* was estimated from the RW pedigree data alone given the positions of the other markers. Three analyses are run on the data set to examine the properties of sequential imputation. For all the analyses presented, the assumption of no interference is made, but this is not a limitation of the method. Genetic distances are measured relative to *ADA*. Computations are performed assuming that MODY is a dominant trait, the penetrance is 0.95, there are no sporadic cases, and the population frequency of the disease allele is 0.0001.

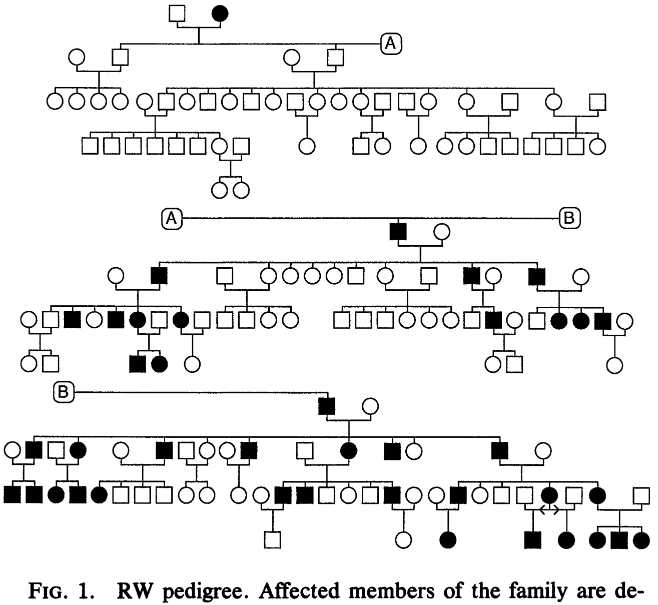

FIG. 1. RW pedigree. Affected members of the family are denoted by solid squares (males) and solid circles (females).

The first analysis uses sequential imputation to calculate a series of three-point location scores for the position of the *MODY* locus between pairs of adjacent markers (four points for the interval between the two *ADA* markers and *L127*). This means that location scores within a marker interval are computed using only data on the flanking markers. This is done because exact computations can be performed for comparisons. For each of the six intervals defined by the markers, a sequential impútation run of $m = 2000$ imputations was done. For each run, the *MODY* locus was placed at four locations: unlinked to the markers, in the middle of the interval, and at the ends of the intervals on top of the markers. Initially, except for one interval, the processing order starts with the marker with the most alleles and finishes with *MODY*. The one exception is the *ADA–L127* interval, with the processing order set to *ADA1, L127, ADA2*, and finally *MODY*. *ADA1* was processed before *L127* because significantly more people were typed for *ADA1* than for *L127*. Two intervals, *S22–S4* and *GPR–GSA*, had large coefficients of variation in the initial runs, and additional runs were performed. For *S22–S4*, the likelihoods were calculated from simulations processing *MODY* first. With *GPR–GSA*, the likelihoods were calculated from a run switching the processing order of *GPR* and *GSA*. Fig. 2 shows the estimated location scores for the whole region together with some exact location scores calculated using the LINKMAP program of the LINKAGE package (5). Sequential imputation is apparently performing well here.

The second analysis utilizes data from all eight markers simultaneously. Location scores for the *MODY* gene are calculated assuming the CEPH locations for the markers. This is hence a nine-point analysis with 155,520 haplotypes. The markers are processed in the order of *RM292, ADA1, L127, GPR, S22, GSA, ADA2, S4*. Then the disease gene is processed at seven different locations: unlinked to the markers and at the midpoints of the six intervals defined by the markers. A total of $m = 10{,}000$ imputations are performed. The results are presented in Fig. 3*A*. Although it can only be clearly seen in the interval *S4–RM292*, there are three curves in each marker interval. The curve in the middle corresponds to the Monte Carlo estimates of the location scores. The top curve and the bottom curve correspond, respectively, to the estimate plus and minus two standard errors. Hence, for each location, the top and bottom values give an approximate 95% confidence interval for the actual location score. The fact that

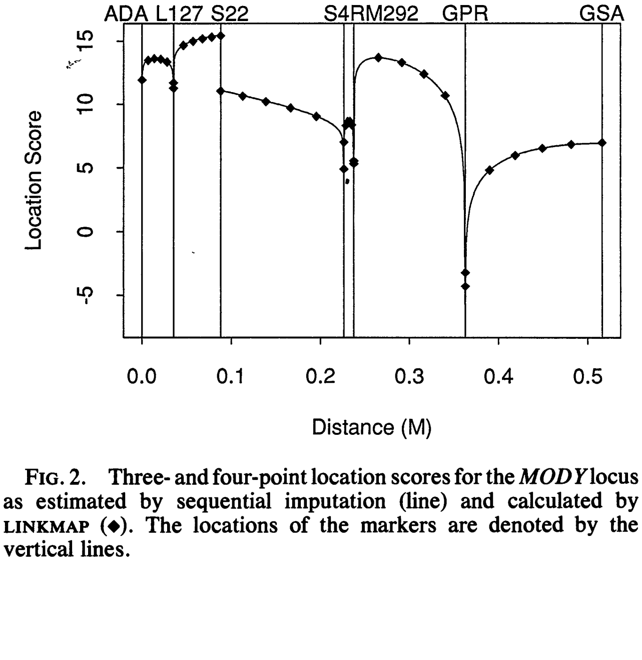

FIG. 2. Three- and four-point location scores for the *MODY* locus as estimated by sequential imputation (line) and calculated by LINKMAP (♦). The locations of the markers are denoted by the vertical lines.

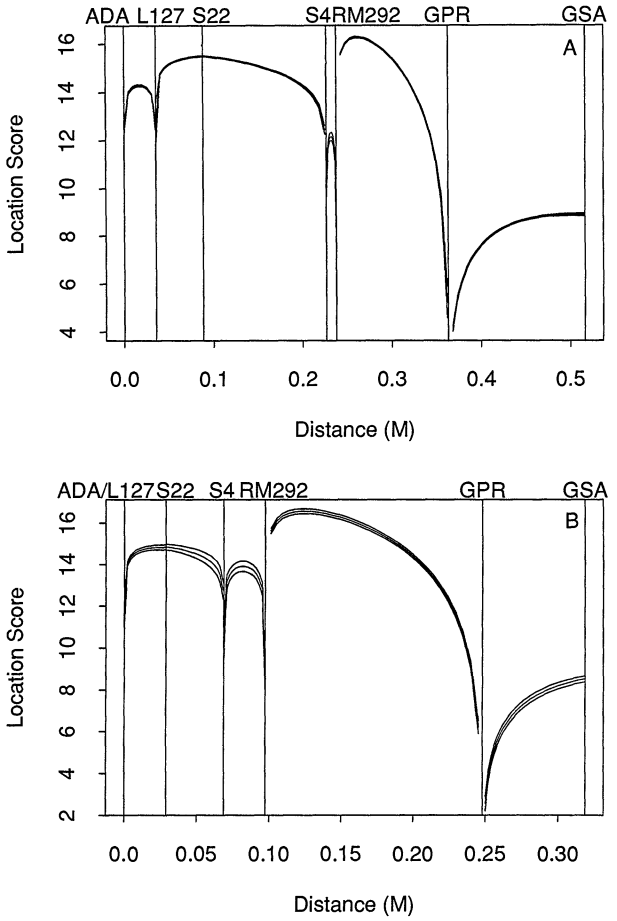

FIG. 3. Estimated nine-point location scores for *MODY* with ±2 SE bounds. (*A*) Under the CEPH distances. (*B*) Under distances estimated by the Monte Carlo expectation-maximization (EM) algorithm.

the three curves are nearly indistinguishable from each other indicates that the standard errors are very small. Note that, compared to the three-point location scores, the interval with the highest location score shifted from *L127–S22* to *S4–RM292*. This indicates the importance of utilizing all marker data simultaneously.

One limitation of exact calculations is that changing parameters such as distances between the markers requires a new set of calculations. In contrast, with sequential imputation, likelihoods under different marker distances can be computed from the same set of simulations by applying Eq. 1. Because the CEPH data for chromosome 20 have not been corrected for typing errors (17), we obtain maximum likelihood estimates of the marker distances based on the RW family data by implementing a Monte Carlo version of the EM algorithm, which also uses sequential imputation (18). These estimated recombination probabilities are 0 between *ADA1* and *ADA2*, 9.5 × 10^{-12} between *ADA* and *L127*, 0.028 between *L127* and *S22*, 0.038 between *S22* and *S4*, 0.028 between *S4* and *RM292*, 0.130 between *RM292* and *GPR*, and 0.066 between *GPR* and *GSA*. Using the same complete data sets and weights generated for the second analysis, location scores for the *MODY* locus are recalculated based on these new marker distances. As shown in Fig. 3*B*, the likelihood is again maximized between *RM292* and *GPR* at a distance of 0.0267 morgans from the *RM292* locus. However, the difference between the maximum location score in the *RM292–GPR* interval and the location score at *S22* has increased from 0.77 to 1.74, which is nearly a factor of 10 on the likelihood scale. Also, while still acceptable, the standard errors here are substantially larger. This is to be expected as the likelihood calculations are being done at a greater distance from the simulation conditions than in the previous analysis.

The nine-point sequential imputation run was performed on a Sun Sparc 10 workstation with 32 megabytes of random access memory (RAM). It took roughly 35 central processing unit hours to run, or ≈13 sec per imputation. (It is noted that the part of the program that performs peeling has not yet been optimized for speed.) With such a high number of haplotypes, exact computations cannot be performed utilizing all the data simultaneously. [In general, for pedigrees without loops, memory requirement is proportional to (number of haplotypes)^4 and computing time is proportional to (number of haplotypes)^6.] Indeed, other investigators working to map the *MODY* locus in this region reported a multipoint location score map containing some of these same markers typed in this same pedigree (7). In these analyses, four-point location scores were calculated by moving *MODY* through intervals of three loci at a time. However, in order to do four-point likelihood calculations, the numbers of alleles were reduced and the pedigree was split. These actions led to a maximum multipoint location score smaller than the highest two-point logarithm of odds score.

Discussion

Because of the limitations of existing computer programs and algorithms that do exact computations of likelihoods, investigators often have to reduce the number of loci and the number of alleles per locus in their analyses (7, 8). This leads to loss of information and sometimes can create bias. In addition, because of the inefficiency of computing likelihoods point by point, careful analysis of the data could be discouraged. The method of sequential imputation introduced in this paper can reduce considerably the burden for multipoint computations and therefore enables more complete analyses, including, for example, assessing the sensitivity of the results to alternative marker allele frequencies or marker distances. Eq. 1 can be applied for most of these purposes, but a note of caution is needed. As mentioned earlier, when θ_1 and θ_0 in the formula are very far apart, the likelihood estimate can be very inaccurate. For example, imputations from one simulation run can be used to find the direction of the effect of modifying the allele frequencies, but the actual estimates of the likelihoods may not be very good if the "new" allele frequencies are very different from those used to perform the simulations.

Sequential imputation is one method to generate multiple samples of the missing data conditioned on the observed data. Alternative methods include the Gibbs sampler (19, 20) and the related Metropolis algorithm (14). Instead of generating independent weighted samples, these methods, based on Markov chain theory, produce correlated samples with equal weights. Because of the special character of pedigree analysis (20), these methods can sometimes be very inefficient because of high correlations among samples. Moreover, while sequential imputation gives direct estimates of likelihoods, these methods only give estimates of likelihood ratios between other values of the parameter and the value used for performing the imputations. Estimates of the likelihood ratios can have very large variances if we are comparing different orderings of the loci. It is however noted that some of the problems being handled by Gibbs sampling, such as inbred pedigrees with many loops and complex traits, do not fall into the area of applications of sequential imputation.

The efficiency of sequential imputation depends on the coefficient of variation of the importance sampling weights. Earlier, a number of ways to improve efficiency were proposed. All were implemented for the analysis of the MODY pedigree, except for the capability of processing more than one locus at a time. Our recent experience with other data makes it clear that this capability will greatly increase the efficiency of the method when marker data are missing for three or more generations at the top of the pedigree. In another direction, a locus can be split into two artificially. For example, a locus with 12 alleles can be considered as two loci right on top of each other with 4 and 3 alleles each. The split can be chosen so that one of these half-loci carries more information than the other half and is processed first. Moreover, two halves of two different loci can be combined during processing to reduce the variation of the weights. Indeed, since only a single peel is performed to process y_1 for all m imputations, one strategy is to have y_1 include all loci by reducing each locus to 2 or 3 alleles. Sequential imputation can then be used to incorporate the *residual* information from each locus. Finally, although the current computer program only handles pedigrees without loops, sequential imputation can in theory be used for pedigrees with a small number of loops.

Appendix I

Note that $\mathbf{z}^*(j)$ is drawn from the density

$$p^*_\theta(\mathbf{z}^*(j)|\mathbf{y})$$

$$= p_\theta(z^*_1(j)|\,y_1) \prod_{t=2}^{n} p_\theta(z^*_t(j)|\,y_1, z^*_1(j), y_2, z^*_2(j), \ldots, y_{t-1}, z^*_{t-1}(j), y_t)$$

$$= \frac{p_\theta(z^*_1(j), y_1)}{p_\theta(y_1)} \prod_{t=2}^{n} \frac{p_\theta(y_1, \ldots, y_t, z^*_1(j), \ldots, z^*_t(j))}{p_\theta(y_1, \ldots, y_t, z^*_1(j), \ldots, z^*_{t-1}(j))}$$

$$= \frac{p_\theta(y_1, \ldots, y_n, z^*_1(j), \ldots, z^*_n(j))}{p_\theta(y_1)} \prod_{t=2}^{n} \frac{p_\theta(y_1, \ldots, y_{t-1}, z^*_1(j), \ldots, z^*_{t-1}(j))}{p_\theta(y_1, \ldots, y_t, z^*_1(j), \ldots, z^*_{t-1}(j))}$$

$$= p_\theta(\mathbf{y}, \mathbf{z}^*(j)) \frac{1}{p_\theta(y_1) \prod_{t=2}^{n} p_\theta(y_t|\,y_1, z^*_1(j), \ldots, y_{t-1}, z^*_{t-1}(j))}$$

$$= \frac{p_\theta(\mathbf{y}, \mathbf{z}^*(j))}{w(j)}.$$

So

$$E_{p^*}[w(j)] = \sum_{\mathbf{z}^*} w(j) p^*_\theta(\mathbf{z}^*(j)|\mathbf{y}) = \sum_{\mathbf{z}^*} w(j) \frac{p_\theta(\mathbf{y}, \mathbf{z}^*(j))}{w(j)}$$

$$= \sum_{\mathbf{z}^*} p_\theta(\mathbf{y}, \mathbf{z}^*(j)) = p_\theta(\mathbf{y}).$$

Appendix II

$$E_{p^*_{\theta 0}}\left[\frac{p_{\theta_1}(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j)} w(j)|\mathbf{y}\right] = \sum_{\mathbf{z}^*} \frac{p_{\theta_1}(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j))} w(j) p^*_{\theta_0}(\mathbf{z}^*(j)|\mathbf{y})$$

$$= \sum_{\mathbf{z}^*} \frac{p_{\theta_1}(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j))} w(j) \frac{p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j))}{w(j)} = \sum_{\mathbf{z}^*} p_{\theta_1}(\mathbf{y}, \mathbf{z}^*(j)) = p_{\theta_1}(\mathbf{y}).$$

This research was supported in part by National Institutes of Health Grant R01-GM46800.

1. Elston, R. C. & Stewart, J. (1971) *Hum. Hered.* **21**, 523–542.
2. Lange, K. & Elston, R. C. (1975) *Hum. Hered.* **25**, 95–105.
3. Cannings, C., Thompson, E. A. & Skolnick, M. H. (1978) *Adv. Appl. Probab.* **10**, 26–61.
4. Lange, K. & Boehnke, M. (1983) *Hum. Hered.* **33**, 291–301.
5. Lathrop, G. M., Lalouel, J. M., Julier, C. & Ott, J. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 3443–3446.
6. Lander, E. S. & Green, P. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 2363–2367.
7. Rothschild, C. B., Akots, G., Hayworth, R., Pettenati, M. J., Rao, P. N., Wood, P., Stolz, F.-M., Hansmann, I., Serino, K., Keith, T. P., Fajans, S. S. & Bowden, D. W. (1993) *Am. J. Hum. Genet.* **52**, 110–123.
8. Schellenberg, G. D., Bird, T. D., Wijsman, E. M., Orr, H. T., Anderson, L., Nemens, E., White, J. A., Bonnycastle, L., Weber, J. L., Alonso, M. E., Potter, H., Heston, L. L. & Martin, G. M. (1992) *Science* **258**, 668–671.
9. Kong, A., Liu, J. S. & Wong, W. H. (1994) *J. Am. Stat. Assoc.* **89**, 278–288.
10. Kong, A., Cox, N., Frigge, M. & Irwin, M. (1993) *Genet. Epidemiol.* **10**, 483–488.
11. Ott, J. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 4175–4178.
12. Ploughman, L. M. & Boehnke, M. (1989) *Am. J. Hum. Genet.* **44**, 543–551.
13. Kong, A. (1991) *Genet. Epidemiol.* **8**, 81–103.
14. Lange, K. & Sobel, E. (1991) *Am. J. Hum. Genet.* **49**, 1320–1334.
15. Bell, G. I., Xiang, K.-S., Newman, M. V., Wu, S.-H., Wright, L. G., Fajans, S. S., Spielman, R. S. & Cox, N. J. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 1484–1488.
16. NIH/CEPH Collaborative Mapping Group (1992) *Science* **258**, 67–86.
17. Keith, T., Swain, P., Serino, K., Yu, H., Ma, N. & Falls, K. (1992) *Am. J. Hum. Genet.* **51**, 191 (abstr.).
18. Kong, A., Irwin, M., Cox, N. & Frigge, M. (1992) Department of Statistics Technical Report 351 (Univ. of Chicago, Chicago).
19. Guo, S. W. & Thompson, E. (1992) *Am. J. Hum. Genet.* **51**, 1111–1126.
20. Kong, A. (1991) *Proceedings of the 23rd Symposium on the Interface Between Computer Science and Statistics* (Interface Found. N. Am., Fairfax Station, VA).