Maximum Likelihood Estimation of Recombination Rates From Population Data

Mary K. Kuhner, Jon Yamato and Joseph Felsenstein

Department of Genetics, University of Washington, Seattle, Washington 98195

Manuscript received December 3, 1999
Accepted for publication June 29, 2000

ABSTRACT

We describe a method for co-estimating $r = C/\mu$ (where C is the per-site recombination rate and μ is the per-site neutral mutation rate) and $\Theta = 4N_e\mu$ (where N_e is the effective population size) from a population sample of molecular data. The technique is Metropolis-Hastings sampling: we explore a large number of possible reconstructions of the recombinant genealogy, weighting according to their posterior probability with regard to the data and working values of the parameters. Different relative rates of recombination at different locations can be accommodated if they are known from external evidence, but the algorithm cannot itself estimate rate differences. The estimates of Θ are accurate and apparently unbiased for a wide range of parameter values. However, when both Θ and r are relatively low, very long sequences are needed to estimate r accurately, and the estimates tend to be biased upward. We apply this method to data from the human lipoprotein lipase locus.

SEQUENCES from loci that do not undergo recombination (for example, mammalian mitochondrial loci) have the same coalescent genealogy at every site. In contrast, when recombination occurs adjacent sites may have different, although correlated, genealogical histories. Reconstructing these genealogies with certainty is impossible. However, we can make inferences about parameters such as population size and recombination rate by considering a sample of plausible genealogies.

We can construct a structure, analogous to a genealogy, which shows the ancestry of each site from a set of sampled individuals. Unlike an ordinary coalescent genealogy this is not necessarily a branching tree; its general form is a directed graph, as described by Hudson (1990). Such graphics have been called "ancestral recombination graphs" by Griffiths and Marjoram (1996); we use the phrase "recombinant genealogies" to emphasize their logical similarity to genealogies. Figure 1 shows a recombinant genealogy for a small case.

A recombinant genealogy can provide information about two parameters: $\Theta = 4N_e\mu$, the product of effective population size and neutral mutation rate per site, and $r = C/\mu$, the ratio between per-site recombination rate and per-site mutation rate. If the genealogy were known with certainty (including all of its branch lengths), we could make a maximum likelihood estimate of these parameters using a simple extension of Kingman's coalescent (Kingman 1982a,b) as suggested by Hudson (1990). Since the true genealogy is unavailable in practice, we estimate the parameters by sampling over many different recombinant genealogies in proportion to each genealogy's expected contribution to the likelihood.

We have previously described similar methods for estimating Θ and for co-estimating Θ and exponential growth rate (Kuhner *et al.* 1995, 1998). The current method is an extension of these, though it is more complex because rearrangement of recombinant genealogies is more difficult. The basic method is maximum likelihood estimation using Metropolis-Hastings sampling (Metropolis *et al.* 1953; Hastings 1970) of candidate genealogies. We sample genealogies on the basis of their posterior probability with regard to the data at a trial value of the parameters and then use the sampled genealogies to evaluate the relative likelihood of other values of the parameters. This importance sampling approach concentrates the sampled genealogies in regions of high posterior probability, which is much more efficient than using random genealogies from the prior, and avoids the potential bias of using only a single genealogy reconstruction as in the UPBLUE method of Fu (1994).

Such an approach is especially useful in cases with recombination, since the space to be sampled is larger (rendering random sampling particularly inefficient) and the chance of making a single genealogy reconstruction that is correct or nearly so is much smaller.

The sampler with recombination is useful both in estimating the recombination rate itself and in allowing accurate estimation of other parameters from data containing recombinations. Attempting to estimate, for example, Θ in recombinant sequences using an approach that does not take recombination into account will lead to an overestimate (Kuhner *et al.* 2000). Thus, in addition to its use in estimating recombination rate, this algorithm opens the way for more precise use of nuclear

Corresponding author: Mary K. Kuhner, Department of Genetics, University of Washington, Box 357360, Seattle, WA 98195-7360.
E-mail: mkkuhner@genetics.washington.edu

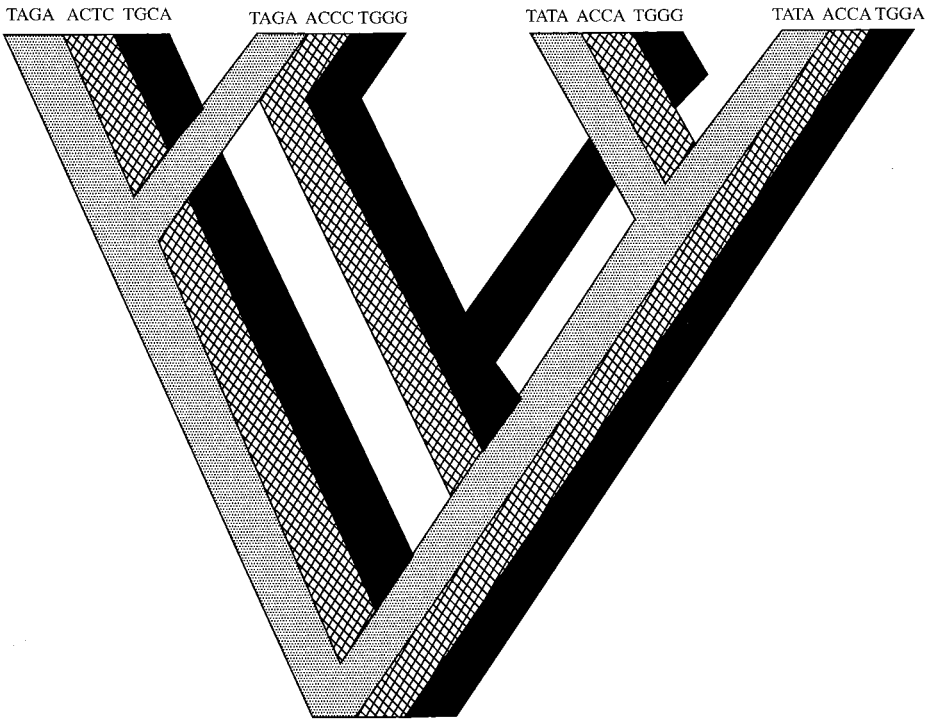

FIGURE 1.—A recombinant genealogy of four sequences, showing two recombination events dividing the genealogy into three subtrees.

gene sequences in assessing other population parameters.

We are aware of at least two other likelihood-based approaches to this problem. GRIFFITHS and MARJORAM (1996) describe a method that also uses importance sampling to estimate recombination rate on the basis of consideration of sampled genealogies, but using the independent-sample algorithm developed by GRIFFITHS and TAVARÉ (1993). FELSENSTEIN *et al.* (1999) compare the conceptual bases of these approaches. R. NIELSEN (unpublished results) describes a sampler somewhat similar to ours, but in a Bayesian framework and without co-estimation of Θ. Further work will be required to compare the strengths and weaknesses of the different approaches.

METHODS

Parameters: Our approach estimates two parameters. The first is $\Theta = 4N_e\mu$, where N_e is the effective population size in individuals and μ is the mutation rate per site per generation. The second is $r = C/\mu$, where C is the recombination rate per unit of intermarker distance per generation. In both cases, μ is present because we cannot directly observe time, only the occurrence of mutations. We assume that the relative distance between markers is known. In the simple case of continuous DNA sequence, for example, all sites could be assumed to be equally spaced. For single nucleotide polymorphism (SNP) data known distances between SNP markers could be provided. We assume that multiple recombinations do not occur simultaneously and therefore that interference between recombinations need not be considered: this is a requirement of the diffusion approximations used in the Kingman-Hudson coalescent.

When the algorithm is applied to DNA or RNA sequences or SNPs, a site is a single aligned column of base pairs. For future applications to microsatellite or electrophoretic data, a "site" is a single microsatellite locus or a single electrophoretic locus, and recombination is estimated between linked "sites."

Definition of a recombinant genealogy: A nonrecombinant coalescent genealogy is defined (KINGMAN 1982a,b) as just those lineages that contribute to the observed individuals. Several ways to extend this concept to a recombinant genealogy can be imagined. We choose to define the recombinant genealogy of a sample as containing only lineages that contribute at least one site to the sample, and only recombinations that separate at least two sites contributed to the sample. Lineages and recombinations that do not fulfill these criteria will not leave any trace in the sampled sequences, and defining the genealogy to exclude them reduces its size, making it easier to analyze.

Further reductions in the size of the recombinant genealogy, without loss of information, are certainly possible. We discuss one such reduction under "final coalescent tactic" below. Another reduction would be to remove from consideration all recombinations that are "loops"; that is, the two sequences that combine in a recombination are nearest neighbors with no intervening coalescences. Loops have no effect on the data and could therefore be discarded, but we have not yet worked out the necessary adjustments to the prior probability.

Calculation of probabilities: The Metropolis-Hastings

genealogy sampler requires two terms, the probability of the data with respect to the genealogy [$P(D|G)$] and the probability of the genealogy with respect to the parameters [in this case, $P(G|\Theta, r)$].

In the recombinant case, $P(D|G)$ can be calculated straightforwardly for any of a large number of mutational models as long as the model assumes that sites are independent and therefore only the genealogy of a specific site is relevant to its likelihood. The independence assumption can be relaxed as long as the correlation between sites does not depend on the genealogy—for example, adjacent sites can be allowed to have autocorrelated rates.

Our implementation for nucleotide sequences is based on the one in PHYLIP (Felsenstein 1993). It uses the substitution model of Felsenstein as described in Kishino and Hasegawa (1989), which allows unequal nucleotide frequencies and differences between transition and transversion rate. Variation of rate among sites, with possible autocorrelation, is accommodated by the Hidden Markov model of Felsenstein and Churchill (1996). Other substitution models could easily be used instead.

For SNP data we could use the same model, omitting autocorrelation (since SNP sites are usually widely spaced and correlation between adjacent sites seems unlikely) and applying a correction to the term $P(D|G)$, accounting for the method by which the SNPs were ascertained (Kuhner *et al.* 2000).

We assume that the mutational model parameters (such as transition/transversion ratio and nucleotide frequencies) are known. Experience with phylogeny estimations suggests that maximum likelihood methods are fairly robust to errors in specification of the mutational model (see, for example, Fukami-Kobayashi and Tateno 1991).

The prior probability of the genealogy, $P(G|\Theta, r)$, is an extension of Kingman's coalescent (Kingman 1982a,b) developed by Hudson (1990). The genealogy is considered as a series of time intervals, starting at the tips; each time interval is bounded by an event (either recombination or coalescence). The rate of coalescence is $\Theta/[k(k-1)]$, where k is the number of lineages in that time interval. The rate of recombination is rs, where s is the length of the region in which a recombination could occur, summed over all k lineages.

Since we have defined the recombinant genealogy to include only recombinations that separate two or more sites destined to contribute to the observed data, we must count as potential locations for recombination only those inter-site links that would produce such a recombination. Thus, only a subset of inter-site links, called "eligible links," are counted in calculating the probability of recombination. Figure 2 shows an example: if a region is destined to be replaced, higher up in the genealogy, with a different sequence, we do not allow any recombinations within it, since they would have no effect on the descent of the actual data.

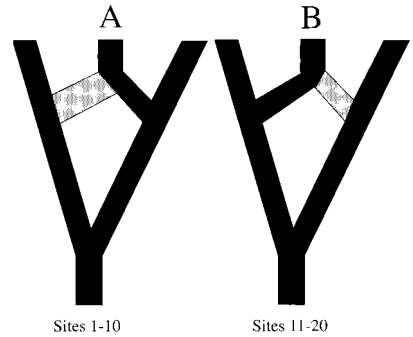

Figure 2.—Eligible sites. Subgenealogies for sites 1–10 and 11–20 are shown. There is a single recombination event; sites 1–10 are contributed by the right-hand sequence and sites 11–20 by the left-hand one. On subtree A, sites 1–10 are not eligible for further recombination in the branch marked in gray because they do not pass through that branch to the tips; similarly, on subtree B, sites 11–20 are not eligible on the gray branch. Such sites are not considered when calculating the probability of recombination.

For such genealogies the Kingman-Hudson prior is

$$P(G|\Theta, r) = \left(\frac{2}{\Theta}\right)^K r^X \exp\left(\sum_i -\left[\frac{k_i(k_i-1)}{\Theta} + rs_i\right]t\right), \quad (1)$$

where K is the total number of coalescences in that genealogy, X is the total number of recombinations, i is the number of the time interval, k_i is the number of lineages in interval i, and s_i is the weighted sum of eligible links in interval i, weighting each link by its length.

This can be understood as the product of three types of terms: the probability density $2/\Theta$ of a specific coalescence, the probability density r of a specific recombination, and the probability $\exp(-(k(k-1)/\Theta + rs)/t)$ of the waiting time until the next event.

Overview of procedure: The Metropolis-Hastings genealogy sampler that we originally derived for nonrecombinant sequences (Kuhner *et al.* 1995, 1998) works by a two-phase process. It begins with an initial genealogy and an initial value of the parameters: in the simplest case, the single parameter Θ, whose initial value is called Θ_0. In the first phase, a new genealogy is created by locally rearranging the previous genealogy in proportion to the coalescent prior probability $P(G|\Theta_0)$ (given by Kingman 1982a,b). In the second phase, this genealogy is accepted or rejected based on $P(D|G)$, the probability of the sequence data on the genealogy (given by Felsenstein 1981). Repetition of this process, which is equivalent to sampling from the posterior probability $P(G|\Theta_0)P(D|G)$, produces a set of genealogies from which a maximum likelihood of Θ can be made. For the case without recombination the relative likelihood becomes

$$\frac{L(\Theta)}{L(\Theta_0)} \approx \frac{1}{n}\sum_i \frac{P(G_i|\Theta)}{P(G_i|\Theta_0)} \quad (2)$$

(Kuhner *et al.* 1995). This estimator is most efficient when Θ_0 is close to its maximum likelihood estimate (MLE), so it is useful to run several iterations of the sampler, using the estimated Θ of each iteration as the starting Θ_0 of the next.

The local rearrangement scheme that was successful for nonrecombinant genealogies (Kuhner *et al.* 1995, 1998) cannot be directly adapted to recombinant genealogies. It confines each rearrangement strictly within a local neighborhood in the genealogy, whereas in recombinant genealogies any lineage can interact with any other. Instead, we adopt a more drastic form of rearrangement: a lineage is broken at random, all lines leading rootward from that lineage into the rest of the genealogy are erased, and then the broken lineage is resimulated from the breakpoint downward toward the root, choosing events based on $P(G|\Theta_0, r_0)$ conditional on the structure of the nonerased tree, until all resulting lineages have coalesced. This approach was originally developed by Beerli and Felsenstein (1999) for use in the sampler with migration. Some additional complexity is needed in the sampler with recombination; this is discussed in the implementation section.

Multiple chains: The importance sampling approach used in this sampler is most effective when Θ_0 and r_0 are near their MLEs. For this reason, it is useful to do a series of short chains, beginning each run with the maximum likelihood estimates of Θ and r from the previous chain. Because the algorithm with recombination rejects a higher proportion of genealogies and is searching a higher-dimensional space, the short chains must be longer than for the nonrecombinant sampler.

This procedure must be modified if the estimate of r from a given chain is 0. Since genealogies generated with $r_0 = 0$ will never contain recombinations, and analysis of Equation 1 shows that a set of genealogies containing no recombinations will always return $r = 0$, 0 represents an absorbing state that the process can never leave once it is reached. We have experimented with setting r to a small positive value (data not shown) but it is difficult to choose an appropriate value. Instead we use an approach developed by Beerli and Felsenstein (1999) in which we add one recombination to the final genealogy of each chain that would otherwise have none, except for the final chain. This prevents the "fatal attraction" of r to 0 and seems to minimize interference in the process, since adding a single recombination represents the least possible unit of change in the estimate. We do not introduce an arbitrary recombination into the final chain, so the biasing effect of this correction is limited to its influence on the final chain's r_0 value.

Construction of likelihood curve: At intervals, genealogies are sampled from the Markov process for use in constructing a likelihood curve. When there are no recombinations in any of the sampled genealogies, the MLE of r should be zero; we fix it to this value and estimate the maximum likelihood value of Θ. Otherwise we co-estimate the maximum for r and Θ.

Multiple loci: If information from multiple unlinked regions of the genome is available, an overall estimate of the parameters can be made by multiplying together the regional likelihood curves, as long as any differences in the values of the parameters are assumed to be known (for example, haploid and diploid data could be combined, using a factor of two difference in the value of N_e). Our previous samplers have been much more efficient with multiple unlinked loci than with a single locus, but we do not expect that to be the case here, since a single lengthy region with recombination combines the advantages of multiple loci (*i.e.*, multiple site genealogies) with maximum ability to detect recombinations.

IMPLEMENTATION

This section discusses the details of the genealogy-rearrangement process. Throughout, "down" is toward the root of the genealogy.

The objective is to erase and redraw a portion of the genealogy in accordance with several goals. New genealogies should be produced in proportion to $P(G|\Theta, r)$ to the extent that this is possible. Any deviations from this proportion must be matched with a compensating Hastings term (Hastings 1970). For the Markov chain to work properly, there must be no irreversible steps; if the process can go from genealogy A to genealogy B in one step, it must be able to go from B to A in one step as well. Finally, for the sampler to be efficient it is desirable not to change the genealogy too much in one step, or the new genealogy will probably be rejected.

We begin by choosing one branch of the genealogy, other than the root branch, at random. This branch is erased, as are all recombinations rootward of and dependent on it. The branch is then resimulated downward according to a conditional recombinant-coalescent prior until all lineages are once again attached to the genealogy. This resimulation includes a rate of recombination (a lineage being simulated splits into two lineages, each containing a subset of the sites) and a rate of coalescence (a lineage under simulation coalesces either with another lineage under simulation or with a lineage already in the tree).

Lineages being resimulated are called "active" lineages, and lineages in the remainder of the genealogy are "inactive" lineages.

Resimulation downward involves drawing a waiting time until the next event (using the Kingman-Hudson prior) and then checking to see if that time falls within the current time interval. If it does, an event is added to the genealogy, and the process is repeated for the new time interval just below the added event. If it does not, the process proceeds down to the next time interval and another time is drawn.

This procedure can create a new genealogy that is outside the defined state space of legal genealogies. A rearrangement in the upper part of the genealogy can render an existing branch lower down meaningless—it no longer contributes any sites to the tips. For example, imagine a branch whose only contribution is to supply site 1 to tip 7. A recombination is now added higher in the tree that supplies site 1 to tip 7 from a different lineage. The original branch now has no sites to contribute.

Since we define recombinant genealogies as having no such branches, it is tempting to reject such rearrangements, but doing so would mean that all rearrangements of the upper parts of highly recombinant trees would tend to be rejected, preventing the sampler from mixing well.

It is also tempting to simply prune out meaningless branches as they arise, but this violates the reversibility of the Markov chain. If we make a change and prune out a lower branch, reversing the change would require "pruning in" the missing branch. To preserve reversibility, if we are going to remove a branch when all its meaningful sites have been taken from it, we must be prepared to add a branch when new meaningful sites become available during the rearrangement process.

This is done by keeping a record of which links were meaningful in each branch of the old genealogy. Construction of the new genealogy occurs from the tips downward, time interval by time interval. For each time interval, each inactive branch is checked to see whether it contains links newly made meaningful by the ongoing rearrangements. If it does, evaluation of the interval includes the possibility that a new recombination may occur at one of the newly meaningful links. At the end of the rearrangement process, the genealogy is checked for meaningless branches and they are removed. This process is fully reversible and preserves the desired sampling from $P(G|\Theta, r)$.

These new recombinations can be thought of as "hidden passages": conceptually they are always present, but they reveal their presence only when, due to rearrangements elsewhere, they have an influence on the descent of sites to the sampled individuals. We do not attempt to keep track of previous hidden passages but instead regenerate them when they are needed (in other words, when a rearrangement elsewhere in the genealogy has made formerly meaningless recombinations meaningful).

The probabilities work out as follows. The active lineages are denoted j and the inactive ones k. Then, s_j are the meaningful links on active lineage j and s_k are the (newly) meaningful links on inactive lineage k, weighted in each case by the length of each link:

$$\text{Prob(coal)} = \frac{k(k-1) + 2jk}{\Theta} \tag{3}$$

$$\text{Prob(rec)} = \sum_j rs_j + \sum_k rs_k. \tag{4}$$

The waiting time formula is then

$$\text{time} = \frac{\log(U)}{\text{Prob(coal)} + \text{Prob(rec)}}, \tag{5}$$

where U is a random number chosen uniformly between 0 and 1.

The only step in the rearrangement process that is not driven by the coalescent prior probability is the decision where to make the initial cut. This algorithm is an example of a "reversible jump MCMC" (Green 1995) in that it moves among state spaces with different dimensionality (in this case, numbers of branches, since each recombination adds two branches to the genealogy). As Green points out, care must be taken to preserve detailed balance when making transitions between states of different dimension.

We choose the branch to be cut uniformly among all branches except the root. This creates an imbalance between forward and reverse transitions that change the number of branches, because the more branches a genealogy has, the less likely it is that the specific branch needed to make the reverse change will be chosen. Thus, a move from a specific genealogy with fewer recombinations to a specific genealogy with more recombinations has a higher probability than the reverse move. To restore balance we introduce a Hastings term (Hastings 1970) representing the ratio of forward to back transition. Our probability of acceptance is decided on the basis of the formula

$$\text{Prob(accept)} = \frac{P(D|G_{\text{new}})}{P(D|G_{\text{old}})} \frac{Q_{\text{new}}}{Q_{\text{old}}}, \tag{6}$$

where Q is the reciprocal of the number of branches in that genealogy, excluding the root. If Prob(accept) is ≥ 1 the new genealogy is always accepted; otherwise it is accepted with that probability.

Recombinations near the rootward end of the genealogy may find themselves below the root after a rearrangement; such structures are removed from the genealogy when encountered, since they are not part of our definition of a recombinant genealogy. To avoid irreversibility, once the resimulation process is below the level of the old root, it is necessary to resimulate the root lineage as well as the other lineages, since its recombinations have been stripped out and must be replaced. This is a specialized case of the hidden passages issue: recombinations on the root are "hidden" and must be revealed when that branch is no longer the root.

One problem encountered in practical testing is that for high values of r, the number of recombinations in the sampled genealogies varies widely and may become very large. Such large genealogies present the risk of running out of computer storage space, or slowing the

program unacceptably. An upper limit may be placed on the allowable number of recombinations, but the estimate will be biased downward if this limit is encountered too frequently.

To increase the practical usefulness of the method, it is possible to define as "uninteresting" any recombination involving only sites that have completed coalescing—that is, sites of which only one copy is present in the genealogy at the point of the contemplated recombination. This method was proposed to us by Richard Hudson (personal communication). Recombinations involving such sites can never change the value of $P(D|G)$, so they represent bookkeeping overhead with no informational value. Conceptually simple changes to the rearrangement process and to the evaluation of the prior can eliminate such recombinations. However, these changes are expensive in terms of computer time and space and are only worth using for high values of r. In such cases they reduce the number of recombinations greatly and therefore allow for a more memory-efficient estimate, and possibly a faster one. We refer to this tactic as the "final coalescence" criterion because it eliminates each site from consideration as that site reaches its final coalescence. Use of the final coalescence criterion should not have any systematic effect on the program's results.

Convergence: Results from a sampler of this type are valid only if the sampler has adequately searched the space of genealogies. As with most Metropolis-Hastings samplers, we do not have an exact test to determine when the sampler has run long enough. Several heuristics, however, can help users of the method decide if they are making a large enough sample. Results from the preliminary short chains should stabilize around a value: if each short chain produces a higher (or lower) estimate of the parameters than the one before, the process has not reached equilibrium. Results from several runs with different initial parameters should converge on similar values. For the simulations, we experimented with different initial values and feel that the given results represent an adequate (although not generous) sampling of the search space.

Simulation testing: To test the accuracy of the method we simulated recombinant genealogies of 10 individuals using a program provided by Richard Hudson (personal communication). This program generates independent random genealogies according to the Kingman-Hudson prior. For each genealogy we simulated DNA data using the same mutational model used by our algorithm. All links were 1 unit long, as would be the case for contiguous nucleotide sequence data.

We simulated data under several values of Θ (0.005, 0.01, and 0.1) and r (0, 0.02, 0.04, and 0.08). In most cases we used sequences of length 1001 and 2001. (We chose these numbers because a sequence of length 1001 has 1000 links available for recombination, facilitating comparison with other studies.) For $\Theta = 0.005$ we used only 2001 sites because data sets with 1001 sites were often invariant. For $\Theta = 0.1$ we used 1001 sites and 1501 sites because with 2001 sites we encountered the limits of computer memory at the higher recombination values; we also omitted $r = 0.08$ for this case.

For each parameter combination, we simulated 100 genealogies. Initial values of Θ were estimated using the method of Watterson (1975). The initial value of r was arbitrarily set to 0.01. A random nonrecombinant genealogy was used as the starting genealogy. RECOMBINE executed five short chains of 20,000 rearrangements each followed by two long chains of 50,000 rearrangements each, sampling every 20th genealogy.

Example data set: To observe the performance of the method on biological data we used human lipoprotein lipase (LPL) data (Clark *et al.* 1998; Nickerson *et al.* 1998), which consist of 9734 bp of intron and exon data from 71 individuals (142 chromosomes) derived from three populations: African Americans from Mississippi, Finns from North Karelia, Finland, and non-Hispanic Whites from Minnesota. We considered only nucleotide substitution differences, disregarding insertion/deletion mutations, and we omitted the 20 rare variable sites (rare allele present in only one or two copies) for which Clark's group did not determine phase.

In this analysis we used a transition/transversion ratio of 1.4 based on the ratio of visible transitions to transversions in the full data. We calculated base frequencies on the basis of the polymorphic sites only, and in the subpopulation analyses we used the frequencies from the corresponding subpopulation. We used two rate categories with a ratio of 1:0.75 and prior probabilities of 9:1, based on the Nickerson *et al.* (1998) estimate that 90% of the sequence was noncoding and that the rate in coding regions was 75% of that in noncoding regions. We analyzed only the polymorphic sites using the "reconstituted DNA" SNP model of Kuhner *et al.* (2000) in which sites not noted were assumed to be invariant sites of unknown nucleotide; this was more convenient than analyzing the full DNA sequence and is expected to give similar results.

We analyzed each subpopulation separately using the following strategy: 10 short chains of 50,000 steps and 1 long chain of 1,000,000 steps. Each chain had a burn-in period of 1000 steps and was subsequently sampled every 20th step. We used the Watterson (1975) estimate of Θ as our starting point for Θ and 0.05 as our starting point for r. For the combined data we doubled the length of all chains (to 100,000 and 2,000,000 steps, respectively) to reflect the larger search space of the full data.

RESULTS

Results from simulations of the method in a variety of cases with increasing values of Θ are presented in

TABLE 1

Simulation results

r	$\hat{r}$	SD of $\hat{r}$	$\hat{\Theta}$	SD of $\hat{\Theta}$
		A. $\Theta = 0.005$, 2001 sites		
0.00	0.0310	0.0752	0.0051	0.0020
0.02	0.0400	0.1150	0.0049	0.0020
0.04	0.0666	0.1252	0.0054	0.0025
0.08	0.0903	0.2154	0.0051	0.0021
		B. $\Theta = 0.01$, 1001 sites		
0.00	0.0197	0.0686	0.0102	0.0043
0.02	0.0388	0.0876	0.0102	0.0040
0.04	0.0857	0.1697	0.0098	0.0036
0.08	0.1260	0.2694	0.0096	0.0040
		C. $\Theta = 0.01$, 2001 sites		
0.00	0.0084	0.0549	0.0101	0.0037
0.02	0.0187	0.0397	0.0100	0.0038
0.04	0.0565	0.0898	0.0101	0.0037
0.08	0.0885	0.1156	0.0096	0.0038
		D. $\Theta = 0.1$, 1001 sites		
0.00	0.0007	0.0022	0.0984	0.0328
0.02	0.0229	0.0151	0.1057	0.0337
0.04	0.0412	0.0212	0.1053	0.0352
0.08	0.0794	0.0305	0.1068	0.0291
		E. $\Theta = 0.1$, 1501 sites		
0.00	0.0004	0.0011	0.1028	0.0353
0.02	0.0238	0.0129	0.1037	0.0295
0.04	0.0415	0.0154	0.0995	0.0309

Means and standard deviations of parameter estimates from 100 simulated data sets of 10 sequences each. Presented as true r, mean estimate of r, standard deviation of r, estimate of Θ, and standard deviation of Θ. The majority of replicates were calculated without the "final coalescence" criterion, but it was used in specific cases where the program would otherwise have run out of space and throughout the subtables with $\Theta = 0.1$.

Table 1. For all of the values of Θ examined, estimation of Θ showed little or no bias and good accuracy. The results with $\Theta = 0.01$ and 1001 bp can be compared with results from the nonrecombinant sampler COALESCE (Kuhner *et al.* 1995). The standard deviations are very similar, suggesting that little power is lost by estimating Θ in the presence of recombination. This confirms previous findings of Hudson (1983) for pairwise estimators of Θ.

When the true value of Θ was relatively low, a pronounced upward bias in the estimate $\hat{r}$ was seen (Table 1, A–C), particularly when r was also low. In contrast, for relatively high values of Θ we recovered good estimates of r (Table 1, D and E).

The estimates of r from the low-Θ cases are a mix of many values near zero and a small population of high values associated with likelihood surfaces that are nearly flat (data not shown). These flat likelihood surfaces come from data sets that are nearly uninformative for recombination, probably because they contain few informative sites (polymorphic sites present in more than one individual). The estimates for $r = 0$ are particularly poor because such troublesome cases arise from underlying genealogies that are, by chance, unusually short. It is more likely for the single genealogy produced in a case without recombination to be unusually short than for a collection of recombinant subgenealogies to be unusually short across the entire sequence. By this reasoning, the best way to improve the accuracy of estimation when Θ and r are low should be to increase the number of base pairs surveyed. Indeed, doubling the sequence length from 1001 to 2001 bp produced substantial improvement in the estimate (compare Table 1B with 1C).

Our artificial addition of one recombination to each chain when there would otherwise be none could also contribute to an upward bias in the estimate of r; however, if this correction is not made the results are only slightly lower (data not shown) so we do not believe this to be a major effect.

As a case study we present an analysis of the human LPL data of Clark *et al.* (1998) and Nickerson *et al.* (1998). These data consist of sequences of length 9734 bp from 71 individuals sampled from three populations (African Americans from Mississippi, Finns from North Karelia, and non-Hispanic Whites from Minnesota). These data show visible evidence of recombination and it is thus interesting to assess them with RECOMBINE. However, it should be noted that one of the program's key assumptions (a single nonsubdivided population) is clearly violated by these data whether they are analyzed as one large population or three smaller ones. A more correct analysis would require combining the logic of MIGRATE (Beerli and Felsenstein 1999) with RECOMBINE, a project we are currently undertaking.

We analyzed the three population data sets separately and together; results are shown in Table 2, with estimates of Θ using the method of Watterson (1975) and of r using the method of Hudson (1987) repeated from Clark *et al.* (1998) for comparison.

DISCUSSION

The method described here is able to recover Θ and r with good accuracy in many cases and appears to have practical advantages over existing pairwise methods. It avoids the bias seen when a nonrecombinant sampler is used on recombinant data (Kuhner *et al.* 2000). The computational burden is not excessive: for example, each of the single-population LPL runs took ~4 hr on a 500-MHz Alpha workstation.

Several points are noteworthy in the analysis of the LPL data. In general, RECOMBINE gives somewhat higher estimates of Θ and lower estimates of r than the pairwise methods. It may be better able than the pairwise methods to distinguish between site inconsistencies

TABLE 2

Analysis of LPL sequences

Data set	Haplotypes	$\hat{\Theta}_W$	$\hat{\Theta}_K$	$\hat{r}_H$	$\hat{r}_K$
Jackson	48	0.0018	0.0072	1.443	0.1531
North Karelia	48	0.0013	0.0027	0.371	0.3910
Rochester	46	0.0014	0.0031	0.335	0.2273
Combined	142	0.0016	0.0073	0.693	0.1521

Results for a human lipoprotein lipase data set (Clark *et al.* 1998; Nickerson *et al.* 1998). Shown are the number of haplotypes in each section of the data set, $\hat{\Theta}$ from Watterson's (1975) estimator (Θ_W) and RECOMBINE (Θ_K), and $\hat{r}$ from Hudson's (1987) estimator (r_H) and RECOMBINE (r_K). Non-RECOMBINE results are taken from Clark *et al.* (1998) with permission.

caused by recurrent mutation and site inconsistencies caused by recombination.

The Watterson (1975) estimator of Θ gives very similar values for all subpopulations, whereas RECOMBINE suggests that the Jackson (African-American) sample and the combined data reflect a larger population than the two European samples. This is consistent with many observations of higher diversity in Africans and also with the possibility of admixture in the African-American population. The homogeneous Watterson results reflect the fact that number of variable sites (the only information used by Watterson's estimator) does not vary much among the populations even though overall haplotype diversity is noticeably higher in the Jackson population. RECOMBINE makes fuller use of the available information.

Hudson's (1987) estimator gives a particularly high value of r for the Jackson population, but Clark *et al.* (1998) suggest that this is due to the greater diversity of the Jackson sample, which provides greater detection power. RECOMBINE, in contrast, does not suggest a higher recombination rate for the Jackson sample. Since RECOMBINE can consider recombinations even in regions where the data are uninformative, rather than inferring recombination only where enough variability exists to reveal it, we would expect it not to be misled by differences in polymorphism level.

A few general conclusions about study design are possible. To estimate the recombination parameter r accurately with a sample of 10 individuals, for most organisms it will be necessary to examine tens of thousands of base pairs. Adding more individuals will be only modestly helpful, since most of the new individuals will be closely related to individuals already in the sample and therefore provide little additional genealogical structure. The computational burden of long sequences could be reduced by use of SNPs or some other type of marker rather than full DNA sequences, at some cost in efficiency (Kuhner *et al.* 2000).

Metropolis-Hastings samplers for nonrecombining data are known to be more effective and less biased if multiple unlinked loci, rather than a single locus, are used (Kuhner *et al.* 1998; Beerli and Felsenstein 1999). By contrast, in data containing recombinations maximal efficiency is obtained by having a single locus with long sequences. This captures some of the advantages of multiple loci, since recombination implies different subgenealogies for different parts of the loci, and the use of continuous sequence maximizes the algorithm's ability to detect recombinations.

A number of extensions to the method described here are possible. It could be combined with the population-growth estimator of FLUCTUATE (Kuhner *et al.* 1998) or the migration-rate estimator of MIGRATE (Beerli and Felsenstein 1999) to allow analysis of more complex population structures. Gene conversion could be added as a supplement to conventional recombination, given an appropriate probability model for conversions.

Any form of data for which map information is available and for which a data likelihood model can be developed, such as microsatellite data, electrophoretic allele data, or restriction site data, can in principle be used to infer recombination rate.

The ability to sample clouds of recombinant genealogies can be used for maximum likelihood linkage disequilibrium gene mapping by computing the probability of the observed pattern of a disease trait on the various genealogies under different hypotheses for the disease gene location (M. K. Kuhner and J. Felsenstein, unpublished results).

The program could provide its user with the cloud of sampled genealogies, but this is difficult to comprehend. It would be desirable to combine information about the cloud into a summary genealogy analogous to a consensus tree. However, it is not obvious how to make a consensus of recombinant genealogies. At the moment, only basic summary information, such as the distribution of number of recombinations or of recombination breakpoints, can be collected.

A final addition that would improve the usefulness of the program would be to attempt an estimate of the location of recombination "hot spots" and "cold spots." In theory this could be done by incorporating a Hidden Markov model (HMM) of recombination frequency into the sampler itself, analogously with the use of an HMM in estimating mutation rate variation among sites (Felsen-

stein and Churchill 1996). A simpler approximation would be to use the HMM after the fact, in analysis of the sampled trees, to produce an estimate of the posterior probability that specific intersite links are in one or another recombination rate category.

Availability of software: The Metropolis-Hastings Monte Carlo algorithm described here is available from the authors as program RECOMBINE in the package LAMARC, which uses an input/output format similar to the PHYLIP package. The program is written in C and can be obtained by anonymous ftp from *evolution.genetics.washington.edu* in directory pub/lamarc or via the World Wide Web at http://evolution.genetics.washington.edu/lamarc.html.

We thank Gary Churchill for pointing out the need for a Hastings ratio and Peter Beerli for extensive help in finding the maxima of surfaces. We thank Richard Hudson for providing the recombinant genealogy simulation program and helping us interpret its results, and for suggesting the final coalescence tactic. This research was supported by National Science Foundation grants BSR-8918333 and DEB-9207558 and National Institutes of Health grant 2-R55GM41716-04 (all to J.F.).

LITERATURE CITED

Beerli, P., and J. Felsenstein, 1999 Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics **152:** 763–773.

Clark, A. G., K. M. Weiss, D. A. Nickerson, S. L. Taylor, A. Buchanan *et al.*, 1998 Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am. J. Hum. Genet. **63:** 595–612.

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:** 368–376.

Felsenstein, J., 1993 Phylip (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle. http://evolution.genetics.washington.edu/phylip.html.

Felsenstein, J., and G. A. Churchill, 1996 A hidden Markov model approach to variation among sites in rates of evolution. Mol. Biol. Evol. **13:** 93–104.

Felsenstein, J., M. K. Kuhner, J. Yamato and P. Beerli, 1999 Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data, pp. 163–185 in *Statistics in Genetics and Molecular Biology* (IMS Lecture Notes-Monograph Series, Vol. 33), edited by F. Seillier. Institute of Mathematical Statistics, Hayward, CA.

Fu, Y.-X., 1994 A phylogenetic estimator of effective population size or mutation rate. Genetics **136:** 685–692.

Fukami-Kobayashi, K., and Y. Tateno, 1991 Robustness of maximum likelihood tree estimation against different patterns of base substitution. J. Mol. Evol. **32:** 79–91.

Green, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82:** 711–732.

Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. **3:** 479–502.

Griffiths, R. C., and S. Tavaré, 1993 Sampling theory for neutral alleles in a varying environment. Proc. R. Soc. Lond. Ser. B **344:** 403–410.

Hastings, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57:** 97–109.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

Hudson, R. R., 1987 Estimating the recombination parameter of a finite population without selection. Genet. Res. **50:** 245–250.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. Futuyma and J. Antonovics. Oxford University Press, Oxford.

Kingman, J. F. C., 1982a The coalescent. Stochastic Process. Appl. **13:** 235–248.

Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

Kishino, H., and M. Hasegawa, 1989 Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. **31:** 151–160.

Kuhner, M., J. Yamato and J. Felsenstein, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421–1430.

Kuhner, M. K., J. Yamato and J. Felsenstein, 1998 Maximum likelihood estimation of population growth rates based on coalescent. Genetics **149:** 429–434.

Kuhner, M. K., P. Beerli, J. Yamato and J. Felsenstein, 2000 Usefulness of single nucleotide polymorphism (SNP) data for estimating population parameters. Genetics **156:** 439–447.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 1953 Equations of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087–1092.

Nickerson, D. A., S. L. Taylor, K. M. Weiss, A. G. Clark, R. G. Hutchinson *et al.*, 1998 DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat. Genet. **19:** 233–240.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Communicating editor: G. A. Churchill