Applications of Metropolis-Hastings genealogy sampling

Mary K. Kuhner, Jon Yamato, and Joseph Felsenstein

Department of Genetics, University of Washington

Running head: Metropolis-Hastings sampling

Corresponding author:

Mary K. Kuhner

Department of Genetics SK-50

University of Washington

Seattle, WA 98195, USA

Phone (206) 543-8751

FAX (206) 543-0754

Internet *mkkuhner@genetics.washington.edu*

Abstract

The genealogy of a random sample of sequences from a population provides information on the parameter Θ (neutral mutation rate times effective population size). When the genealogy is unknown, Θ could in principle be estimated by summing over possible genealogies, but there are too many of them. We use a Metropolis-Hastings sampler to concentrate attention on the genealogies of high posterior probability for a given value of Θ; a likelihood curve for nearby values of Θ can be constructed from the sampled genealogies, and its maximum is a maximum likelihood estimate of Θ. This method can in principle be extended to populations with migration and population growth, and to sequences with recombination and selection. It can also be applied to types of data other than nucleotide sequences in cases where an appropriate likelihood model is available.

Introduction

The relationships among individuals from a population can be summarized by a genealogy showing the common ancestors of the individuals and the times at which those ancestors lived. If the individuals were sampled at random from a randomly mating population, the times back to the common ancestors have an approximate expected distribution called the *coalescent* which was derived by Kingman (1982a, b). The coalescent distribution depends only on the number of individuals sampled and the parameter $4N_e\mu$, also called Θ, which combines the effects of effective population size N_e and the neutral mutation rate μ. (We use capital Θ rather than the conventional lowercase θ in this paper because we are considering mutation rate per site, not per locus as in most previous studies.) Using the coalescent approximation, Θ can be estimated from the genealogy. Such an estimate is potentially more efficient than one made from the sequence data without reference to the genealogy (Felsenstein 1992b). Unfortunately the genealogy is generally unknown.

At least three approaches have been previously proposed for estimates of Θ incorporating genealogical information. The method of Fu (1994) makes an estimate based on a single reconstructed genealogy, combined with a correction factor. The method of Tavaré and Griffiths (1993a, b) sums over all (in smaller cases) or a random sample of possible genealogical explanations for the data. The bootstrap Monte Carlo method of Felsenstein (1992a) makes an estimate based on genealogies produced by bootstrapping the original data set. (We have recently found the bootstrap Monte Carlo method to be biassed (Kuhner, Yamato and Felsenstein 1995, submitted) and do not recommend its use.)

The approach proposed in the current paper is importance sampling: we sample among all genealogies, but concentrate on those which are expected to make substantial contributions to the estimate. This can be done by means of a Metropolis-Hastings sampler. New genealogies are created by small modifications of an initial one, and accepted or rejected based on their fit to the data. Samples from this Markov Chain of genealogies can be used to construct a likelihood curve for Θ and to find its maximum. This approach avoids any bias which may be caused by using a single reconstruction of the genealogy, since such reconstructions unavoidably involve some degree of error.

We have designed such a sampler, which can be used for non-recombining nucleotide sequences (such as mitochondrial DNA) from a constant-size, randomly mating population. The details of the sampling procedure are presented in another paper (Kuhner, Yamato, and Felsenstein 1995, submitted): this paper will briefly review the sampler, and discuss ways in which it can be extended to other types of data.

Methods

The Metropolis-Hastings sampler on genealogies is a form of importance sampling: we sample genealogies from a known distribution which we hope will be similar to the unknown true distribution in Θ, in order to gain as much information about Θ per sampled genealogy as possible. In evaluating the genealogies, we then correct for the known distribution from which we sampled.

The known distribution we will use incorporates the probability of the sequence data given the genealogy ($P(D|G)$), and the probability of the genealogy given a chosen value of Θ which we will call Θ_0 ($P(G|\Theta_0)$). The true distribution which we are trying to estimate is identical to this except that the unknown true Θ replaces Θ_0. The importance sampling function is:

$$\frac{P(D|G)P(G|\Theta_0)}{P(D|\Theta_0)}$$

which is the posterior probability of the genealogy. Since only ratios of posterior probabilities are needed, the unknown constant $P(D|\Theta_0)$ need not be considered.

The prior probability, $P(G|\Theta_0)$), can be readily computed using the coalescent approximation of Kingman (1983a, b). The probability with respect to the data, $P(D|G)$, is the quantity maximized by ML methods of phylogeny estimation, and can be computed for a variety of models: we chose the nucleotide sequence model of Felsenstein (described by Kishino and Hasegawa 1989) which allows the specification of the ratio of transitions to transversions and of the frequencies of the four bases. Under Extensions we describe alternative models which could be used.

To carry out this sampling, we construct a new genealogy by modifiying the current genealogy, drawing the new coalescence times from a distribution proportional to $P(G|\Theta_0)$. (This procedure by itself, if repeated many times, would eventually sample all genealogies in proportion to how well they fit coalescent expectations for the given Θ_0.) Having constructed a new genealogy, we decide whether to accept or reject it by comparing the relative probability of the sequence data, $P(D|G)$, on the old and new genealogies. The outcome of this process, repeated many times, is a Markov Chain of genealogies where each possible genealogy is represented in proportion to the product $P(G|\Theta_0)P(D|G)$, its posterior probability. Details of the rearrangment procedure follow.

To make a local rearrangement, we choose a random node in the tree and consider a neighborhood around it: its two descendent nodes, its parent, and its parent's other descendent. A series of neighborhood rearrangements of this type can eventually transform any topology into any other.

The interior of the neighborhood is erased and replaced with a new set of coalescences chosen based on a conditional coalescent distribution. Two factors must be considered: (1) the probability of coalescence at a given moment depends on the number of lineages present in the entire genealogy at that moment; (2) the three lineages within the neighborhood must coalesce with one another, and not with any other lineage, by the end of the neighborhood (this condition limits the process to local rearrangements). We use a modification of the state-array approach of Viterbi (1967). Briefly, we construct a lattice containing the conditional probabilities that three, two or one lineages are present in the neighborhood at each moment, then trace a weighted random path through this lattice beginning with one lineage at the bottom of the neighborhood and ending with three lineages at the top. This path defines a set of coalescence times: the lineages coalescing are then chosen at random from the neighborhood lineages available at the given times. (Complete details of this algorithm are given in Kuhner, Yamato, and Felsenstein 1995, submitted.)

Now that a modified genealogy has been constructed, the probability of the data on the old and new genealogies is evaluated under the chosen model. The decision to accept or reject the new genealogy depends on the ratio r of these probabilities; if $r \geq 1$ the new genealogy is accepted with probability 1, whereas if $r < 1$ it is accepted with probability r (Metropolis et al. 1953). The method of lineage erasure and redrawing does not introduce any bias towards particular topologies, so the corrective terms described by Hastings (1970) need not be computed in this form of the sampler. However, more complex future extensions of the algorithm may require corrective terms to compensate for any tendency to over-propose certain genealogies.

At intervals, genealogies can be sampled from this Markov Chain, and the sampled genealogies can be used to construct a likelihood curve for Θ via the following relationship, which compensates for the use of the importance sampling function:

$$L(\Theta) = \sum_G \frac{P(G|\Theta)}{P(G|\Theta_0)} \tag{1}$$

The Metropolis-Hastings sampler should produce the correct value of $L(\Theta)$ for any value of Θ_0, asymptotically as the number of steps along the Markov chain approaches infinity. In practice, for finite numbers of steps there is a bias towards Θ_0 if it is too distant from Θ, since genealogies informative for the true value of Θ will very rarely be produced. A useful approach is to run several chains, each using as its Θ_0 the maximum of the curve for the previous chain. Genealogies from all chains can be used to make a single estimate by treating them as if they were sampled from a mixture distribution (Geyer 1991).

Applications

The Metropolis-Hastings genealogy sampler was originally designed in order to estimate Θ, but the ability to sample genealogies in proportion to their posterior probability has other potential uses.

Phylogeny estimation. The maximum likelihood method of phylogeny estimation searches for the phylogeny maximizing $P(D|G)$. Maximum likelihood is one of the most effective phylogeny methods (Kuhner and Felsenstein 1994), but it is computationally intensive. Current heuristics such as the one implemented in Felsenstein's DNAML program from PHYLIP are not guarenteed to find the maximum likelihood phylogeny, and they are very slow on large data sets.

The Metropolis-Hastings sampler is not attempting to maximize $P(D|G)$. However, in the course of its search the sampler can easily make note of the highest value encountered, and the genealogy producing it. In practice, this genealogy often has the same topology as the maximum likelihood genealogy. Its branch lengths can then be optimized by standard methods. Thus, the Metropolis-Hastings sampler can be used as a heuristic method for finding the genealogy of maximum likelihod with respect to the data. It is much faster on large data sets (50 sequences and up) than DNAML. We plan to compare their performance on simulated data. Gary Olsen (pers. comm.) reports that DNAML's heuristic frequently becomes trapped in local maxima when analyzing large data sets. Preliminary results suggest that the sampler is less prone to become trapped.

For data drawn from a randomly mating population (as opposed to data representing multiple populations or species) the genealogy of maximum posterior probability at the best value of Θ may in fact be a superior reconstruction of the true genealogy, since the coalescent distribution contributes some information about expected branch lengths. However, this approach is probably not appropriate for interspecific data, since speciation may not follow coalescent dynamics.

Bootstrap-like resampling. The genealogies sampled during a Metropolis-Hastings run with the best value of Θ represent a weighted random sample from the space of all genealogies, and can be used to test hypotheses about that space. For example, they can be used in the same way as a set of bootstrap reconstructions to test the strength of support for a particular branch or rooting. However, since successive steps of the Markov Chain are not independent, care must be taken to sample at sufficiently long intervals that successive sampled genealogies are close to independence. Even with a very long sampling interval, the Metropolis-Hastings sampler is much faster than extant likelihood bootstrapping approaches for medium to large numbers of sequences. It should be feasible to apply it to questions such as the degree of support for an African root in the human mitochondrial DNA data of Vigilant et al. (1991).

Extensions

Other likelihood models. Any likelihood model can be substituted into this algorithm, leaving the genealogy-rearrangement machinery intact. For example, the sequence likelihood model of Yang (1993), which assumes that mutation rates follow a gamma distribution, can be used if an efficient method of computing it is developed. Types of data other than nucleotide sequences can be accomodated: for example, the protein likelihood model of Kishino et al. (1990), or the restriction site likelihood model of Felsenstein (1992c). These models will, however, be more computationally intensive than the current one.

Variable population size. Change in population size changes the expected shape of the coalescent, and thus reconstructions of the genealogy can be used to explore past population size changes. We are currently exploring estimation of an exponential growth rate g. The approach is to assume an initial estimate of g (called g_0) and an initial estimate of the present-day Θ (called Θ_0). Modification of genealogies is then done in much the same way as the original algorithm, except that times are rescaled in terms of g_0. From the collection of sampled genealogies it is possible to make a joint estimate of g and Θ, though there will be bias unless g_0 and Θ_0 are fairly close to g and Θ.

Preliminary results indicate that this method produces an upward bias in estimation of both Θ and g (i.e. the population is estimated to be bigger than it really is, and to have grown more rapidly), but that use of multiple unlinked loci can decrease this bias.

A further possibility in this direction is to allow several different episodes of exponential growth or decline, and attempt to estimate the entire collection of growth rates and time periods. This is complicated by the fact that the model with the largest number of episodes will tend to fit the data best unless the model explicitly penalizes rate changes. It is difficult to suggest a likelihood model incorporating changes in growth rate, since little is known about the statistical distribution of such changes. An alternative approach would be to run independent Markov Chains with different numbers of episodes, and then accept the results of the first chain for which adding a given number of additional episodes does not significantly improve the results.

Recombination. The coalescent distribution is very stochastic, and a single realization of it (such as is provided by mitochrondrial DNA) cannot provide much certainty about the history of the population. Multiple independent samples are desirable, but in order to analyze nuclear DNA we will need to deal with recombination.

Recombinant genealogies can be thought of as collections of ordinary genealogies, each applying to a part of the sequence (Figure 1). There is no difficulty in evaluating the likelihood of sequence data on such a genealogy, since each individual site "sees" a normal genealogy. The major issue that must be resolved is how to make rearrangements that sample among recombinant as well as non-recombinant genealogies in proportion to $P(G|\Theta_0, c_0)$ (where c is the recombination rate per site). The concept of the neighborhood of rearrangement, used in the non-recombinant algorithm, no longer applies, since any lineage can potentially recombine with any other, including lineages not currently represented in the genealogy.

One approach, which we are currently investigating, is to allow two different types of modifications: one type which modifies branching order and length (analogous to the modifications in the non-recombinant version), and one which adds and removes recombinations. Some simplification of the full recombinant coalescence process will probably be required. The non-recombinant coalescent need only consider the probability that three lineages will coalesce into two or one during an interval of time. In a recombinant genealogy, three lineages could split into an unlimited number of lineages via recombination. An explicit solution for this infinite family of probabilities (three lineages become four, five, ...) is probably too difficult. A truncated version that allows only a limited number of new recombinations to occur during a time interval may be an adequate substitute.

Migration. Takahata (1985) explored the coalescent theory of populations that exchange migrants, and found it intractable except for very small cases. The difficulty arises in allowing for all possible combinations of migration events. As an alternative approach, a Metropolis-Hastings sampler which explicitly included migration events as well as coalescences in its genealogies could be used. Such a sampler would allow two types of modification to a genealogy: a change of branching order or length (as in the original algorithm), and a modification of the putative migration events. Such a search could be made fairly efficient by a modification scheme which did not propose migration events inconsistent with the present-day subpopulation structure, although it would still need to search through a very large state space of genealogies with migration.

Selection. General models of selection are probably too complex for the Metropolis-Hastings approach, but some special cases can be analyzed in this fashion. Consider a mutation at a single site which is strongly favored and thus rapidly goes to fixation in the population. Sites which are closely linked to the favored site will also tend to be dragged to fixation with it ("hitchhiking"), while more distant sites will retain polymorphism due to recombination. (In non-recombining DNA a single selected site will draw the entire molecule to fixation, providing no information about which site was selected.) In essence, the selected site sees an extremely small Θ (because the only individuals who contributed to the current population were those carrying the site destined to be fixed) while adjacent sites see a larger and larger Θ approaching the neutral Θ of the genome as a whole. This type of pattern should be detectable with Metropolis-Hastings sampling of recombinant genealogies that explicitly include instances of hitchhiking.

Another potentially tractable type of selection is balancing selection at a specific site. A population could be considered to consist of two subgroups, each with a specific base at the selected site: the two subgroups would be maintained at constant size by natural selection. Mutation at the key site could move an entire sequence from one subgroup to another; recombination could move the part of the sequence not containing the key site. It should be possible to detect which site is defining the subgroups, although this will require the machinery of both recombination and migration.

Disequilibrium mapping. The Metropolis-Hastings approach could be used for a type of disequilibrium linkage mapping. Conventional linkage mapping uses recombinations observed in family data to locate a gene of interest with respect to marker loci. The resolution of conventional mapping is limited: the closer one is to the gene of interest, the rarer are the informative recombinations. Disequilibrium mapping attempts to use information from recombinations occuring throughout the history of the region, which should potentially make many more recombinations available for study.

Current methods of disequilibrium mapping are based on pairwise combinations of loci, losing some of the available information. A Metropolis-Hastings method with recombination could be used for mapping by computing a likelihood function across the sampled genealogies for each possible location of the gene to be mapped. This will presumably require a likelihood model for non-sequence data such as restriction fragment polymorphisms (Felsenstein 1992c), since sequence data across the entire area to be mapped are unlikely to be available.

One problem to be overcome is that a sample with many copies of a rare trait (such as a genetic disease) is likely to be quite different from a random population sample, and this ascertainment bias must be allowed for in constructing the Metropolis-Hastings sampler. Furthermore, if the trait in question is not recessive allowance must be made for uncertainty about which of the affected individual's two chromosomes carries it. However, even if the estimates of recombination rate and population size are biassed by such ascertainment issues, the assignment of the trait to a particular chromosome region may still be made correctly.

Discussion

The Metropolis-Hastings sampler is a relatively efficient way to explore the space of possible genealogies for a given data set. In preliminary simulations (Kuhner, Yamato, and Felsenstein 1995, submitted) we used samples of 100 sequences of length 1000, performing a total of 105,500 steps divided among 7 Markov Chains: the entire estimation took approximately 181 minutes on a DECstation 5000/125, a mid-speed workstation. The run time does not increase quickly with increasing numbers of sequences or sites. Samples of over 800 sequences are feasible on a workstation, although many steps along the chain will be necessary in order to adequately survey the space of plausible genealogies.

In principle the Metropolis-Hastings genealogy sampling approach can be extended to a number of key problems, including variable population size, recombination, migration and selection. The more complex models will probably demand large numbers of relatively long sequences, and possibly multiple loci, in order to simultaneously estimate several parameters, but the recent explosion in sequencing technology should soon make such data sets available. It is interesting to contemplate a "grand unified model" of sequence evolution incorporating all the listed processes, and perhaps insertion, deletion, rearrangment, and interactions among sites as well. Only time will tell whether such a unified model is feasible.

Software availability We are making available a new package, LAMARC, containing our Metropolis-Hastings programs. The first program, COALESCE, implementing the Metropolis-Hastings genealogy sampler for the simplest case (neutral non-recombining sequences in a constant-size, randomly mating population) is available by anonymous ftp from *evolution.genetics.washington.edu* in directory /pub/lamarc. It is written in C. Future programs will also be available in the same directory.

Acknowledgements

We thank the Institute for Mathematics and its Applications for inviting us to the Workshop on Mathematical Population Genetics, supported by funds provided by the National Science Foundation. The research presented here was supported by National Science Foundation grants BSR-8918333 and DEB-9207558 and National Institute of Health grant 2-R55GM41716-04 (all to J. F.).

Literature Cited

FELSENSTEIN, J., 1992a Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. Genet. Res. **60:** 209-220.

FELSENSTEIN, J., 1992b Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. **56:** 139-147.

FELSENSTEIN, J., 1992c Phylogenies from restriction sites, a maximum likelihood approach. Evolution **46:** 159-173.

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

GEYER, C. J., 1991 Estimating normalizing constants and reweighting mixtures in Markov Chain Monte Carlo. Technical Report No. 568, School of Statistics, University of Minnesota.

GRIFFITHS, R. C., and S. TAVARÉ, 1993a Sampling theory for neutral alleles in a varying environment. Proc. R. Soc. Lond. B in press.

GRIFFITHS, R. C., and S. TAVARÉ, 1993b Inference for the infinitely-many-sites model. Genetics in press.

KISHINO, H., T. MIYATA and M. HASEGAWA, 1990 Maximum likelihood ingerence of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. **31:** 151-160.

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

KINGMAN, J. F. C., 1982a The coalescent. Stochastic Processes and Their Applications **13:** 235-248.

KINGMAN, J. F. C., 1982b On the genealogy of large populations. J. Applied Prob. **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. **29:** 170-1790.

KUHNER, M. K., and J. FELSENSTEIN, 1994 A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. in press.

KUHNER, M. K., J. YAMATO, and J. FELSENSTEIN, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics submitted.

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.

TAKAHATA, N., 1985 The coalescent in two partially isolated diffusion populations. Genet. Res. **52:** 213-222.

VIGILANT, L., M. STONEKING, H. HARPENDING, K. HAWKES, and A. C. WILSON, 1991 African populations and the evolution of human mitochondrial DNA. Science **253:** 1503-1507.

Viterbi, A. J., 1967 Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans. Inform. Theory **IT-13:** 260-269.

Yang, Z., 1993 Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10:** 1396-1401.

Figures

Figure 1: A recombinant genealogy. Recombination has occured between sites 9 and 10. The upper genealogy represents the entire sequence; the two lower genealogies show its decomposition into multiple non-recombinant genealogies, each applying to a subset of the sites.