BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 6 2006, pages 768–770
doi:10.1093/bioinformatics/btk051

Genetics and population analysis

LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters

Mary K. Kuhner[1]

[1]Department of Genome Sciences, Box 357730, University of Washington, Seattle, WA 98195-7730, USA

Received on November 15, 2005; revised and accepted on January 9, 2006
Advance Access publication January 12, 2006
Associate Editor: Frank Dudbridge

ABSTRACT

Summary: We present a Markov chain Monte Carlo coalescent genealogy sampler, LAMARC 2.0, which estimates population genetic parameters from genetic data. LAMARC can co-estimate subpopulation $\Theta = 4N_e\mu$, immigration rates, subpopulation exponential growth rates and overall recombination rate, or a user-specified subset of these parameters. It can perform either maximum-likelihood or Bayesian analysis, and accomodates nucleotide sequence, SNP, microsatellite or elecrophoretic data, with resolved or unresolved haplotypes. It is available as portable source code and executables for all three major platforms.

Availability: LAMARC 2.0 is freely available at http://evolution.gs.washington.edu/lamarc

Contact: lamarc@gs.washington.edu

1 INTRODUCTION

Inference of population parameters (such as effective population size, growth rate or immigration rate) from sequence data is often done using summary statistics in order to avoid dealing with the unknown genealogy relating the sampled sequences. Such genealogies are difficult to infer accurately, and nearly impossible in cases with recombination.

The Lamarc package addresses this difficulty by approximate integration over the space of possible genealogies using Markov chain Monte Carlo (MCMC) sampling. This avoids both the loss of power from using summary statistics and the difficulty of inferring the true genealogy. Previous programs in the package include COALESCE (Kuhner *et al*., 1995), estimating $\Theta = 4N_e\mu$ and several programs co-estimating Θ and one additional type of parameter: FLUCTUATE (exponential growth rate) (Kuhner *et al*., 1998), MIGRATE (immigration rates) (Beerli and Felsenstein, 1999), (Beerli and Felsenstein, 2001) and RECOMBINE (recombination rate) (Kuhner and Felsenstein, 2000; Kuhner *et al*., 2000a).

When multiple evolutionary forces act on a population, analyzing them one at a time may lead to bias and loss of power. We have developed an integrated program, LAMARC 2.0, which can infer multiple forces simultaneously for greater accuracy.

Previous Lamarc package programs have performed maximum-likelihood analysis, using the sampled genealogies to construct a likelihood surface for the parameters of interest. LAMARC 2.0 retains this capability, but can also perform a Bayesian analysis in which the sampler searches among parameter values as well as genealogies.

2 ALGORITHM

2.1 Statistical approaches

LAMARC's maximum-likelihood estimation uses a set of driving values, working values of the population parameters, to construct an importance sampling function which will guide the search among genealogies. This procedure can be inefficient for finding the maximum-likelihood estimates (MLEs) of the parameter values unless the driving values are close to the unknown true parameters, so the search is iterated using the previous estimates as new driving values.

Bayesian estimation searches simultaneously among genealogies (guided by the current working values of the population parameters) and among values of the population parameters (guided by the current genealogy). Most probable estimates (MPEs) and credibility intervals are produced by recording the parameter values visited by the search and doing one-dimensional curve-smoothing to obtain the posterior probability curve for each parameter.

For both forms of analysis, LAMARC estimates parameters for each unlinked genomic region separately, as well as a joint estimate over all regions.

2.2 Evolutionary models

LAMARC estimates $\Theta = 4N_e\mu$, where N_e is the effective diploid population size and μ is the neutral mutation rate per site per generation. (The estimated Θ can also be interpreted in a haploid, mitochondrial or alternative ploidy context.) It can co-estimate the exponential growth rate g. In subdivided populations it estimates Θ and optionally g for each subpopulation, and immigration rate into each subpopulation from each of the others. Finally, it can optionally estimate the overall recombination rate $r = c/\mu$, where c is the recombination chance per site per generation. Customized models where specific rates are omitted, held constant or forced to be equal to one another are possible for all parameters.

3 IMPLEMENTATION

3.1 Search strategy

LAMARC 2.0 provides several mechanisms to improve its search efficiency. Metropolis-Coupled MCMC or 'heating' allows auxiliary searches with more permissive acceptance criteria to act as 'scouts' for the main analysis (Geyer, 1991a). For likelihood analyses, multiple replicated searches can be combined using reverse logistic regression (Geyer, 1991b). For Bayesian analyses, the Bayesian priors and the ratio of parameter change steps to genealogy change steps can be set by the user.

3.2 Mutational models

LAMARC 2.0 offers the Felsenstein 84 (F84) and General Time-Reversible (GTR) models for DNA or RNA data, and for SNP data when information about the total sequence length surveyed is available. The SNP model used is correct only if all variable sites in the data have been captured; there will be an ascertainment bias if SNPs were surveyed based on their presence in an external panel. Multiple substitution rate categories (including an invariant category) and potential autocorrelation between rates at adjacent sites are accomodated using a hidden Markov model (Felsenstein and Churchill, 1996).

For microsatellites, four models are available: a stepwise mutation model (Ohta and Kimura, 1973); a Brownian-motion approximation to the stepwise model (Beerli and Felsenstein, 2001) which is much faster, but may be inaccurate when polymorphism is low; a K-allele model and a mixture model of the stepwise and K-allele models, with the mixture parameter potentially optimized based on the data. The K-allele model is also suitable for analyzing elecrophoretic data.

Separate genetic regions with different forms of data (e.g. a DNA locus and an unlinked microsatellite locus) may be combined in a single analysis. The user must provide information on the expected relative μ and/or N_e of the various regions if they differ. For example, mitochondrial and nuclear DNA may be combined in one analysis, but the program must be informed of the expected 4× difference in N_e.

3.3 Haplotype uncertainty

Phase-unknown data may be used, although they are less powerful than phase-known data. The genealogy search is extended to search among haplotype resolutions as well, so that the estimate takes into account haplotype uncertainty as well as genealogy uncertainty (Kuhner and Felsenstein, 2000; Kuhner *et al*., 2000b).

3.4 Availability

LAMARC 2.0 is freely distributed as portable C++ source code and as executables for Windows, Mac OSX and Linux. It provides a utility to convert PHYLIP, RECOMBINE and MIGRATE input files. The file converter's graphical user interface uses a multiplatform windowing system which works on all three major platforms, but a pure text file converter is also available. The major requirements for the use of LAMARC are availability of memory and time. For example, estimation of recombination rate using 60 16 kb mtDNA sequences required 2 GB of memory and 3–4 weeks of workstation time. Smaller analyses will often take 1–2 days.

4 DISCUSSION

4.1 Model assumptions

LAMARC 2.0 assumes that individuals are drawn from panmictic subpopulations and that the subpopulation structure has been constant throughout the lifespan of the underlying coalescent tree. It is not suitable for populations which have recently diverged from a common ancestor. It assumes that the rate at which a lineage immigrates into a population is independent of the size of both source and recipient populations. It also assumes that exponential growth rates and immigration rates have been constant throughout the lifespan of the coalescent tree and that recombination rate does not vary by position, subpopulation or with time. Finally, it assumes that the variation being observed is neutral, though purifying selection removing harmful mutations does not disrupt the analysis much.

Violation of these assumptions will potentially result in biased estimates and inaccurate confidence intervals.

4.2 Bayesian versus likelihood analysis

In most cases examined so far (Kuhner and Smith, manuscript submitted) LAMARC's Bayesian and likelihood methods produce similar point estimates and confidence intervals. The Bayesian method is vulnerable to a poor choice of priors, but with good priors it may search among genealogies more efficiently, especially in cases where one or more parameters are close to zero. Our current curve-smoothing method does not allow the Bayesian algorithm to assess correlation among parameters, whereas the likelihood algorithm can. Speed requirements of the two methods are similar; the Bayesian sampler must perform more search steps, but its curve-smoothing is faster than likelihood maximization.

4.3 Data requirements.

LAMARC 2.0 assumes that individuals are sampled randomly within each subpopulation, but it does not require equal sample sizes among subpopulations.

If some subpopulations are not genetically differentiated ($4N_em$ much greater than one) results will be unsatisfactory. Such subpopulations are best pooled into a single subpopulation.

A sample of 20 individuals per subpopulation is fully adequate and results are often satisfactory with as few as eight, especially if multiple loci are available. For estimation of any parameter except recombination rate, adding unlinked loci will improve the estimate more than adding individuals or lengthening sequences. For estimating recombination rate, lengthening sequences or adding linked loci are preferable.

4.4 History

LAMARC 1.0 was released in 2001. LAMARC 2.0 corrects several deficiencies in the previous versions, particularly errors in likelihood maximization and handling of multi-locus data. LAMARC 2.0 adds Bayesian analysis, the ability to constrain parameters and new mutational models.

ACKNOWLEDGEMENTS

The author thanks the *Lamarc* team, including Peter Beerli, Joseph Felsenstein, Eric Rynes, Lucian Smith, Elizabeth Walkup, Jon Yamato and Wang Yi. Development of this program was supported by NIH grant 5R01GM51929-11 to M.K. Funding to pay the Open Access publication charges was provided by the National Institutes of health grant 5R01CM51929-11 to M.K.

Conflict of Interest: none declared.

REFERENCES

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

Beerli,P. and Felsestein,J. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations using a coalescent approach. *Proc. Natl Acad. Sci. USA*, **98**, 4563–4568.

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

Geyer,C.J. (1991a) Markov chain Monte Carlo maximum likelihood. In Keramidas (ed.), *Computing Science and Statistics: Proceedings of 23rd Symposium on the Interface,* Interface Foundation, Fairfax Station, pp. 156–163.

Geyer,C.J. (1991b) Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. *Technical Report No. 568,* School of Statistics, University of Minnesota, MN revised 1994.

Kuhner,M.K. and Felsenstein,J. (2000) Sampling among haplotype resolutions in a coalescent-based genealogy sampler. *Genet. Epidemiol.*, **19** (Suppl. 1), S15–S21.

Kuhner,M.K. *et al.* (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, **140**, 1421–1430.

Kuhner,M.K. *et al.* (1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, **149**, 429–434.

Kuhner,M.K. *et al.* (2000a) Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics*, **156**, 439–447.

Kuhner,M.K. *et al.* (2000b) Maximum likelihood estimation of recombination rates from population data. *Genetics*, **156**, 1393–1401.

Ohta,T. and Kimura,M. (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.*, **22**, 201–204.