BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–5

Comparison of Bayesian and maximum likelihood inference of population genetic parameters

Peter Beerli

School of Computational Science and Department of Biological Sciences, Florida State University, Tallahassee FL 32306-4120

ABSTRACT

Comparison of the performance and accuracy of different inference methods, such as maximum likelihood and a Bayesian inference, is difficult because the inference methods are implemented in different programs often written by different authors. I implemented both methods in the program MIGRATE, that estimates population genetic parameters, such as population sizes and migration rates using coalescence theory. Both inference methods use the same Markov chain Monte Carlo algorithm and differ from each other in only two aspects from each other: parameter proposal distribution and maximization of the likelihood function. Using simulated data sets, the Bayesian method generally fares better than the ML approach in accuracy and coverage. Although for some values the two approaches are equal in performance.

Motivation: The Markov chain Monte Carlo-based maximum likelihood framework can fail on sparse data and can deliver non-conservative support intervals. A Bayesian framework with appropriate prior distribution is able to remedy some of these problems.

Results: The program MIGRATE was extended to allow not only for maximum likelihood based estimation of population genetics parameters but also to use a Bayesian framework. Comparisons between the Bayesian approach and the ML approach are facilitated because both modes estimate the same parameters under the same population model and under the same assumptions.

Availability: The program is available from http://popgen.csit.fsu.edu.

Contact: beerli@csit.fsu.edu

1 INTRODUCTION

Population genetics changed considerably after Kingman (1982*b*,*c*,*a*) introduced the *n-coalescent*. The *n-coalescent* (coalescent for short) allows us to calculate probabilities of relationships among a random population sample. This in turn facilitates calculations of probabilities of whole genealogies under a specific population model, for example two populations exchanging migrants at a constant rate. The first applications that calculated the likelihood of the population size parameter based on DNA samples were described by Griffiths and Tavaré (1994) and Kuhner *et al.* (1995). Bahlo and Griffiths (2000) and Beerli and Felsenstein (1999, 2001) extended the basic estimation of a single parameter to joint estimations of migration rates and population sizes, whereas Kuhner *et al.* (2000) allowed for the estimation of recombination rate. These maximum likelihood approaches were complemented by several Bayesian approaches (Nielsen, 1998, 2000; Hey and Nielsen, 2004; Beaumont, 1999, and others). All of these approaches try to estimate population genetic parameters. They typically treat the genealogy as a nuisance parameter and summarize over all possible genealogies G; to be precise, they sample over all possible labeled histories T and branch lengths B, taking into account the genetic data and the population genetic model. The likelihood of the data given the model parameters is

$$\mathrm{L}(D|\pi) = \sum_{T} \int_{B} k(T,B|\pi)\mathrm{L}(D|T,B)dB \qquad (1)$$

where $k(T,B|\pi)$ is the Kingman coalescent probability density and $\mathrm{L}(D|T,B)$ is the likelihood of the data given the genealogy.

Nielsen (pers. comm., 2001) suggested that the maximum likelihood approach is hampered by several problems. Maximizing the likelihood function $\mathrm{L}(D|\pi)$ for complicated scenarios with many parameters is VERY difficult. The Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970) with static driving values π_0, as implemented in MIGRATE and other programs can take a prohibitively long run time required to full explore all possible genealogies. These problems have been shown by Abdo *et al.* (2004), although Abdo *et al.* apparently failed to recognize that the problems are far less serious when using biologically reasonable data sets and when the guidelines about convergence outlined in the MIGRATE-manual (available from http://popgen.csit.fsu.edu) are followed.

2 APPROACH

The program MIGRATE uses a Metropolis-Hastings algorithm to explore all possible genealogies (Beerli and Felsenstein, 1999). The adaptation of the program to a Bayesian

framework was not difficult because only a module handling the prior distributions and a minor change in the program flow needed to be added together with changes in the input and output user interfaces.

The program MIGRATE calculates the posterior probability distribution per locus, treating each locus as completely unlinked to the others. This assumption is reasonable because most biologists would prefer to sample such loci rather than partially linked or completely linked loci because unlinked loci can be treated as independent replicates of the genealogical history. MIGRATE approximates the posterior distribution

$$\mathrm{f}(\pi|D) = \frac{r(\pi)\int_G k(G|\pi)\mathrm{L}(D|G)dG}{\mathrm{P}(D)} \tag{2}$$

using a Metropolis-Hastings approach. The integral over G is a condensed expression of the sum over topologies and integral over all branch lengths. The denominator is

$$\mathrm{P}(D) = \int_{\pi\in\Omega} r(\pi)\int_G k(G|\pi)\mathrm{L}(D|G)dGd\pi$$

where we integrate over all possible parameter values π.

The updating scheme of the genealogies is the same in the ML and the Bayesian approach and was described by Beerli and Felsenstein (1999). The updating scheme of the parameters is based on arbitrary prior distributions $r(\pi)$. MIGRATE allows the user to choose between a small number of prior distributions

- Uniform prior distribution between a minimum and a maximum value for each parameter;
- Exponential prior distribution with a minimum, mean, and maximum value for each parameter;

The incorporation of additional prior distributions, such as a gamma distribution, are planned.

A key issue in Metropolis-Hastings algorithms is the acceptance or not of a change of the current state in the Markov chain. The algorithm should accept fairly often so that the chain can explore the solution space more efficiently; poor algorithms will reject often and force very long runs to achieve equilibrium and an appropriate sample of the possible states. Typically, the acceptance or rejection of a move in the Markov chain is based on a ratio that consists of two parts: (1) the ratio of probabilities to move from an old state to a new state using a prior distribution and the effect of the data (Metropolis *et al.*, 1953), the Metropolis ratio r_M; (2) the ratio of probabilities to be in the old or new state and go to the new or old state (Hastings, 1970), the Hastings ratio, r_H. In the Bayesian implementation in MIGRATE the ratio of accepting a move suggested by the parameter prior is only dependent on the Kingman coalescent probability density.

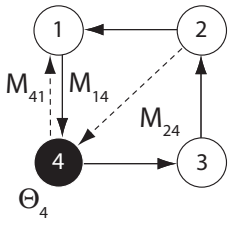

Fig. 1. Population scenario used in the example: Four populations exchange migrants unidirectionally as follows: from population 1 to 4 (M_{14}), from 4 to 3 (M_{43}), from 3 to 2 (M_{32}) and from 2 to 1 (M_{21}). Parameters are scaled effective population sizes Θ_i ($4\times$ effective population size $\times$ mutation rate per site per generation), and scaled immigration rates M_{ji} (immigration rate divided by mutation rate). Migration along routes indicated by solid arrows was simulated using "true" values of $M = 100$; migration along all eight other migration routes was simulated with a value of M=0. Migration along the dashed arrows are discussed in the Result section

The acceptance/rejection ratio is

$$\mathrm{r} = r_M r_H = \frac{r(\pi_i^{(n)})k(G|\pi_i^{(n)})\mathrm{L}(D|G)}{r(\pi_i^{(o)})k(G|\pi_i^{(o)})\mathrm{L}(D|G)}\frac{\mathrm{prob}(\pi_i^{(o)}|\pi_i^{(n)})}{\mathrm{prob}(\pi_i^{(o)}|\pi_i^{(n)})}, \tag{3}$$

which reduces to

$$\mathrm{r} = \frac{k(G|\pi_i^{(n)})}{k(G|\pi_i^{(o)})}. \tag{4}$$

If we consider the uniform random prior distribution (URP) then

$$r(\pi_i^{(n)}) = r(\pi_i^{(o)}) \tag{5}$$

and the Hastings ratio r_H will turn into

$$\frac{\mathrm{prob}(\pi_i^{(o)}|\pi_i^{(n)})}{\mathrm{prob}(\pi_i^{(n)}|\pi_i^{(o)})} = \frac{r(\pi_i^{(o)})}{r(\pi_i^{(n)})} = 1. \tag{6}$$

For the exponential-prior distribution a similar logic applies, although moving from $\pi_i^{(o)}$ to $\pi_i^{(n)}$ versus from $\pi_i^{(n)}$ to $\pi_i^{(o)}$ will not have equal probability as with the URP (6). In this case the prior probabilities in the Hastings-ratio will cancel with the prior probabilities in the Metropolis ratio (formula 3).

I illustrate the performance of the improvements on a data set for four populations with a unidirectional migration pattern (Figure 1). Simulated DNA sequence alignments, generated using the population model described in figure 1, were analyzed to show the performance of the Bayesian and the ML approach. One data set with 10 loci, and 4 groups of 100

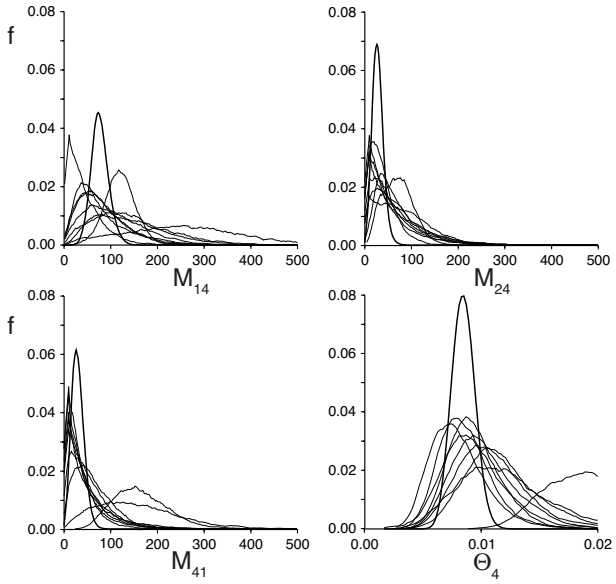

Fig. 2. Posterior distributions f estimated using exponential priors: expected mode for the scaled migration rate M_{14} is 100, expected modes for M_{41} and for M_{24} are zero, expected mode for the scaled effective population size Θ_4 is 0.01. The posterior distributions of ten independent loci (thin lines) and the combined posterior distribution (thick line) are shown. The relationship among the populations is explained in Figure 1

single locus data sets were analyzed. Each data set contained 20 individuals from each of the four populations. Using a coalescence-based simulator (cf. Hudson, 2002) "true" genealogies using population sizes (Θ_T) for all populations of $0.1, 0.01, 0.001$, and 0.0001 and M_{ji} referenced in Figure 1 were created. DNA sequences of 10,000 bp length were then simulated on this true genealogies using an F84 model with equal base frequencies and transition/transversion ratio of 2.0. These data sets were then analyzed using either the maximum likelihood (ML) inference mode (Beerli and Felsenstein, 1999, 2001) or the Bayesian the inference mode in MIGRATE. The ML mode was run for 10 short chains visiting $100,000$ genealogies and storing 5000, updating the driving parameter after each chain, and 2 long chains with $10,000,000$ visited genealogies and sampled $50,000$ using an adaptive heating scheme. The Bayesian inference was run for $10,000,000$ updates, approximately half of which were updates of the 16 parameters and approximately half ($\sim 5,000,000$ because of random switching between genealogy and parameter updates) were genealogy updates. New parameters were proposed using an exponential prior distribution with population size mean of $2\Theta_T$ and boundaries of $\Theta_T/10$ and $10\Theta_T$, and scaled migration rate mean M of 200 and boundaries of 0 and 1000. Results for uniform priors with the same boundaries were very similar, and therefore are not shown.

I show results and problematic issues only for population 4, but the pattern is identical for the three other populations. The scenario chosen for an example is difficult for any gene flow estimator because it requires the estimation of 12 migration rates and 4 population sizes. With high migration rates, haplotypes are distributed evenly over all populations, so that establishing the directionality of gene flow from estimated migration rates is difficult. With low migration rates, however, the difference from zero, and thus the directionality, is difficult to establish. The number of variable sites or the number of alleles in the data set is crucial for accurate estimation of population size and migration rates of any magnitude. Single locus data sets with low variability do not allow estimating migration rates with great precision.

Despite these difficulties, with sufficient data, estimates are expected to be useful for inferring direction and magnitude of gene flow and magnitude of population size. Using the 16-parameter model analyzed here will produce very variable parameter estimates from single locus data, however, and such analyses are not advisable for real biological data.

2.1 Multilocus analysis:

Figure 2 shows that the variability of individual loci resulting from the coalescent and difficulties in reaching convergence can be large, but the combined estimate over all loci gives a rather accurate picture. The variability for migration rate estimates is much larger than for the population size estimates. It is difficult to establish the gene flow direction (M_{41} versus M_{14}) for the single locus estimates. The estimate over all loci clearly allows the distinction between the two directions: M_{14} is much bigger than M_{41}. The estimation of migration parameter values between populations with no direct connections, for example migration rate rate M_{24} between population 2 and 4, is consistently low (Figure 2).

2.2 Comparison of Bayesian and maximum likelihood inference

MIGRATE allows direct comparison of the success of parameter inference using the Bayesian approach and the maximum likelihood approach. In theory the results should be very similar. Table 1 shows medians and quartiles of 100 single locus runs. I chose medians and quartiles because they are a better indicator of the distribution of the results than mean and standard deviation because these are heavily influenced by large outliers. The median of the maximum posterior probabilities is similar to the median of the maximum likelihood estimates for moderate values of the population size ($\Theta = 0.01$). The results for the low variability data sets are mixed; the medians of the two methods are still comparable but the range of the quartiles of ML M estimates are very large, standard deviations (not shown) were even larger because of outliers in the ML analysis. Several of the 100 runs reported values that were very different from the true value. The data sets with the smallest true $\Theta(0.0001)$

Table 1. Medians and quartiles of 100 single-locus data sets for the two inference methods (I): maximum likelihood (M) and Bayesian (B). Simulated data sets that were generated with 4 different values of "true" population sizes ($\Theta^{(t)}$). Θ is $4\times$ effective population size $\times$ mutation rate per site per generation, and M is immigration rate over mutation rate. The range of number of migrants per generation $N_e m = \Theta M/4$ covers a wide range from 0.0025 (corresponding to a $\Theta = 0.0001$) to 2.5 (corresponding to a $\Theta = 0.1$) migrants per generation. Run conditions for ML and Bayes inferences are specified in the text.

$\Theta_4^{(t)}$	$M_{14}^{(t)}$	I	Θ_4			M_{14}		
			25%	Med	75%	25%	Med	75%
0.0001	100	M	0.0004	0.00092	0.0028	0.0	0.2	643.6
		B	0.00006	0.00009	0.00013	7.0	9.0	41.0
0.001	100	M	0.0010	0.0017	0.0036	0.0	46.3	171.5
		B	0.0013	0.0015	0.0017	65.0	79.0	117.0
0.01	100	M	0.0089	0.0104	0.0128	20.0	53.7	108.1
		B	0.0085	0.0101	0.0012	63.0	90.0	125.0
0.1	100	M	0.0295	0.0573	0.0825	36.1	66.5	100.5
		B	0.0698	0.0891	0.1143	45.5	69.0	116.5

shows even more problems with the ML approach because the medians for Θ is strongly overestimated and the range of the quartiles for M is huge. In contrast to ML the Bayesian runs recover the population size, but report very low values for the migration rate. Figure 3 shows a comparison of posterior distributions of the scaled migration rate M of the first data sets of each population size category (Table 1). The power to make inferences about the magnitude of the migration rate is directly correlated with the magnitude of the population size. For very small population sizes there is no power to estimate such low migration rates in the chosen 16 parameter problem with a single locus data set of 10,000 bp for each of the 100 individuals. The posterior distribution is similar to the exponential prior distribution used. In contrast to the problems encountered in the migration rate estimations, the posterior distributions for Θ are strongly peaked near 0.0001 (results not shown).

The ML method has difficulty recovering the expected values when the data set is very variable, whereas the Bayesian inference is closer to the "true" values for all scenarios. The range of the quartiles of the ML approach is often much larger than the range of the Bayesian approach.

The coverage of the Bayesian approach is rather conservative and includes the "true" values in the 95% credibility interval with frequencies of 0.85 to 1.00 for the migration and population size parameters (Table 2), whereas the ML approach has difficulty with convergence, especially on low variability data sets, and so has a rather low coverage (frequencies between 0.06 and 0.94).

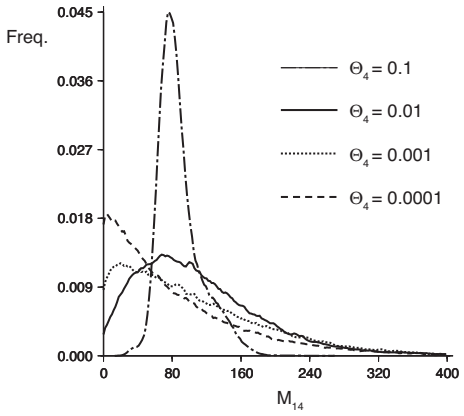

Fig. 3. Posterior distribution of the scaled migration rate M_{14} for four different values of Θ_4 of a single locus data set. The population model is explained in Figure 1. Graphs are results from the first replicate of the four replicate groups shown in Table 1. Data was simulated with $M_{14} = 100$.

Table 2. Coverage of maximum likelihood and Bayesian inferences using simulated data sets that were generated with 4 different values of "true" population sizes ($\Theta^{(t)}$). Θ is $4\times$ effective population size $\times$ mutation rate per site per generation, and M is immigration rate over mutation rate. Coverage is measured as the percentage of times the true value is within the estimated 95% support interval. Run conditions for ML and Bayes inferences are specified in the text.

$\Theta_4^{(t)}$	$M_{14}^{(t)}$	Coverage [%]			
		Θ_4		M_{14}	
		ML	Bayes	ML	Bayes
0.0001	100	6	98	33	100
0.001	100	47	100	55	99
0.01	100	94	96	62	96
0.1	100	51	91	49	85

3 DISCUSSION

The scenario chosen for an example is difficult for any gene flow estimation program that uses only a single sample in time. The problem stems from the fact that the only information about the directionality are the mutations in the data set. If the migration rate is high, all mutations, even the rare ones, are distributed over all populations and any directionality estimation based on a single locus will fail. With low migration rates among the populations, each population will acquire unique mutations and in principle the magnitude and directionality can be estimated even for single locus data sets, if there is enough variability in the data set. In reality, however, such an estimation has proven difficult because the difference between the migration rates between two populations is small and often close to zero. Estimate based on

single locus data sets thus often cannot recover the directionality, but multilocus estimates will allow the inference of the migration direction (Figure 2.

The power to estimate migration rate is crucially dependent on the number of variable sites or number of alleles in the data set. Too little variation leads to haphazard results in the ML method because the MCMC process has no strong guidance whether to insert or remove migration events during the course of the analysis; the process is more dependent on the static driving parameters. Comparison of several runs will deliver very different results and therefore show non-convergence. The only remedy is to run these analyses much longer to get a better estimate of the uncertainty of the estimate. Bayesian analysis is straightforward in such cases because when the posterior distribution is similar to the prior distribution, we can conclude that the data set does not contain enough information for the inference. The ML method also has difficulties exploring the distribution around the maximum likelihood estimate with highly variable data because the genealogy is very well defined by the large number of variable sites: the static driving value and the updating scheme (Beerli and Felsenstein, 1999) will not explore many different migration scenarios and therefore the tails of the distribution are not visited. This results in too narrow support intervals with small coverage values. In contrast, Bayesian inference manipulates the parameters using a diffuse prior. This forces more changes of the genealogy, therefore exploring more different migration scenarios and visiting the tails of the posterior distribution more efficiently.

The coverage shown for the Bayesian runs might be conservative but this is preferable to the coverage reported for ML, especially in the low variability data sets ($\Theta \leq 0.001$, Table 2). Some ML-runs did not really converge and were estimating either very large or zero migration rates.

4 CONCLUSION

Many users of MIGRATE have reported in numerous email queries that achieving convergence with the ML approach with low-information data, such as single locus data sets or data with a low mutation rate, is difficult and needs special attention. Bayesian inference seems to allow such users to achieve reliable results with less effort than the ML approach. It seems appropriate that if only the parameters and their support interval are of interest, then biologists should prefer the Bayesian approach, although it will be interesting to see whether this will hold for all biological data sets.

ACKNOWLEDGEMENT

I thank Rasmus Nielsen for suggesting that I implement a Bayesian method into MIGRATE. At first, I was uncertain about its success for complicated scenarios, like the one used (Figure 1). Thomas Uzzell, Koffi Sampson, and two anonymous reviewers helped to improve the manuscript. Funding for this research was supplied through Florida State University and National Science Foundation grant DEB-0108249 to Scott Edwards and PB.

REFERENCES

Abdo,Z., Crandall,K.A. and Joyce,P. (2004) Evaluating the performance of likelihood methods for detecting population structure and migration. *Molecular Ecology,* **13** (4), 837–851.

Bahlo,M. and Griffiths,R.C. (2000) Inference from gene trees in a subdivided population. *Theoretical Population Biology,* **57** (2), 79–95.

Beaumont,M.A. (1999) Detecting population expansion and decline using microsatellites. *Genetics,* **153** (4), 2013–29.

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** (2), 763–73.

Beerli,P. and Felsenstein,J. (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the United States of America,* **98** (8), 4563–4568.

Griffiths,R. and Tavaré,S. (1994) Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society London B Biological Sciences,* **344** (1310), 403–10.

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

Hey,J. and Nielsen,R. (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis. Genetics,* **167** (2), 747–760.

Hudson,R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics,* **18**, 337–338.

Kingman,J.F.C. (1982*a*) Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics*, (Koch,G. and Spizzichino,F., eds),. North-Holland, Amsterdam p. 97112.

Kingman,J.F.C. (1982*b*) On the genealogy of large populations. *Journal of Applied Probability,* **19A**, 27–43.

Kingman,J.F.C. (1982*c*) The coalescent. *Stochastic Processes and their Applications,* **13**, 235–248.

Kuhner,M.K., Yamato,J. and Felsenstein,J. (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics,* **140** (4), 1421–30.

Kuhner,M.K., Yamato,J. and Felsenstein,J. (2000) Maximum likelihood estimation of recombination rates from population data. *Genetics,* **156** (3), 1393–1401.

Metropolis,N., Rosenbluth,A.W., Rosenbluth,M.N., Teller,A.H. and Teller,E. (1953) Equation of state calculations by fast computing machines. *Journal of Chemical Physics,* **21**, 1087–1092.

Nielsen,R. (1998) Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Journal of Theoretical Population Biology,* **53** (2), 143–151.

Nielsen,R. (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics,* **154** (2), 931–942.