3

How to use MIGRATE or why are Markov chain Monte Carlo programs difficult to use?

PETER BEERLI

Population genetic analyses often require the estimation of parameters such as population size and migration rates. In the 1960s, enzyme electrophoresis was developed; it was the first method to gather co-dominant data from many individuals in many populations relatively easily. Summary statistics methods, such as allele-frequency based F-statistics (Wright 1951), were used to estimate population genetics parameters from these data sets. These methods matured and expanded into many variants that were enthusiastically accepted by many researchers. F-statistics are still a hallmark of any population genetic study, especially in conservation genetics, although over the years, limitations have become evident (Neigel 2002). Many of these methods use restrictive assumptions, for example, disallowing mutation. F-statistics, such as F_{ST} methods, are often employed on pairs of populations; this can lead to biased parameter estimates (see Beerli 2004; Slatkin 2005) and the reuse of data in these pairwise methods is undesirable from a statistical viewpoint.

In 1982, Sir John Kingman developed the coalescence theory (Kingman 1982a, b). His overview of the developments of this theory (Kingman 2000) gives an interesting insight into the development of new ideas. This new development opened the door to methods in population genetics that go beyond the F-statistics methods and have led to several theoretical breakthroughs (Hein et al. 2005; although inferences based on coalescence theory were not practicable until about 1995 because of computational constraints). In recent years, computer-intensive programs that can estimate parameters using genetic data under various coalescent models have been developed; for example, programs that estimate gene flow (Beerli and Felsenstein 1999, 2001; Bahlo and Griffiths 2000; Wilson et al. 2003; De Iorio and Griffiths 2004; Hey and Nielsen 2004; Beerli 2006; Ewing and Rodrigo 2006; Kuhner 2006). These programs use different

Population Genetics for Animal Conservation, eds. G. Bertorelle, M. W. Bruford, H. C. Hauffe, A. Rizzoli and C. Vernesi. Published by Cambridge University Press.

models and different approaches, but in all of them, the quantities of interest are difficult to calculate. Very generally, the goal of these applications is to calculate the probability of the parameters of the chosen model given the data. Population genetics methods often use the relationship among the sampled individuals to get accurate estimates of population size, migration rate or other parameters. These relationships, called genealogies, are typically unknown. Therefore, an optimal approach is to look at all genealogies and weight them using the data. Such approaches can be expressed as integrals over all possible relationships. Unfortunately, there are too many possible genealogies and such an integral cannot be solved exactly. Several numerical integration methods have been developed over the centuries, but only recently Metropolis *et al.* (1953) developed a general approach allowing the integration of complicated multidimensional functions and named this approach the 'Markov chain Monte Carlo method'. Their original algorithm, the Metropolis algorithm, was extended by Hastings (1970) and Green (1995). Many coalescence-based programs use the Metropolis–Hastings or the Metropolis–Hastings–Green algorithm to approximate this integral over all possible genealogies. In the following explanations, I will focus on the program MIGRATE (Beerli and Felsenstein 1999, 2001; Beerli 2006) but all discussions of Markov chain Monte Carlo approximations and most, if not all, problems are shared with the other programs that use such an approximation.

WHAT IS `MARKOV CHAIN MONTE CARLO´?

The Markov chain Monte Carlo (MCMC) method is an integration technique for problems that have no simple analytical solution. Instead of exploring the function to integrate in a systematic manner, as in standard numerical integration techniques, MCMC is an autocorrelated method, where each step or sample depends on the last one, but it also has no memory because no step prior to the last one is remembered and thus, cannot influence the choice of the next step. Requirements for the method to work are

- It must be possible to calculate the integration-function up to a constant. We can often reduce the function of interest to two functions: one that we can calculate and another one that we cannot solve analytically but can hold constant throughout the analysis. Replacing this constant with 1 typically does not change the relationship among the steps or the steepness of the function but only the height of the function.

- Each point on the probability-landscape must be reachable from any other point, if necessary in multiple steps.
- Moves from an old point to a new point on this probability-landscape are reversible and equally likely; if not, this directional bias needs to be corrected.

An almost too simple example

Integration takes a central role for calculating the expectation of a probability distribution. It is standard procedure to calculate the integral analytically or to solve it piecewise, most often by discretizing the continuous distributions. The only requirement for such an approach is that we must be able to calculate the function at any point. With many discrete pieces this function can be integrated with high accuracy. Unfortunately, with many parameters (many dimensions) this approach does not work very well. Often, the function cannot be calculated on an absolute scale but only relative to an arbitrary quantity; therefore, all evaluations using this unscaled function will be off by a constant. When we compare function-values within the same analysis, the differences of these unscaled function-evaluations are the same as those using the correctly scaled function, which we typically cannot calculate easily. This new unscaled function can, however, be used in an MCMC context. The algorithm works like this

Step 1.1: Start with a random assignment of parameters (for example migration rates, population sizes, and genealogy)
Step 1.2: Evaluate the function for this first step (L_{old})
Step 2.1: Change the parameters (or a single parameter at a time)
Step 2.2: Evaluate the function for this step (L_{new})
Step 3.1: Evaluate the ratio $R = L_{new} / L_{old}$
Step 3.2: Draw a random number r from a uniform distribution between 0 and 1.
Step 3.3: If $r < R$ then accept the parameter change and record the new state; otherwise stay at the old state, and record it.
Step 4: Go to 2.1 and repeat many, many times.

For a simple illustration of the steps above, I used a convolution of two normal distributions: in this case the absolute probability density function is known and can be calculated (smooth curve in Fig. 3.1). The histograms were built up using a very simple MCMC procedure that was optimized for this problem. Figure 3.1 shows an MCMC run for a single parameter after 3 steps, 300 steps, 300 000 steps, and 3 000 000 steps. Improvement of the approximation to the area under the curve of the function is obvious.

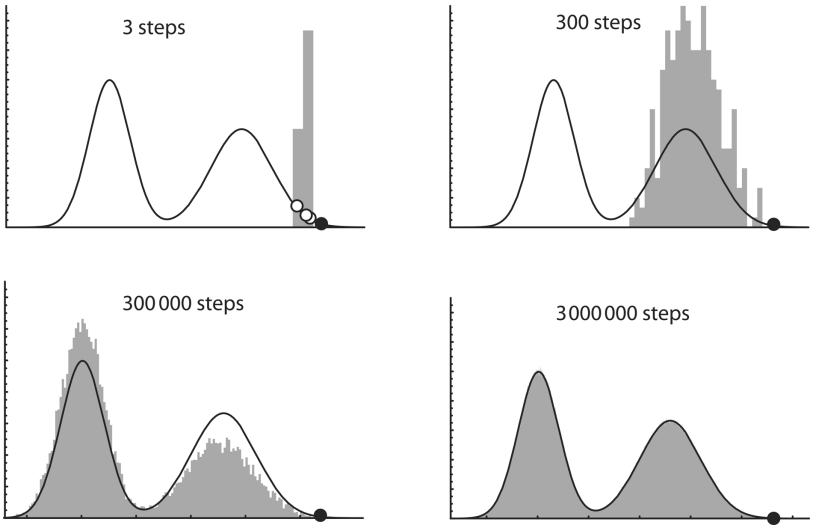


Figure 3.1. Approximation of the area under a curve using MCMC: The curve is the exact function, the grey area is the approximation using MCMC. The black dot marks the starting point of the run, the white dots in the top left panel show the three sampled states that make up the histogram.

Figure 3.1 clearly shows that without running many steps, the approximation is very crude. However, the problem is that there are no clear stopping rules; for example if we are only interested in the maxima of the function, a sample of 300 000 steps would be fine, but the area under the curve is still not approximated very well. If we do not know the function well enough, we would still not know whether there are more than two peaks. This example is very simple and it is important to remember that any integration in the context of multiple parameter estimation will almost certainly be more difficult and less accurate.

Migrate – a program for inferring population genetic parameters

I will use my program Migrate to explain some general difficulties of using software that employs MCMC, and will also give some ideas on how to analyse data using such software.

Migrate uses two frameworks: (1) coalescence theory to model population genetics forces, such as population sizes and migration rates, and (2) mutation models that explain the change of alleles or nucleotides at sites over time. Both models are simplistic, but for many reasons, no better

alternatives are available. It certainly is a strong assumption that Kingman's population genetic model fits all natural populations, but comparisons with other statistics, for example F_{ST} (Beerli 1998), have shown that coalescence theory recovers population scenarios at least as well as or better than some of the other methods. The mutation models are borrowed from phylogenetics (cf. Swofford *et al.* 1996; Felsenstein 2004) or 'old-fashioned' population genetics (Kimura and Crow 1964; Kimura and Ohta 1978a; Ohta and Kimura 1973). In phylogenetics, the distinction of the terms substitution and mutation is important, but, within this population genetics framework, we assume that mutations are neutral or nearly neutral, and therefore, substitution and mutation are equivalent.

Coalescence theory

Kingman (1982a, b) extended Sewall Wright's observation (1951) that it takes two randomly chosen chromosomes in a population of size N about $2N$ generations until they meet in their most recent common ancestor. Kingman showed that it is possible to calculate the probability of a genealogy of any number of individuals. His findings allowed the use of a random sample of individuals to infer parameters for the whole population. Hudson (1991) popularized Kingman's n-coalescent among biologists and today, many extensions of the basic n-coalescent exist; for example, models on recombination (Hudson and Kaplan 1988), gene flow (Hudson *et al.* 1992; Notohara 1990; Wilkinson-Herbots 1998), speciation (Nielsen 1998), selection (Kaplan *et al.* 1988; Neuhauser and Krone 1997; Felsenstein 2004) and many more. The coalescent was derived using a rather general population model, the Cannings model, which is a generalization of the Wright–Fisher population model. The Cannings model allows for variance in the offspring function, whereas the Wright–Fisher model fixes this variance at 1 (Ewens 2004). The coalescent fits simulated data that were generated using a time-forward process almost perfectly when the population model is the Wright–Fisher model. Although the coalescent is robust, caution is needed because it is a diffusion approximation and holds in principle only when the population size is much larger than the sample size, because with either large sample size or very small population size, we expect an increased probability of multiple coalescence per generation, which Kingman's n-coalescent ignores. The effects of multiple coalescences in a generation and effects of sample numbers were explored by several authors. Additions to the coalescence theory by Pitman (1999), Möhle (2000), Schweinsberg (2000), Möhle and Sagitov (2003) and Fu (2006) allow for situations in which more than two lineages merge in the same generation and therefore, for a less restrictive

ratio of sample size and population size. Fu (2006) compared the standard coalescent with his multiple-merger coalescent and found that the standard coalescent works astonishingly well even with small populations and large sample sizes; this corroborates the finding of Wakeley and Takahashi (2003) that the standard coalescence is robust as long as the sample size is smaller than the effective population size. If the reproductive success is very uneven among individuals, the concept of effective population size could, in principle, become meaningless, for example, if one individual produces all the offspring for the next generation (Eldon and Wakeley 2006). Such a 'neutral sweep' would be indistinguishable from a selective sweep. The risk for such a sweep decreases as the size of the population increases. It is perhaps most pronounced in species that can have small population sizes and produce millions of gametes per individual, as is the case for many fish species.

Mutation models

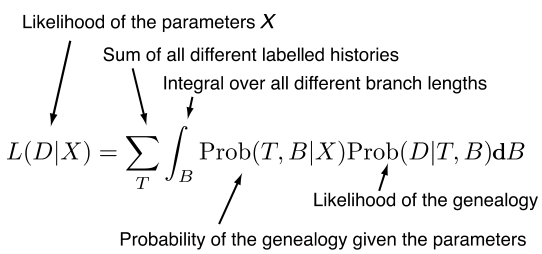
Readers familiar with phylogenetics know that many studies are preoccupied with using the best substitution model. In population genetics, the problem of misspecification of the mutation model is less severe because the gene trees (genealogies) typically occupy a much shorter time period than phylogenetic trees. MIGRATE accommodates only a few nucleotide mutation models; the default is the Felsenstein 84 model (F84: Hasegawa *et al.* 1985). This model is similar to the Hasegawa–Kishino–Yano (HKY) model: both allow for different nucleotide frequencies and uneven transition rates between purines and pyrimidines (see Swofford *et al.* 1996). Restricting the F84 model, for example by setting all base frequencies equal to 0.25, makes it equivalent to simpler models. This model is not very sophisticated, but it incorporates important features of sequence evolution without many additional parameters. Population genetic inference uses a much more recent time window than phylogenetics and more sophisticated models are warranted only for very rapidly evolving microbes. Researchers in population genetics often accept much simpler models for sequence data, such as the infinite sites model or no-mutation models. MIGRATE does not estimate mutation model parameters, such as transition-transversion ratio and site rate-variation parameters. To get good results, it is better to input specifics about the mutation model and whether rate variation among sites should be assumed. Such parameters can be derived using other programs such as PAUP* (Swofford 2003) or MODELTEST (Posada and Crandall 1998). Recently, single nucleotide polymorphism data were used to investigate population genetics features in humans (Wakeley *et al.* 2001). Programs like MIGRATE and LAMARC (Kuhner 2006)

can adjust for the fact that only variable sites are used in the analysis. This is important because, without correction, population genetics parameters would be overestimated (Kuhner *et al.* 2000; Nielsen 2000; Nielsen and Signorovitch 2003; Clark *et al.* 2005).

The models for electrophoretic markers and microsatellite markers are even less sophisticated than the sequence models, although a large number of possible models is known (Calabrese and Sainudiin 2005). Most of these more sophisticated models are difficult to apply many millions of times during a single run: each might need a separate MCMC run to estimate a single branch length. MIGRATE allows the use of mutation models for allozyme data (Kimura and Crow 1964) and for microsatellites (single-step mutation model: Ohta and Kimura 1973; Kimura and Ohta 1978b) and a Brownian motion model that approximates the single-step mutation model (Beerli 1997; Blum *et al.* 2004). DNA or RNA sequence data often contain more information about the history of mutations in the sample and therefore, usually allow for better inferences than other types of data. Nevertheless, these other data types (allozymes, microsatellites) still contain useful information about the population genetics processes. The genealogies generated with such data may look uninformative but, as the example in this section shows, allow us to make inferences that go beyond F_{ST}-based analyses.

How are these pieces combined?

MIGRATE infers parameters either by (1) maximum likelihood or (2) Bayesian inference. A central probability in MIGRATE is the probability of the parameters for a specific data set and a specific genealogy. This probability is calculated as the product of the probability of the data given the parameter and the probability of a genealogy for a given parameter value. Finally, the likelihood is the sum over all genealogies (topologies and branch lengths) of this weight:



Bayesian inference uses an arbitrary prior distribution for each parameter and the coalescent as a prior distribution for the genealogy, but it also

needs the likelihood machinery to sum over all genealogies. Details were given by Beerli and Felsenstein (1999, 2001) and Beerli (2006). This sum over all genealogies is approximated using MCMC and the likelihood is scaled by an unknown constant: it is a relative likelihood. It is important to recognize that a specific log-likelihood value is uninformative, and that the likelihoods of different independent runs with MIGRATE typically cannot be compared. This topic is discussed in the section 'Likelihood ratio tests and related test statistics'.

Running in maximum likelihood mode

Maximum likelihood analysis (ML) and Bayesian inference (BI) use different schemes to estimate parameters. The likelihood method starts with arbitrary values for parameters and genealogy. A new set of genealogies is found with these arbitrary parameter settings using MCMC (these parameter are called the driving parameters because they drive the MCMC). Maximum likelihood estimates of the parameters are then found using this new set of genealogies. These maximum likelihood estimates are probably quite different from the driving parameter values because the data are pushing the likelihood function (and thus the parameter values) towards values that are compatible. A second MCMC chain uses these new parameter values as driving parameters and samples a new set of genealogies after which a new set of parameter values is estimated. This iterative procedure inches towards parameter values that are compatible with the data. By trial and error we (Mary Kuhner, Jon Yamato, Joseph Felsenstein and Peter Beerli, unpubl.) found that several chains that are relatively short allow the exploration of the parameter space. It typically takes about five to ten chains to find sufficiently good driving values, as marked by small changes of parameters between consecutive chains; then two or three very long chains are run and the last chain is used to report the maximum likelihood estimates. Approximate confidence intervals are calculated using profile likelihoods.

Running in Bayes inference mode

For Bayesian inference, it seems most profitable to run one single long chain with a prior distribution for each parameter or combinations of parameters. Parameters and genealogy are updated randomly using a user-specified frequency of genealogy-changes. For likelihood, the driving values needs adjusting, whereas in a Bayesian framework the prior distribution of the parameters provides a mechanism for exploring different parameter values to change the genealogy during the MCMC run. The

parameter values recorded during the run of this single long chain are then used to generate a posterior probability density for each parameter. Migrate displays these posterior distributions as histograms and also tabulates quantiles, mode, median, and mean. The most important features are the mode of the posterior distribution (i.e. the maximum posterior estimate), and the 2.5% and the 97.5% quantile, the borders of the 95% credibility interval.

In ML, the success of run depends on the length and number of short and long chains, whereas in BI the choice of the prior distribution is critical. This prior distribution is often a simple distribution that reflects our knowledge of the parameters before the analysis. Researchers often apply uninformative prior distributions, such as the uniform distribution, perhaps hoping not to bias the posterior distribution. However several Bayesian statisticians suggest using prior information and advocate the use of informative prior distributions. Informative data will overpower any reasonable prior distribution, but informative priors will influence the result when the data is weak. Effects of choices of prior boundaries are discussed using an example in a later section. In Migrate, several prior distribution are implemented: a uniform distribution with lower and upper bounds that need to be chosen more extreme than any parameter compatible with the data, and two types of exponential distributions that put more emphasis on small values dependent on the mean of the distribution.

A SHORT EXPLANATION OF WHAT Migrate DOES AND DOES NOT DO

Migrate, like other population genetic model-based methods, is based on several assumptions. It shares almost all of these assumptions with other programs that infer population sizes or magnitude of gene flow. These assumptions are:

- *Population sizes are constant through time or are randomly fluctuating around an average population size.* This assumption is very common for many population genetics analyses, especially F_{ST}-based analyses. Only a few programs that estimate gene flow relax this assumption, for example Lamarc (Kuhner 2006), and IM (Hey 2005). The program Beast (Drummond *et al.* 2005) estimates varying population sizes through time for a single locus and a singe population. Additionally, some tests are now available for detecting whether a drastic decrease in population size occurred in the past (for example Cornuet and Luikart 1996); however, many loci are needed and the effects of the

population bottleneck must be severe for it to be recognized. Such tests often ignore gene flow among populations or other population genetic forces.

- *Individuals within a population are randomly mating, and each individual has the same potential to have offspring.* Therefore, it is assumed that no selection is acting on the loci under study. The creation of programs for the inference of selection coefficients with a coalescence-based framework is underway.
- *Mutation rate is constant through time and is the same in all parts of the genealogy.* Although MIGRATE assumes rate constancy on the genealogy, it allows using of site rate variation among nucleotide sites and mutation rate differences among loci. Only phylogenetic methods, for example *r8s* (Sanderson 2002), and the program BEAST (Drummond *et al.* 2005) allow for different rates on different branches, but these programs either do not account for population parameters at all or only population sizes.
- *Immigration rate is constant through time, but can differ among populations.* All programs that allow for the estimation of migration rates force rate constancy through time or some segments of time (for example IM: Hey and Nielsen 2004); in addition, F_{ST} -based analyses also impose symmetric rates or symmetric numbers of migrants.
- *Populations exchange genetic material only through migrants, so no population divergence is allowed.* If the time of the most recent common ancestor is younger than the divergence time then MIGRATE is a perfect tool. If you have a data set with two populations that have split only very recently you might want to compare your MIGRATE results with the results from IM (Hey and Nielsen 2004). In contrast to IM, MIGRATE can analyse one, two, or more than two populations; using only population pairs can lead to overestimations of parameters (Beerli 2004; Slatkin 2005).

What happens when the population history violates the assumptions?

One of the most frequent comments from of users of MIGRATE is that it is not applicable because the population history of their species violates the assumptions of MIGRATE. However, it is important to remember that no program will be able to relax all assumptions, and practitioners need to assess whether an assumption violation will harm their conclusions. Figure 3.2 highlights the direction in which the program will err when assumptions are violated. Several population scenarios that deviate from the assumption that the population size is constant through time were simulated (see

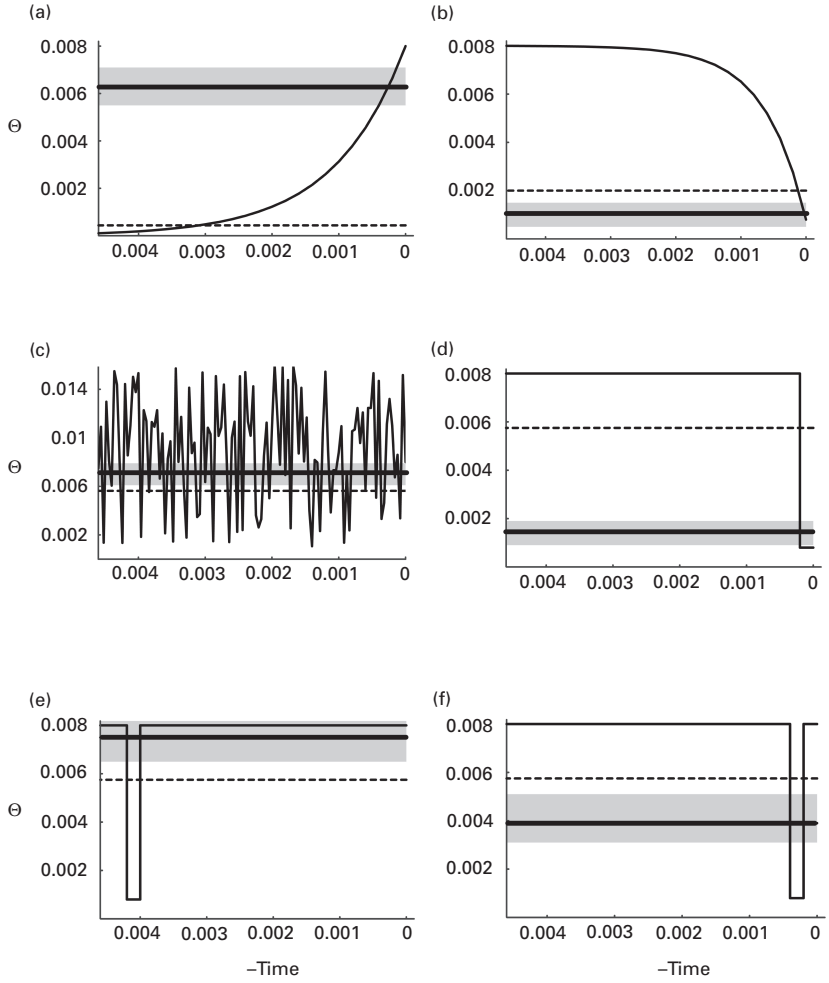


Figure 3.2. Estimation of population size under different population histories. The *x*-axis shows time scaled by mutation rate: past to the left, today is at 0. The *y*-axis shows the mutation scaled population size Θ that is 4 * effective population size * mutation rate per site. Thin lines show the true population size through time; the dashed line was calculated from the true population sizes using a harmonic mean to estimate the average long-term population size; the grey area is the 95% credibility interval and the thick line is the value at the mode of the posterior distribution evaluated by MIGRATE using simulated data sampled at time 0 (1 population with 50 individuals sampled; 10 loci each 10 000 base pairs long; details in Appendix).

Appendix for the simulation and run details). With growing or shrinking populations, MIGRATE will under-or overestimate the effective population size, respectively (Fig. 3.2a, b). The results show that the estimates are mainly influenced by the situation close to the sampling date. On a genealogy with concurrent tips, most lineages are present close to the tip date and will contribute more to the final estimate. With randomly fluctuating population sizes (Fig. 3.2c), the estimate will roughly track the average size. Interestingly, before this experiment, I had expected this estimate to be the harmonic mean, which is believed to track the long-term population size; however, the most recent fluctuations contribute more to the estimate and so many replicates might show an average at the harmonic mean. Short bottlenecks in the past have little effect on the estimate (Fig. 3.2e), whereas recent bottlenecks might mimic a smaller population size (Fig. 3.2f). If the population decline to moderate numbers is very sudden and very recent, MIGRATE is strongly influenced by the bottleneck (Fig. 3.2d). These outcomes need to be explored in more depth, and more simulations with different number of sampled individuals need to be done (Beerli, unpubl.). In any case, it is already possible to say that MIGRATE is influenced by recent changes in population size despite the fact that it delivers long-term estimates.

Example data set

As an example a data set, I will us the one for water frogs from my Ph.D. thesis (Beerli 1994). The data are listed in the Appendix and include five populations and 31 electrophoretic marker loci; Beerli *et al.* (1996) and Beerli (1994) provide details about the different loci. Today, electrophoretic marker data may seem outdated, but it has only recently become easy to sample more than 30 anonymous sequence loci (Brumfield *et al.* 2003), or microsatellites for most species groups. A complete analysis is difficult because of uneven sampling, uneven distribution of alleles, and (perhaps even worse) lots of missing data. The localities are mapped in Fig. 3.3. This data set is interesting because additional information about the geological history of this area is available. After the last glaciation period (Würm period) ended, the water level rose about 120 m and so isolated the island Samos from the mainland around 10 000 years ago (R. A. Rohde at http://globalwarmingart.com/wiki/Image:Post-Glacial_Sea_Level_png based on Fleming *et al.* 1998; Fleming 2000; Milne *et al.* 2005). The salt water barrier between Samos and Anatolia is shallow. However, the sea between Samos and Ikaria is rather deep and the two islands were probably only connected during the most severe of the more recent glaciation periods (Mindel period) about 200 000 years ago.

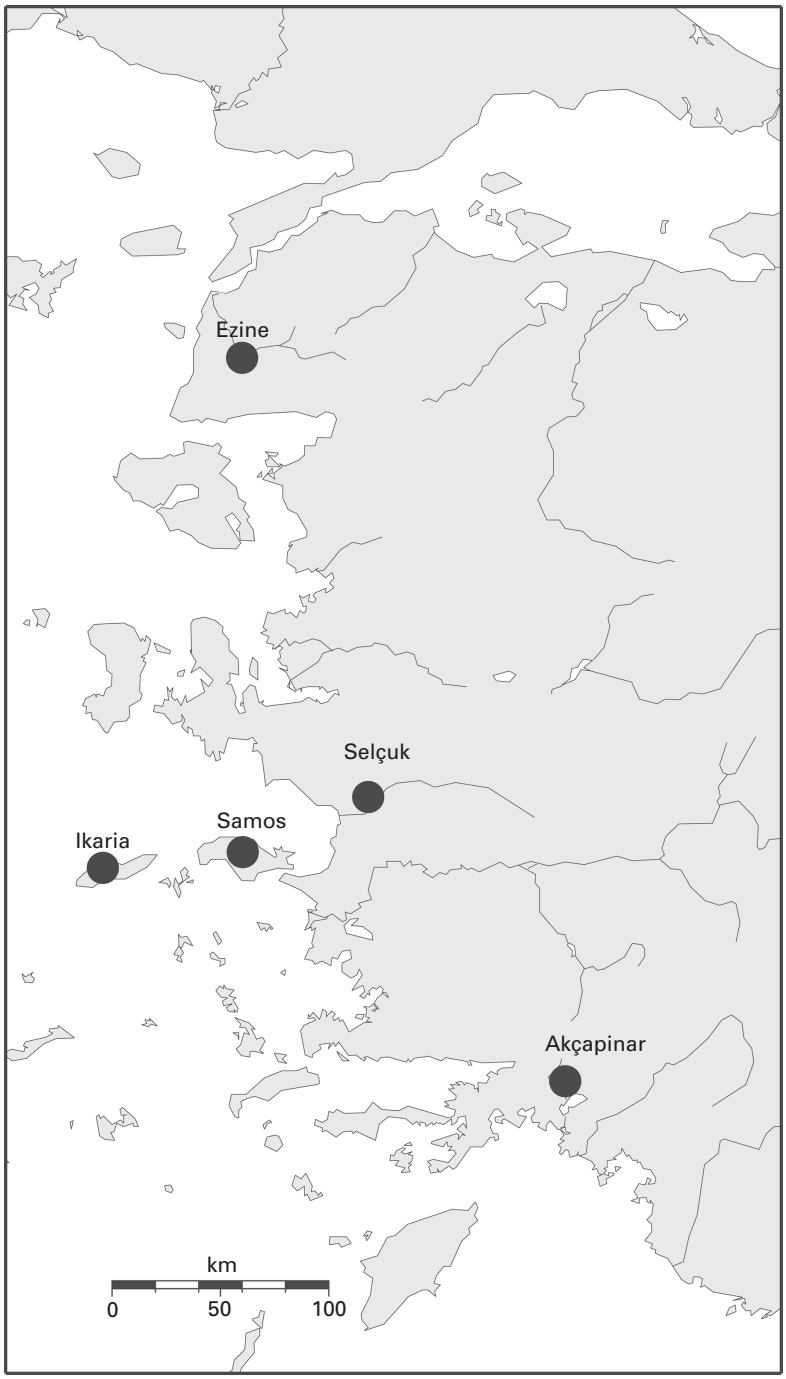


Figure 3.3. Map of water frog sampling locations on Anatolia, Samos and Ikaria.

ANALYSIS USING MIGRATE

I will now analyse the frog data set to estimate the gene flow pattern to and from the mainland (Selçuk) and islands Samos and Ikaria. We will assume that more gene flow occurs from the mainland to the islands than from the islands to the mainland, and in the following sections we will explore this hypothesis. The analysis in this chapter is incomplete, but reveals both difficulties and successes.

Basic analysis – getting familiar with MCMC-based software and data

MIGRATE version 2.0 and newer (Beerli 2006) has the capability of inferring the parameters using either maximum likelihood (ML) or Bayesian inference (BI). For a first analysis, BI is preferred over ML because simulations have shown that, with non-informative data, results using MCMC-based ML analyses are more error-prone (Beerli 2006). This chapter will give a sketch of a possible way to analyse any data and gain confidence that the results are correct. In a first encounter with the program and the data set, I suggest experimenting with the program using the default values for the run conditions. Once you are convinced that the data has been read correctly and the program runs to completion, run the program with the default values. Be aware that default values are chosen so that the program can finish in a reasonable time frame for small to moderate data sets. Depending on the number of parameters to explore, such defaults can be inappropriate and should only be considered as the roughest guide. The number of populations in the example data set is five, so there are 5 population-size and 20 migration parameters. The default values, and so the first default ML or BI run, will not be very trustworthy because these defaults were set for much smaller data sets. With 25 parameters, the MCMC runs will be 'too short'. The MCMC procedure adds variance to the variance introduced by the data, and only multiple runs of different lengths will help to evaluate the magnitude of this variance.

One of the common mistakes of such analyses is that researchers want to do it right on the first try; they will run all the data on very long chains and are disappointed when the program fails or the reported end of that single run is in the following month. A better practice is to use several trial runs to see how the software behaves (this is true for any program that uses MCMC). For BI, change the settings in the Strategy menu of MIGRATE and make sure to visit all submenus, especially the menu entries on the prior distributions. For a first run, choose one 'long' chain to explore around a million steps and save around 100 000 steps. On small data sets with few

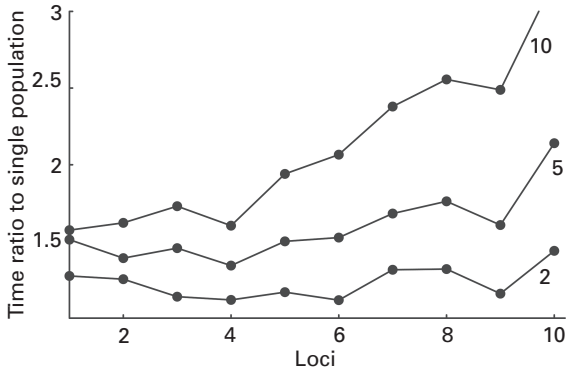


Figure 3.4. Comparison of the runtime of simulated data sets with 1, 2, 5 and 10 populations and 1 to 10 loci. The *y*-axis shows the runtime ratio of the multi-population parameter estimation compared with the single population. The effort for each run was the same: each run used a total of 100 sampled individuals with a total of 10 000 base pairs each. For example the last data point for the 5-population setting uses 20 individuals per population and 10 loci, each 1000 base pairs long.

loci and few populations this will take minutes, but might take a couple of hours on data sets with more than four populations and a single locus. Figure 3.4 gives a rough comparison of runtime of different population scenarios and number of loci compared to a single-population run. With 10 populations and 10 loci, the runtime is about three times longer than with a single population when the amount of data is the same for all scenarios. In reality, researchers will have 10 times more data from 10 populations than from one population, therefore, runtime will be probably about 30 times longer.

We can think of this first run with the default values as a baseline run. We expect that the resulting posterior distribution will not be smooth, and it is quite possible that some parameters will show strange posterior distributions (Fig. 3.5a). For example, if your data suggest a population size of 0.1, but your prior distribution is uniform on the interval 0 to 100, then most proposals will be rejected because most of the suggested population sizes are incompatible with the data. In such cases, we need to shrink the upper bounds of the uniform prior, increase the number of samples considerably, or use another prior, for example, an exponential prior. Figure 3.5 gives examples of what could go wrong with prior specification. Once we get an idea how long to run the MCMC chains, set up an even longer chain and use this to report results. For ML analyses, a similar iterative approach is useful. The default settings will often work for two-population data sets that are moderately or highly variable. The example data set needs longer

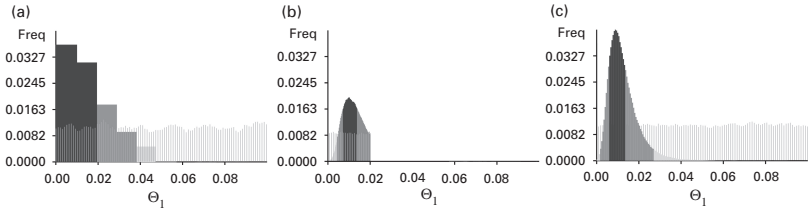


Figure 3.5. Effect of mis-specification of prior distribution on the posterior distribution. A sample of the prior distribution is shown with thin lines; histograms are posterior distributions: shading marks approximate 50% (black), 95% (dark grey) credibility sets. (a) A uniform prior in the range between 0.0 and 10.0, which is too diffuse combined with too few samples from the MCMC, does not lead to an informative posterior distribution. (b) A prior distribution that has too slow an upper limit (0.02) cuts off the posterior distribution at that upper limit. (c) Uniform prior distribution that facilitates fast convergence without truncation for this data set (upper limit 0.1, many more steps saved). Detailed run condition in Appendix.

runs than the defaults and the sampled chains for the short and long chains should be large. ML uses an iterative scheme of several short and long chains because it does not change the parameter values that drive the MCMC. If these driving parameters are too small, convergence to good estimates is very slow. An iterative improvement of the driving values with several shorter chains moves these driving values towards the 'true' values (Wilson *et al.* 2000). When the driving values are sufficiently close to the 'true' values the ML approach delivers good estimates. ML estimates are very useful for establishing a likelihood ratio test framework (as discussed in the section 'Likelihood ratio test and related test statistics').

Comparison of effect of gene flow using the Bayesian framework

In contrast to a DNA sequence locus, an individual allozyme locus is not very informative because the history of the sampled mutations cannot be inferred; but with many loci there is a good chance that we can recover directionality in gene flow. Figure 3.6 shows such an analysis. MCMC run-conditions are specified in the Appendix. The migration rates were calculated assuming that migration (gene flow) is only possible between nearest neighbors and geographic distance is also taken into account. A user can supply a geographic distance matrix between the localities and these distances will scale the migration rate. If migration rates are only a function of distance then all values should be similar. For frogs, salt water is a barrier; therefore, we expect lower migration rates than over land. Hence, I expected lower migration rates between Samos and Selçuk, and Samos and Ikaria, compared to migration rates between mainland locations. In fact, the

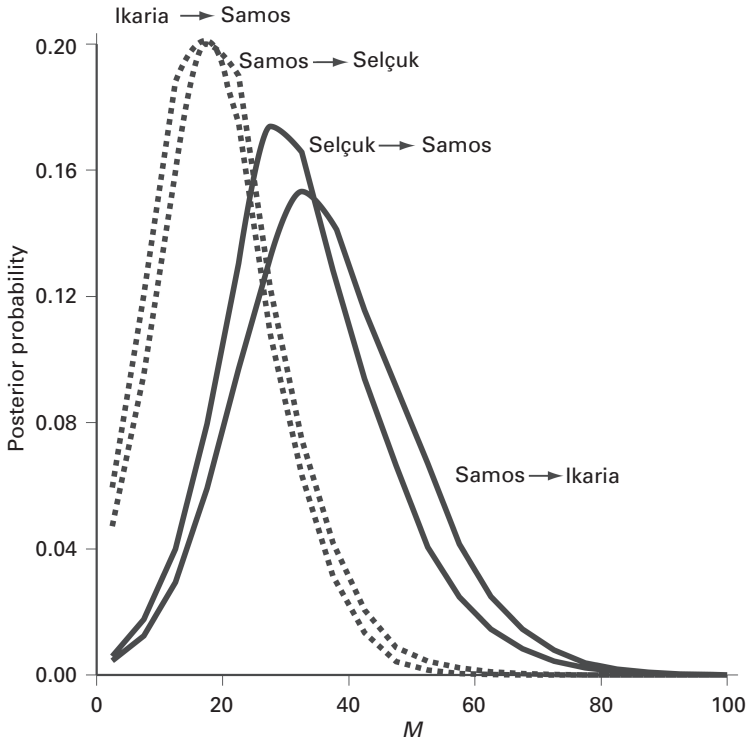


Figure 3.6. Posterior probability distributions of the mutation-scaled migration rate $M_{ji} = m_{ji}/\mu$ where m is the immigration rate per generation into a population i from j and μ is the mutation rate. All six pairwise migrations between the mainland (Selçuk) and the island of Samos (close to the mainland) and between Samos and Ikaria are shown.

migration rate between Samos and Ikaria should be the smallest because the sea strait separating Ikaria persisted for the longest time. The migration rates from the mainland (Selçuk) to the islands is much larger than from the islands to the mainland; for example the rate from Samos to Selçuk is about half of the rate from Selçuk to Samos (Fig. 3.6). The difference in geographic distance between Samos and Ikaria is larger than between Samos and the mainland, so we would expect a difference in gene flow; in this case, however, the difference seems smaller than expected.

Comparison of Bayesian inference and maximum likelihood

It is difficult to make a fair comparison between BI and ML, because each program use slightly different models and programs. Recently, the

programs MIGRATE (Beerli 2006) and LAMARC (Kuhner 2006) were improved and can run both BI and ML. Only the portions of the program that constitute the individual statistics are different. ML works well with very variable data (Beerli 2006; Kuhner and Smith 2006), but has problems with low-variability data (Beerli 2006; Kuhner and Smith did not evaluate low-variability cases). When the data do not contain many variable sites the ML approach has difficulties in converging and needs very long MCMC chains. Often with such data, the ML approach does not give good guidance whether the data can support or reject a population model. In contrast, BI calculates posterior distributions that are similar to the prior distribution, thus alerting the user that the data may not support a complicated population model. In a Bayesian context, it is possible to use the distribution similar to that of the prior distribution to assess whether the data are overfitted with too complicated a model. When the posterior is identical to the prior then the data do not contribute to the result. In fact, programmers use this no-data case as one test to check whether the programs run correctly. In the ML analysis this is somewhat trickier: in current implementations, the MCMC algorithms describe a Brownian motion walk because the data have no influence. Running from the same starting point many times will produce results that are 'normally' distributed around the starting value.

Runs using BI and ML of the water frog data set reveal some differences, but the overall picture is about the same. A comparison of Figs. 3.6 and 3.7 shows that the two approaches agree that the gene flow to islands is higher than from the islands to the mainland.

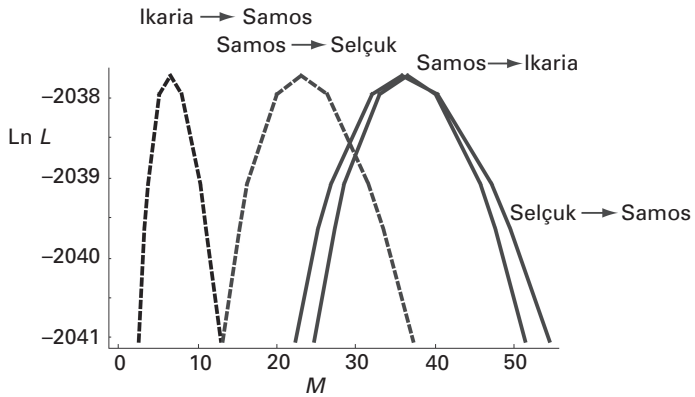


Figure 3.7. Log profile likelihood (Ln L) of mutation-scaled migration rates $M_{ji} = m_{ji}/\mu$ where m is the immigration rate per generation into a population i from j. The two curves closer to zero are for gene flow towards the mainland.

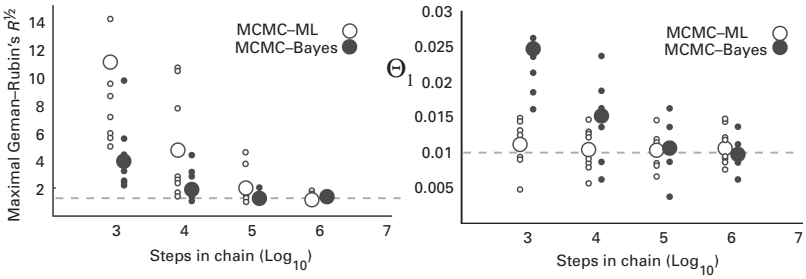


Figure 3.8. (a) Gelman–Rubin statistic of the Bayesian and ML schemes when run for different numbers of sampled steps in the last chain. Values below the dashed line show convergence. (b) Values of Θ estimates using the same runs as in (a). The dashed line in (b) is the population size used to simulate the datasets. Large dots are averages of 10 independent runs (small dots). The data were modelled using two populations; only the size of population 1 was reported.

How long to run

MCMC runs of complicated models need much longer to converge than simple models. The convergence rate is dependent on the data: when the true branching pattern and the mutation events are well distributed, convergence is fast; with low variability or very long terminal branches, the convergence is typically slow. The example data set needs longer chains than the default in MIGRATE. Although the program calculates the Gelman–Rubin convergence diagnostic (Brooks 1998), the best test is longer and longer trial runs. For example, increase the run-length by a factor of 10, until different runs return similar, consistent, results. This exercise is also useful because you become more familiar with the output file format and the program in general. Convergence diagnostics can show successful convergence, but the results may still be very different among runs when too few samples are taken. In a two-population scenario with simulated data from 10 loci (Fig. 3.8), BI seems to converge faster than ML when judged by the convergence diagnostic, but the estimates of ML converge faster to the true value than BI. This is only a single, very simple example, but still it needs to run for at least 105 steps. For most data sets, simple MCMC runs do not achieve good results because the chain does not explore the possible solutions very easily and improvements of the MCMC strategy are needed.

Replication and heating

Geyer (1991; Geyer and Thompson 1992) developed a replication scheme that allows combining different MCMC chains for ML estimation. This

scheme calculates relative weights for each chain and so adjusts the contribution of each chain to the final ML. This replication scheme is used in MIGRATE and LAMARC (see Wilson *et al.* 2000).

Geyer and Thompson (1995) and others developed a method that uses several chains run with different acceptance ratios powered by the inverse of a 'temperature' (Metropolis-coupled MCMC or MCMCMC). With a temperature of 1.0, standard acceptance ratios are used; with a temperature of ∞, all changes in the MCMC are accepted. This powering up of the acceptance ratio essentially flattens the solution space and so makes it easier to cross deep valleys and descend from very steep peaks. After each chain has made a step, a random pair of temperatures is compared using a Metropolis algorithm-based acceptance ratio and, if the move is accepted the chains running at different temperatures swap parameter states.

With more than two populations, I suggest exploring heating very early in the experimental runs because you do not know what the solution space looks like. It might be jagged and then you need chains that can jump between peaks. MCMCMC is a possible solution to such problems. MIGRATE allows to set arbitrary temperatures, and a static or an adaptive heating scheme. The adaptive heating scheme takes the start temperatures and decreases the temperature difference by 10% between chains that do not swap for a preset number of trials. If the chains swap more than once in the preset number of trials, the temperature difference increase by 10%. Adaptive heating with a fixed number of heated chains is not the cure-it-all for difficult mixing problems; a system that allows insertion or deletion of chains would be superior over simply increasing or shrinking the temperature difference of existing chains.

How long to wait

Runtime on a single CPU machine depends on the number of loci and the number of replicates. As a simple rule of thumb you can expect that time to increase linearly with the number of loci; for example, if one locus takes a couple of hours then with 31 loci, expect a run of several days on a single CPU machine. The run-length is highly dependent on the number of populations: the time to evaluate genealogies depends on the number of possible events on the genealogies. With n populations there are n different coalescent events, and with the default connection matrix among populations there are $n(n-1)$ possible migration events. Increasing the population number by 1 increases the possible number of events by a factor of $2n-1$ (Fig. 3.4). This increase is typically accompanied by an increase of the total number of individuals, which results in an additional slow-down.

For data sets with many populations, many loci are needed to get accurate estimates. Figures provided by Beerli and Felsenstein (1999) and Beerli (2006) show the reduction of the variance when using more than one locus. Estimates based on many loci take a long time and for such data sets, it is often more convenient to run them on a computer cluster. MIGRATE can run on a large number of computer systems. Difficulties arise when users have a large data set with many loci and want to run it on their laptop or desktop computer. Runs as outlined in this chapter will often take much too long and either the machines are needed for some other tasks or the power goes out.

The program can use symmetric multiprocessing (multiple threads) for running parallel chains with different temperatures. The use of a threaded program is not different from a non-threaded program. This is an efficient use of many high-end desktop machines with two CPUs or, very recently, with dual-core CPUs that can be found even in laptops. Typical gain in speed over non-thread runs is about 1.6 for Bayesian runs, and a little less than that for ML runs because the calculations for the approximate confidence intervals are not threaded.

The fastest way to run MIGRATE is to compile it for use on a computer cluster. The program can take advantage of large clusters running multiple loci and replicates on different CPUs. It uses the message passing interface (MPI: Gropp *et al.* 1999a, b). Several free programs, such as OPENMPI (Gabriel *et al.* 2004), LAM-MPI (Burns *et al.* 1994; Squyres and Lumsdaine 2003) and MPICH2 (http://www-unix.mcs.anl.gov/mpi/mpich/index.htm) are available to set up a virtual cluster on top of the real computer cluster. This real computer cluster can be a single machine or a network of idle lab computers, or a dedicated set of machines connected with a very fast network. Once the virtual cluster is functional, it is only a matter of compiling MIGRATE for such a cluster and running it. The MIGRATE manual gives details of installing and running MIGRATE on such machines. The speed gain depends on the number of loci, number of replicates, and how many real CPUs are available. I typically run MIGRATE on a small cluster of 15 computers with 30 single core 2 GHz AMD Opteron CPUs. The runtime difference is remarkable: the default run of the example data set took about 1 hour and 17 minutes whereas an Intel Core Duo (dual core) 2.16 Ghz machine took about 15 hours. For a researcher with some computer administration knowledge it is rather simple to establish an ad hoc cluster using desktop computers if they run some form of the UNIX operating system (for example LINUX or MacOS X); Windows might be trickier.

Can we trust the support intervals in a MCMC-assisted maximum likelihood analysis?

The support or approximate confidence interval of the maximum likelihood estimate is evaluated using profile likelihoods. In contrast to maximum likelihood, which finds the set of parameters with the highest likelihood, profile likelihood fixes one parameter at an arbitrary value and then finds the set of other parameters that maximize the likelihood. Often, we assume that the likelihood function approximates a χ^2 distribution. Significance levels of this χ^2 distribution then allow specifying quantiles and, thus, support intervals. With short MCMC runs the landscape of genealogies is not well explored and, therefore, the uncertainty of the parameters might be underestimated. This is somewhat disturbing because it means we will be overconfident in our results. With informative data, very long runs often allow a good approximation of the support intervals. Recently, Abdo *et al.* (2004) claimed that the profile likelihood tables of MIGRATE are inadequate. Their simulation study used the program defaults and ignored guidelines in the manual about how long to run MIGRATE. They showed that the 95% support interval in MIGRATE is often too narrow. In simple scenarios, such as the one they tested, it should be possible to achieve appropriate confidence limits with informative data. Beerli (2006) showed in a much more complicated four-population scenario that, with certain parameter configurations, the data do not contain enough information to estimate migration rates with confidence. Such data sets typically do not produce consistent results when run several times using ML in MIGRATE, and therefore fail to deliver consistent support intervals. Using BI, we can recognize that the posterior distribution is similar to the prior distribution. The example data set does not contain much information per locus, but the 31 loci produce consistent results using BI. ML produces somewhat more variable results but the directionality and magnitude are the same (compare the modes of Figs. 3.6 and 3.7).

LIKELIHOOD RATIO TESTS AND RELATED TEST STATISTICS

Often, we might want to test one migration scenario against another. The MCMC approximations makes this rather cumbersome because only relative likelihoods are calculated, and in normal (default) runs there is no control about the driving values that define the denominator of the relative likelihood. MIGRATE allows estimating an approximate likelihood ratio test (LRT) by using the sampled trees to test nested migration models. For example, using the ML scheme, many genealogies are sampled using

the default connection matrix among populations: all can connect directly. By supplying an alternative to the most general model, we can test whether the power of the more restricted model to explain the migration scenario is similar to that of the full model. Accepting a parsimony criterion, we would choose the model with fewer parameters.

Comparison of two different migration models

Can we exclude migration from the islands to the mainland (Fig. 3.9)? Running Migrate using the likelihood ratio test allows us to make a comparison, but this comparison is only approximate because the full (or the more complete) model is used to sample genealogies. These are then used to evaluate the likelihood both of the model that was used to sample the genealogies and of the model with fewer parameters. Such a procedure seems likely to reject the null hypothesis that there is no difference between the two models too often. In a first application of the built-in LRT, Miura and Edwards (2001) successfully compared several scenarios and could exclude some but not all alternative models.

I describe a different approach that seems more appropriate but is much more time consuming and might be prohibitive without good computing resources. Carstens *et al.* (2005) described an even better, but even more expensive method to evaluate migration models. The reported likelihood in a single program run is a relative likelihood: it is relative to the likelihood of the last chain times an unknown constant. A procedure to make the runs for both models using the same unknown constant is outlined here:

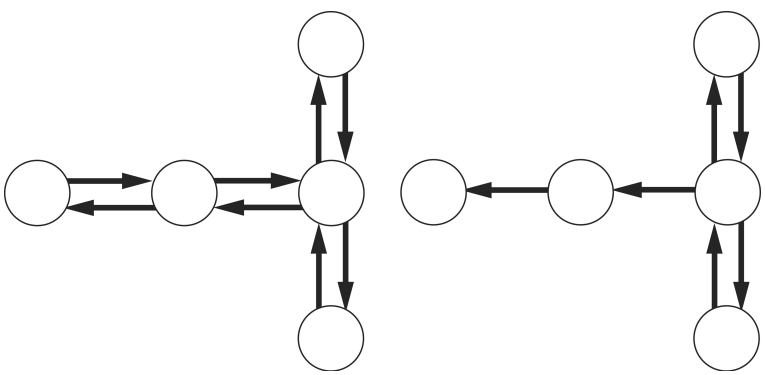


Figure 3.9. A possible, testable hypothesis: is gene flow between the islands and the mainland bidirectional (M_A) or unidirectional (M_B) resulting in the null hypothesis: $M_A = M_B$ and the alternative hypothesis $M_A \neq M_B$. Data are from the example data set; geography as in Fig. 3.3.

(1) Run data under model A; record parameters. This run needs to sample the MCMC chains appropriately and needs to be run for many steps (compare with section 'How long to run').
(2) Run data under model B; record parameters. This run needs to sample the MCMC chains appropriately and needs to be run for many steps (compare with section 'How long to run').
(3) Run data under model A for one very long single chain: No short chains, only one very long one, sampling, for example, the same number of genealogies as the total of run 1 or 2. Use the average parameter estimates from runs 1 and 2 for start parameters.
(4) Run data under model B for one very long single chain. Use the average parameter estimates from runs 1 and 2 for start parameters.
(5) Evaluate the likelihood ratio; calculate the degrees of freedom, which is the number of parameters that are different between the hypotheses; under some normality condition we can compare the LRT statistic with a χ^2 distribution with the same degree of freedom. MIGRATE calculates the probability of acceptance of the null hypothesis. Alternatively we can compare the LRT with tabulated χ^2 values for different significance levels typically printed in the Appendix of many introductory statistics texts.

The example data sets allow testing of whether there is only unidirectional migration from the mainland to Samos (the closest island) and from Samos to Ikaria. First, we set the model that allows for migration in both direction between mainland and the islands as the full model A (M_A) in which the unidirectional model B (M_B) is nested. Our null hypothesis specifies that there is no difference between the two models, and the alternative hypothesis is that the two models are different.

$$\begin{aligned} \mathrm{LRT} &= -2\,\mathrm{In}(L(D|\mathrm{B})/\,L(D|\mathrm{A})) \\ &= -2[\mathrm{In}L(D|\mathrm{B}) - \log L(D|\mathrm{A})] \\ &= -2(144.767 - 149.162) = 8.79, \quad (p = 0.012, \mathrm{df} = 2) \end{aligned}$$

where LRT is the likelihood ratio test statistic for the two models. The probability that improvement in likelihood for model B is caused by chance is small. Therefore, we reject the null hypothesis that assumes equality of model M_A and M_B. Therefore, we should use the full model (having a higher likelihood) and not the smaller model. We used a fair number of parameters and in these cases the likelihood ratio test may be conservative (Burnham and Anderson 2002). In addition, the LRT assumes nested hypotheses, whereas other model selection criteria, such as Akaike's information

criterion (AIC: Akaike 1972) or Schwartz' Information Criterion (the Bayesian information criterion – BIC: Schwartz 1978) can be applied to nested and non-nested models. These information criteria use the number of parameters to penalize the likelihood ratio favoring models with fewer parameters. Burnham and Anderson (2002) gave an extensive discussion of LRT, AIC and other information criteria and suggested using a version of AIC that corrects for small sample size, the AIC_c (Hurvich and Tsai 1989). Applying AIC and AIC_c to the models A and B we get the following values:

$$\begin{aligned}
AIC^{(A)} &= -2\ln L(D|A) + 2k_A = -2 \times 149.162 + 2 \times 13 = -272.32 \\
AIC^{(B)} &= -2\ln L(D|B) + 2k_B = -2 \times 144.767 + 2 \times 11 = -267.53 \\
AIC_c^{(A)} &= -2\ln L(D|A) + 2k_A c_A = -2 \times 149.162 \\
&\quad + 2 \times 13 \times 31/(31 - 13 - 1) = -250.91 \\
AIC_c^{(B)} &= -2\ln L(D|B) + 2k_A c_B = -2 \times 144.767 \\
&\quad + 2 \times 11 \times 31/(31 - 13 - 1) = -253.639
\end{aligned}$$

where k_i is the number of parameters in the model i, n_L is the number of samples, and the small sample correction factors $c_A = n_L/(n_L - k_A - 1)$ and $c_B = n_L/(n_L - k_B - 1)$. For AIC_c , I chose the number of loci in the study as samples, ignoring the number of individuals in the study. It is not clear how to specify n_L when the samples are correlated. The different information criteria cannot be mixed for comparison. The model with the lowest score is the best model in the set. The example compares two models and using AIC we choose model M_A with a score of −272.32 over the model M_B with −267.32. Using AIC_c we choose model M_B with −253.639 over model M_B with −250.91. Burnham and Anderson (2002) suggested that for most cases we should use AIC_c because it corrects for small sample size and is equivalent to the original AIC with large sample sizes. For these data it might be a tough call to decide whether we should prefer the simpler model M_B as suggested by AIC_c or the full model M_A as suggested by the LRT. Given the large number of parameters in the models, the few informative loci, the quality of the data (many values missing), and the use of MCMC, it might be wise to explore both models further before concluding that there is no gene flow from the islands to the mainland.

The likelihoods are approximated by MCMC; it is important to show that the chains have converged and that one has sampled enough genealogies, either by replicated runs and/or convergence diagnostics. Replicated runs from random starting points (for example random genealogies and different parameter values) that arrive at similar estimates after long runs are most promising. Carstens *et al.* (2005) developed an even better method to

estimate a more accurate likelihood ratio test than the procedure shown, but their method is very time intensive and requires bootstrapping the LRT because the commonly used assumption that the test-statistic is χ^2 distributed might be incorrect; as a consequence, the null hypothesis will be rejected too often. In MIGRATE, the described LRT comparisons and the built-in LRT-approximation are used to justify the replacement of a more complicated model with a simpler model. In a worst-case scenario, we would use the test with too narrow confidence intervals and, therefore, inflated differences of the two likelihood values caused by insufficient MCMC runs or lack of congruence with a χ^2 distribution. The outcome would be conservative because we would reject the null hypothesis that the full model and the simpler model are equivalent, and we would stick with the more complicated (full) model.

Use of the coalescent in conservation genetics

In conservation genetics, most of the tools used with a single genetic sample in time are derivatives of the coalescence theory, and can be explained summary statistics based on the coalescent, or are simply derived expectations of the coalescent, for example F_{ST} -based measures (Slatkin 1991; Neigel 2002). One of the biggest concerns in conservation biology is the long-term maintenance of variability in a population and, therefore, large effective population sizes, but changes in population size are difficult to estimate. With a single locus, positive growth in exponential growth models is often reported, but this result is strongly biased (Felsenstein *et al.* 1999). Populations that fluctuate randomly are often not distinguishable from estimates of populations with constant population sizes, and so an analysis using a model assuming constant population size will trace an average population size that is influenced by recent generations.

Programs such as MIGRATE that assume constant population sizes over time, average the population size over time. Even programs such as LAMARC and IM, which allow for other models than constant population size through time average over time: LAMARC averages out fluctuations to fit an exponential growth model, and IM forces constant or linear growing population sizes before and after the population split. Only the program BEAST allows for changes of a set of time segments with different population sizes in the past for a single population and a single locus. It is very versatile in the treatment of past population size variability, but needs to allow the use of multiple loci to achieve precise results. With a constant population size model, the population size is averaged over the time interval between the date of the most recent common ancestor and the date of the sample. The

expected time of the most recent common (diploid) ancestor is $4N_e$ generations in the past. In large populations the average is, therefore, over a longer time than in small populations. The coalescence-based population genetic parameter estimates are based on the number of mutation events, and also the frequencies of these alleles in the populations. Therefore, very recent changes in population size or migration rate are not necessarily visible using genetic data. Still, these long-term estimates deliver baselines for further management of these populations, for example protection or (moderate) harvest. For example, estimates of past population sizes of humpback whales estimated from mtDNA data (Roman and Palumbi 2003) are very different from current population sizes and from estimates using whaling logbooks. If the differences are real and not an artifact of the analysis, then management of whale populations should increase their protection. The whale study is based on a single locus, and further studies using multilocus data are urgently needed to corroborate Roman and Palumbi's findings. Using the probability distribution of the most recent common ancestor (Tavareé 1984) with the whaling logbook value as the true populations size of humpback whales reveals a tiny probability ($p < 10^{-10}$) for a population size value at the 2.5% quantile of Roman and Palumbi's data. This result suggests that it will be difficult to justify the logbook values even with multiple loci. Still, studies based on a single locus are easy to criticize because different population genetic forces can deliver similar signatures; for example, a small population size estimate can be the result of a population bottleneck, a long-term small population size, or a recent selective sweep. Only studies with multiple unlinked loci will be able to distinguish the selective sweep from the small effective population size. Recently, the program BEAST (Drummond *et al.* 2005) working with single-locus sequence data from a single population is able to estimate population size changes over time using samples from different times.

Researchers often contrast results from census sizes (N_c) with effective population sizes (N_e) using the ratio of N_e/N_c. In some marine fishes these ratios are very small (for example Turner *et al.* 2002). We can interpret this result in a variety of ways, including the following:

- The population size today as measured with the census size could have increased strongly in the last generation or two, so that there are not enough new mutations to see this same increase in the effective population size measured by genetic variability. Given the dire situation for most species this is a rather unlikely scenario, and can be excluded

rather easily with a historical observation that does not need to be based on genetics, although randomly fluctuating population size over genealogical time scale could well explain the difference.

- The effective population size and the census size are measured on a different population scale: census size is measured over a structured population and the genetic measurements came only from a single subpopulation. This is a highly unlikely scenario, even with unknown structure.
- Very few individuals have far more offspring than others. This will result in a small effective population size, and if the carrying capacity is large, large numbers of closely related individuals could be maintained. A comparison of multiple species with known life histories should reveal that when this sweepstakes scenario is correct, we would expect a correlation between number of eggs and ratio of N_e/N_c.

It will be important to explore these effects of high variance of reproduction success on the estimates of population sizes not only practically but also theoretically (Eldon and Wakeley 2006).

SUMMARY

Many powerful new methods for population genetic analysis have been developed in recent years. Almost all of them use heuristic techniques to calculate probabilities of model parameters given the observed data. Researchers that use such methods not only need to explore the variability in their data, but need to understand the variance introduced by the heuristic strategy. In this chapter, I have tried to point to ways that can help to minimize the error introduced by MCMC. The most important lesson is that such programs need to be run for a long time. If a convergence diagnostic is supplied, use it, but remember that convergence diagnostics only detect the grossest errors. Sometimes the diagnostic shows convergence, but the parameter estimates of interest still are not optimal. Run the program multiple times increasing run length. If you get different results, then you either need to run longer or resort to use MCMCMC. Replication is only useful when you have multiple computers to distribute the work. If you get different results using different prior distributions, try to understand why. Possible sources of the problem, ordered from the least likely to the most likely, are: (a) programming error; (b) in BI: bounds of priors are misspecified; in ML: driving values are not at equilibrium; (c) program has not been run long enough.

ACKNOWLEDGEMENTS

I thank Thomas Uzzell, two anonymous reviewers, Nathaniel K. Jue, Sonali Joshi and Michal Palczewski for many helpful comments and Heidi Hauffe for her patience. This work was supported by grant GM078985–01 of the National Institutes of Health through the joint NSF/NIGMS Mathematical Biology Program.

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**, 837–851.

Akaike, H. (1972). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*. Budapest: Akademiai Kaidó, pp. 267–281.

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

Beerli, P. (1994). Genetic isolation and calibration of an average protein clock in western Palearctic water frogs of the Aegean region. Ph.D. thesis, University of Zurich (http://www.scs.fsu.edu/~beerli/ownpapers/phd-thesis-beerli-1994.pdf).

Beerli, P. (1997). Migrate v. 0.3: a maximum likelihood program to estimate gene flow using the coalescent. http://popgen.scs.fsu.edu

Beerli, P. (1998). Estimation of migration rates and population sizes in geographically structured populations. In *Advances in Molecular Ecology*, NATO Science Series A: Life Sciences vol. 306, ed. G. R. Carvalho. Amsterdam: IOS Press, pp. 39–53.

Beerli, P. (2004). Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Molecular Ecology*, **13**, 827–836.

Beerli, P. (2006). Comparison of Bayesian and maximum likelihood inference of population genetic parameters. *Bioinformatics*, **22**, 341–345.

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 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 USA*, **98**, 4563–4568.

Beerli, P., Hotz, H. and Uzzell, T. (1996). Geologically dated sea barriers calibrate a protein clock for Aegean water frogs. *Evolution*, **50**, 1676–1687.

Blum, M. G. B., Damerval, C., Manel, S. and François, O. (2004). Brownian models and coalescent structures. *Theoretical Population Biology*, **65**, 249–261.

Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society Series D*, **47**, 69–100.

Brumfield, R. T., Beerli, P., Nickerson, D. A. and Edwards, S. V. (2003). The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology and Evolution*, **18**, 249–256.

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*, 2nd edn. New York: Springer Verlag.

Burns, G., Daoud, R. and Vaigl, J. (1994). LAM: an open cluster environment for MPI. In *Proceedings of Supercomputing Symposium*, pp. 379–386.

Calabrese, P. and Sainudiin, R. (2005). Models of microsatellite evolution. In *Statistical Methods in Molecular Evolution*, ed. R. Nielsen. New York: Springer-Verlag, pp. 289–305.

Carstens, B., Bankhead, A., Joyce, P. and Sullivan, J. (2005). Testing population genetic structure using parametric bootstrapping and migrate-n. *Genetica*, **124**, 71–75.

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H. and Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, **15**, 1496–1502.

Cornuet, J. and Luikart, G. (1996). Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*, **144**, 2001–2014.

De Iorio, M. and Griffiths, R. C. (2004). Importance sampling on coalescent histories. II: Subdivided population models. *Advances in Applied Probability*, **36**, 434–454.

Drummond, A., Rambaut, A., Shapiro, B. and Pybus, O. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, **22**, 1185–1192.

Eldon, B. and Wakeley, J. (2006). Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, **172**, 2621–2633.

Ewens, W. J. (2004). *Mathematical Population Genetics*. New York: Springer-Verlag.

Ewing, G. and Rodrigo, A. (2006). Coalescent-based estimation of population parameters when the number of demes changes over time. *Molecular Biology and Evolution*, **23**, 988–996.

Felsenstein, J. (2004). *Phylogenetic Inference*. Sunderland: Sinauer Associates.

Felsenstein, J., Kuhner, M. K., Yamato, J. and Beerli, P. (1999). Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. In *Statistics in Molecular Biology and Genetics*, ed. F. Seillier-Moiseiwitsch. Hayward, California: Institute of Mathematical Statistics and American Mathematical Society, pp. 163–185.

Fleming, K., Johnston, P., Zwartz, D., Yokoyama, Y., Lambeck, K. and Chappell, J. (1998). Refining the eustatic sea-level curve since the last glacial maximum using far-and intermediate-field sites. *Earth and Planetary Science Letters*, **163**, 327–342.

Fleming, K. M. (2000). Glacial rebound and sea-level change constraints on the Greenland ice sheet. Ph.D. thesis, Australian National University.

Fu, Y. X. (2006). Exact coalescent for the Wright–Fisher model. *Theoretical Population Biology*, **69**, 385–394.

Gabriel, E., Fagg, G. E., Bosilca, G., *et al.* (2004). Open MPI: Goals, concept, and design of a next generation MPI implementation. In *Proceedings, 11th European PVM/MPI Users' Group Meeting*, Budapest, Hungary, pp. 97–104.

Geyer, C. J. (1991). *Estimating Normalizing Constants and Reweighting Mixtures in Markov Chain Monte Carlo*, Technical Report No. 568. St Paul: School of Statistics, University of Minnesota.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte-Carlo maximum-likelihood for dependent data. *Journal of the Royal Statistical Society Series B*, **54**, 657–699.

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, **90**, 909–920.

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

Gropp, W., Lusk, E. and Skjellum, A. (1999a). *Using MPI Portable Parallel Programming with the Message-Passing Interface: Scientific and Engineering Computation*, 2nd edn. Cambridge, Mass.: MIT Press.

Gropp, W., Lusk, E. and Thakur, R. (1999b). *Using MPI-2 Advanced Features of the Message-Passing Interface: Scientific and Engineering Computation*. Cambridge, Mass.: MIT Press.

Hasegawa, M., Kishino, K. and Yano, T. (1985). Dating the human–ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160–174.

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

Hein, J., Schierup, M. H. and Wiuf, C. (2005). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford: Oxford University Press.

Hey, J. (2005). On the number of new world founders: a population genetic portrait of the peopling of the Americas. *PLoS Biology*, **3**, 965–974.

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**, 747–760.

Hudson, R. R. (1991). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, **7**, 1–44.

Hudson, R. R. and Kaplan, N. L. (1988). The coalescent process in models with selection and recombination. *Genetics*, **120**, 831–840.

Hudson, R. R., Slatkin, M. and Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, **132**, 583–589.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Kaplan, N. L., Darden, T. and Hudson, R. R. (1988). The coalescent process in models with selection. *Genetics*, **120**, 819–829.

Kimura, M. and Crow, J. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, **49**, 725–738.

Kimura, M. and Ohta, T. (1978). Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proceedings of the National Academy of Sciences USA*, **75**, 2868–2872.

Kingman, J. (1982a). The coalescent. *Stochastic Processes and Their Applications*, **13**, 235–248.

Kingman, J. (1982b). On the genealogy of large populations. In *Essays in Statistical Science*, ed. J. Gani and E. Hannan. London: Applied Probability Trust, pp. 27–43.

Kingman, J. F. (2000). Origins of the coalescent: 1974–1982. *Genetics*, **156**, 1461–1463.

Kuhner, M. (2006). LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics*, **22**, 768–70.

Kuhner, M. and Smith, L. (2006). Comparing likelihood and Bayesian coalescent estimators of population parameters. *Genetics*, **75**, 155–165.

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

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

Milne, G. A., Long, A. L. and Basset, S. E. (2005). Modelling Holocene relative sea-level observations from the Caribbean and South America. *Quaternary Science Reviews*, **24**, 1183– 1202.

Miura, G. I. and Edwards, S. V. (2001). Cryptic differentiation and geographic variation in genetic diversity of Hall's babbler *Pomatostomus halli*. *Journal of Avian Biology*, **32**, 102–110.

Möhle, M. (2000). Ancestral processes in population genetics: the coalescent. *Journal of Theoretical Biology*, **204**, 629–638.

Möhle, M. and Sagitov, S. (2003). Coalescent patterns in diploid exchangeable population models. *Journal of Mathematical Biology*, **47**, 337–352.

Neigel, J. E. (2002). Is F_{ST} obsolete? *Conservation Genetics*, **3**, 167–173.

Neuhauser, C. and Krone, S. M. (1997). The genealogy of samples in models with selection. *Genetics*, **145**, 519–534.

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

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

Nielsen, R. and Signorovitch, J. (2003). Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theoretical Population Biology*, **63**, 245–255.

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology*, **29**, 59–75.

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

Pitman, J. (1999). Coalescents with multiple collisions. *Annals of Probability*, **27**, 1870–1902.

Posada, D. and Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

Roman, J. and Palumbi, S. (2003). Whales before whaling in the North Atlantic. *Science*, **301**, 508–510.

Sanderson, M. (2002). Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution*, **19**, 101–109.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Schweinsberg, J. (2000). Coalescents with simultaneous multiple collisions. *Electronic Journal of Probability*, **5**, 1–50.

Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genetical Research*, **58**, 167–175.

Slatkin, M. (2005). Seeing ghosts: the effect of unsampled populations on migration rates estimated for sampled populations. *Molecular Ecology*, **14**, 67–73.
Squyres, J. M. and Lumsdaine, A. (2003). A Component Architecture for LAM/MPI. In *Proceedings, 10th European PVM/MPI Users' Group Meeting, number 2840 in Lecture Notes in Computer Science*. Venice, Italy, pp. 379–387.
Swofford, D. (2003). PAUP*: phylogenetic analysis using parsimony (*and other methods), v. 4. Sunderland: Sinauer Associates.
Swofford, D., Olsen, G., Waddell, P. and Hillis, D. (1996). Phylogenetic inference. In *Molecular Systematics*, ed. D. Hillis, C. Moritz and B. Mable. Sunderland: Sinauer Associates, pp. 407–514.
Tavareé, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, **26**, 119–164.
Turner, T. F., Wares, J. P. and Gold, J. R. (2002). Genetic effective size is three orders of magnitude smaller than adult census size in an abundant, estuarine-dependent marine fish (*Sciaenops ocellatus*). *Genetics*, **162**, 1329–1339.
Wakeley, J. and Takahashi, T. (2003). Gene genealogies when the sample size exceeds the effective population size of the population. *Molecular Biology and Evolution*, **20**, 208–213.
Wakeley, J., Nielsen, R., Liu-Cordero, S. N. and Ardlie, K. (2001). The discovery of single-nucleotide polymorphisms – and inferences about human demographic history. *American Journal of Human Genetics*, **69**, 1332–1347.
Wilkinson-Herbots, H. M. (1998). Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology*, **37**, 535–585.
Wilson, I. J., Weale, M. E. and Balding, D. J. (2003). Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society Series A*, **166**, 155–188.
Wilson, L., Stephens, D. A., Harding, R. M., *et al.* (2000). Inference in molecular population genetics: discussion. *Journal of the Royal Statistical Society Series B*, **62**, 636–655.
Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

Appendix

Data set

This chapter used a data set from my thesis (Beerli 1994) as an example how I would analyse such a dataset. The data set is imperfect, probably like most real data set, and simple but complex enough to highlight difficulties in its analysis.

```
5 31 Electrophoretic loci data: Anatolian water frogs
16 SELCUK
SELC ?1072 EE BB BB BB BB AA AA BC BB ?? ?? CD AA ?? AA CC
   DD DD BB BB ?? CC BB AA AA ?? EE BD ?? BC
```

BB
SELC ?1071 EE BB BB BB AB AA AA CJ BB ?? ?? CC AA ?? AC CC DD DD BB BB ?? CC BB AA AA ?? EE BB ?? BB
BB
SELC ?1074 EE BB BB BB AB AA AA BC BB ?? BB CC AA ?? BC CC DD DD BB BB ?? CC BB AA AA BB EE BD ?? BC
BB
SELC ?1073 EE BB BB BB AA AA AA CC BB ?? ?? CC AA ?? CC CC DD DD BB BB ?? AC BB AA AA ?? EE BB ?? BC
BB
SELC ?1068 EE BB BB BB AA AA AA CC BB ?? ?? CC AA ?? ?? CC DD DD BB BB ?? CC BB AA AA ?? EE BB ?? BC
BB
SELC ?1067 EG BB BB BB AB AA AA CJ BB ?? ?? CC AA BB BB CC CD DD BB BB ?? CC BB AA AA ?? EE DD ?? BB
BB
SELC ?1070 EE BB BB BB AB AA AA BC BB ?? BB CC AA ?? CC CC DD DD BB BB ?? CC BB AA AA BB EE DD ?? BC
BB
SELC ?1069 EE BB BB BB AB AA AA CC BB ?? ?? CC AA ?? BC CC CD DD BB BB ?? AC BB AA AA ?? EE BB ?? BB
BB
SELC ?1080 EE BB BB BB BB AA AA ?? BB ?? ?? CC AA ?? CC CC DD DD BB BB ?? AC BB AA AA ?? EE DD ?? BB
BB
SELC ?1079 EE BB BB BB AA AA AA CC BB ?? ?? DD AA ?? CC CC DD DD BB BB ?? CC BB AA AA ?? EE DD ?? BC
BB
SELC 17164 EE BB BB BB AA AA ?? CC BB ?? ?? CC AA ?? BB CC ?? ?? BB BC ?? ?? BB AA ?? ?? EE BB ?? BB
BB
SELC 17163 EE BB ?? BB BB AA ?? CC BB ?? ?? CC AA ?? CC CC ?? ?? BB BB ?? ?? BB AA ?? ?? EE BB ?? BB
BB
SELC ?1076 EE BB BB BB AA AA AA CC BB ?? ?? CD AA ?? CC CC DD DD BB BB ?? CC BB AA AA ?? EE BD ?? BC
BB
SELC ?1075 EE BB BB BB AA AA AA CC BB ?? ?? CC AA ?? CC CC DD DD BB BB ?? AC BB AA AA ?? EE BD ?? BB
BB

SELC ?1078 EE BB BB BB AB AA AA BC BB ?? ?? CC AA ?? CC CC
DD DD BB BB ?? AC BB AA AA ?? EE DD ?? BB
BB
SELC ?1077 EE BB BB BB AA AA AA CC BB ?? BB CC AA ?? AC CC
DD DD BB BB ?? AA BB AA AA BB EE BD ?? BB
BB
15 AKCAPINAR
AKCA 16804 EE BB AB BB ?? AA AA CC BB ?? ?? CC AA BB BB ??
DD DD BB BB AA AC BB AA AC BB EE DD AA CC
BB
AKCA ?1065 EE BB BB BB EE AA AA CC BB ?? ?? CC AA ?? DD CC
DD DD ?? BC AA AC BB AA AA BB EE DD ?? CC
BB
AKCA ?1064 EE BB BB BB AB AA AA BB BB ?? ?? CC AA ?? DD CC
DD DD ?? CC AA CC BB AA AA BB EE DD ?? CC
BB
AKCA 16805 ?? ?? BB ?? ?? AA AA BC BB ?? ?? CC AA BB BB ?? DE
DD BB CC AA AC BB AA AA BB EE DD AA CC
BB
AKCA 16808 ?? ?? AB ?? BE AA ?? ?? BB ?? ?? ?? ?? BB ?? ?? DD DD BB
BB AA CC BB AA AA ?? ?? DD ?? CC
BB
AKCA 16807 ?? ?? ?? BB AB AA ?? ?? BB ?? ?? ?? ?? BB ?? ?? DD DD BB
BC AA AC BB AA AA ?? ?? DD ?? CC
BB
AKCA 16806 EE BB BB BB ?? AA AA BJ BB ?? ?? CC AA BB BC ?? DD
DD BB BB AA CC BB AA AA BB EE DD AA CC
BB
AKCA ?1063 EE BB BB BB AB AA AA BC BB ?? ?? CC AA BB DD CC
DD DD BB BC AA CC BB AA AA BB EE DD ?? CC
BB
AKCA ?1058 EE BB BB BB AB AA AA ?? BB ?? BB CC AA BB CD CC
?? DD BB BC AA AA BD AA AC BB ?? DD ?? CC
BB
AKCA ?1057 EE BB AB BB BB AA AA ?? BB ?? BB CC AA BB DD CC
?? DD BB BE AA AA BB AA AJ BB ?? DD ?? CC
BB
AKCA ?1066 EE BB AB BB BB AA AA ?? BB ?? ?? CC AA ?? CD CC
DD DD BB BB AA ?? BB AA ?? BB EE ?? ?? CC
BB

AKCA ?1059 EE BB AA BB AA AA AA ?? BB ?? BB CC ?? BB BC CC ??
DD BB BB AA AC BB AA AJ BB ?? DD ?? CC
BB
AKCA ?1062 EE BB AA BB AB AA AA BC BB ?? ?? CC AA BB DD CC
DD DD BB BC AA CC BB AA AA ?? EE DH ?? CC
BB
AKCA ?1061 EE BB BB BB AB AA AA BC BB ?? ?? CC AA BB CC CC
DD DD BB BB AA CC BB AA AA BB EE DH ?? CC
BB
AKCA ?1060 EE BB BB AB BE AA AA ?? BB ?? ?? CC ?? BB DD ?? ??
DD BB BB AA AC BB AA AC BB ?? DD ?? CC BB
11 EZINE
EZIN?1081 EE BB BB BB ?? AA AA CC BB ?? ?? CC AA ?? BC CC DD
DD BB ?? AA CC BB AA AA BB EE BB AA BB
BB
EZIN 16782 EE BB ?? BB BB AA ?? CC BB ?? BB CC ?? ?? BB CC ??
DD BB BB ?? ?? BB AA AA BB EE BB AA ??
??
EZIN 16781 EE BB BB BB AA AA ?? CC BB ?? BB CC ?? BB BB CC CD
DD BB BB AA AC BB AA CC BB CE BD AA BB
BB
EZIN 16783 EE BB BB ?? BB AA ?? CC BB ?? ?? CD ?? ?? BB CC DD
DD BB BC AA CC BB AA AA ?? EE BD AA BB
BB
EZIN 16785 ?? BB BB ?? AB AA ?? CC BB ?? ?? CC AA ?? BB CC DD
DD BB BB AA AC BB AA AC BB EE BD AA BB
BB
EZIN 16784 EE BB BB ?? ?? AA ?? CC BB ?? ?? CC AA ?? BB CC DD
DD BB BC AA CC BB AA AC ?? CE BD AA BB
BB
EZIN ?1083 EE BB BB BB AB AA AA CC BB ?? ?? CC AA ?? BB CC
CD DD BB BB AA CC BB AA AA BB EE BB AA BB
BB
EZIN ?1082 EE BB AB BB AA AA AA CC BB ?? ?? CC AA BB BB CC
DD DD BB BB AA CC BB AA AA BB EE BB AA BB
BB
EZIN ?1084 EE BB AB BB AA AA AA CC BB ?? ?? CC AA ?? BC CC
CD DD BB BB AA CC BB AA AC AB EE BB AA BB
BB

EZIN 16780 ?? BB AB ?? AB AA ?? CC BB ?? ?? CC AA BB BB CC DD DD BB BB AA AC BB AA AA BB EE BD AA BB
BB
EZIN ?1085 EE BB AB BB BB AA AA CC BB ?? ?? CC AA BB BB CC BB DD BB BB ?? CC BB AA AC BB EE BB AA BB BB
11 IKARIA
IKAR 17331 EE BB BB BB AA AA ?? CC BB ?? ?? CD AA BB DD ?? DD DD BB BB AA CC BB AA AJ ?? EE BB AA CC
BB
IKAR 17330 EE BB BB BB AA AA ?? CC BB BB ?? CD AA BB DD ?? DD DD BB BB AA CC BB AA AA ?? EE BB AA CC
BB
IKAR 17332 EE BB BB BB AA AA ?? CC BB BB BB CD AA BB DD ?? ?? DD BB BB AA CC BB AA AA ?? EE BD AA CC
BB
IKAR 17379 EE BB BB ?? AA AA ?? CC ?? ?? ?? CC AA ?? DD CC DD ?? ?? BB AA CC BB AA AA BB EE BD AA CC
BB
IKAR 17378 EE BB BB BB AA AA ?? CC ?? ?? BB DD AA ?? DD CC DD ?? BB BB AA CC BB AA AA BB EE BB AA CC
BB
IKAR 17329 EE BB BB BB AA AA ?? CC BB BB ?? DD AA BB DD ?? DD AD BB BB AA CC BB AA AA ?? EE DD AA CC
BB
IKAR 17325 EE BB BB BB AA AA ?? CC ?? BB ?? DD AA ?? CC ?? DD AD ?? BB AA CC BB AA AA ?? EE BB AA CC
BB
IKAR 17324 EE BB BB BB AA AA ?? CC ?? BB BB DD AA ?? DD CC DD DD BB BB AA CC BB AA AA BB EE BB AA CC
BB
IKAR 17326 EE BB BB BB AA AA ?? CC ?? BB BB DD AA ?? DD ?? DD DD ?? BB ?? CC BB AA AA ?? EE BB AA CC
BB
IKAR 17328 EE BB BB BB AA AA ?? BC BB BB ?? CD AA BB DD ?? DD DD BB BB AA CC BB AA AA ?? EE BB AA BC
BB
IKAR 17327 EE BB BB BB AA AA ?? CC ?? BB BB CD AA ?? CD CC DD DD ?? BB ?? CC BB AA AA BB EE BD AA CC
BB

```
4 SAMOS
SAMO 17320 EE BB AB BB AA AA ?? CC BB ?? ?? DD AA BB BB ??
   DD DD ?? BB AA AC BB AA AA BB EE DD AA BC
BB
SAMO 17321 EE BB BB BB AA AA ?? CC BB BD BB DD AA BB DD
   CC DD ?? ?? BB AA CC BB AA AA BB EE DD AA BC
BB
SAMO 17323 EE BB AB BB AA AA ?? CC BB BB ?? ?? AA BB BB CC
   DD DD BB BB AA AC BB AA AA ?? EE DD AA CE
BB
SAMO 17322 EE BB AB BB AA AA ?? ?? BB BD BB CD AA BB DD ??
   DD DD ?? BB AA AA BB AA AA AA EE DD AA CC BB
```

Run conditions for specific examples in this chapter

Figure 3.2: For each of the six panel a data set for a single population with 50 sampled individuals, each with 10 unlinked loci, each 10 000 bp long, was generated.

MIGRATE was run using the Bayesian inference mode. The runs were done on a computer cluster with one master and 10 compute nodes. Four parallel heated chains using an adaptive heating scheme were run for each locus. Each chain sampled 10 000 MCMC updates of parameters and genealogies every 200 steps, after discarding the first 100 000 updates. Only the values of the cold chains were used for the posterior distributions. Each run took about 10 minutes.

Figure 3.5: For each of the three panels MIGRATE was run twice, first with an mtDNA data set from 10 individuals of *Rana lessonae* (Plötner *et al.*, unpubl.) to generate the posterior distribution, and then with no data (all nucleotides were replaced by '?') to generate a sample from the prior distribution. Run condition: The runs were done on a computer cluster with one master and four compute nodes and combinations (replicates) of four parallel long chains, each chain sampled 10 000 MCMC-updates of parameters and genealogies every 200 steps, after discarding the first 10 000 updates. The optimal strategy to run this on a single computer would have been different: one long chain, sampling 40 000 every 200, and discarding 10 000. This would have run about four times longer. The prior distribution for the scaled population size Θ was uniform with bounds for (A) at 0 and 10, (B) 0 and 0.02, and (C) 0 and 0.1. The histograms were copied from the PDF result file and combined with the program Adobe Illustrator.

Figure 3.6: Allozyme data set was run on a parallel computer cluster with a total of 72 compute nodes for about 2.5 hours. The run used a customized migration matrix that allowed gene flow only between geographic neighbours, the distance between neighbours was adjusted using a geographic distance file. One cold chain and three heated chains were used during the run: temperatures were 1.0, 1.2, 3.0, and 6.0. Ten replicates of one long chain were used to visit 10 000 000 steps per locus and saving 50 000 steps (50% genealogy change trials, 50% parameter change trials). The recorded parameters were then used to generate the posterior distributions.

Figure 3.7: Allozyme data set was run on a parallel computer cluster with a total of 72 compute nodes for about 1.5 hours. The run used a customized migration matrix that allowed gene flow only between geographic neighbours, the distance between neighbours was adjusted using a geographic distance file. For each locus a total of 10 short chains each visiting 10 000 genealogies and using 500 to improve the driving values for the next chain. Finally, 3 long chains each visiting 100 000 sampling are used. The last chain delivers the MLE and profile likelihood curves shown in Fig. 3.7. To improve mixing, I used a heating scheme with four chains with temperatures of 1.0, 1.2, 3.0 and 6.0.