Analysis of geographically structured populations: Estimators based on coalescence

Peter Beerli
Department of Genetics, Box 357360,
University of Washington, Seattle WA 98195-7360,
Email: beerli@genetics.washington.edu

The rapid increase in the collection of population samples of molecular sequences, plus the great expansion of the use of microsatellite markers, makes it possible to investigate the patterns and rates of migration among geographically subdivided populations with much greater power than was previously possible. The difficulty with methods for analyzing these data has been that they do not allow the researcher to observe the genealogical tree of ancestry of the sampled sequences, but only make an estimate of it which has a great deal of uncertainty. Taking the uncertainty in our estimate of the genealogy into account is the major challenge for a proper statistical analysis of these data. The statistical approach of maximum likelihood is used to infer these rates and patterns, using the Markov Chain Monte Carlo (MCMC) method of computing the likelihoods. This method samples genealogies from the space of possible genealogies, using an acceptance-rejection method to concentrate the sampling in the regions which contribute most to the outcome. Even though the number of possible genealogies is vast, the MCMC sampling can avoid wasting computer time on possibilities that can have made little contribution to the observed outcome. This sampling of different genealogies in computing a likelihood for the parameters correctly accounts for our lack of knowledge of the true gene tree.

It can be shown that these ML-methods are superior to methods based on F_{ST}. Additionally, ML-methods can take into account variability in mutation rate and can estimate all relevant population parameters jointly and also analyze cases with different population sizes and migration rates. Comparison of different data types reveals that number of loci sampled is a key factor in reducing the variability of the parameter estimates.

The coalescent

Most current population genetics analyses are using theoretical findings of Sewall Wright and R. A. Fisher which were made in the early 20th century. Their work is based on a view which uses discrete generations of idealized individuals passing their genes to offspring in the next generation. This "looking forward" strategy implies that calculation of the probability of a given genotype is rather difficult. Kingman (1982a,b) formalized a "looking backward" strategy: the coalescent. Hudson (1990) and Donnelly and Tavaré (1997) give comprehensive reviews on the subject. Coalescence theory takes the relatedness of the sample into account, so it incorporates random genetic drift and mutation. This approach makes it very easy to calculate probabilities of a genealogy of a sample of individuals with a given effective population size, $\mathrm{P}(g|\Theta)$. Hudson (1990) and others showed that we can extend this single population approach to multiple populations and estimate migration rates and also that we can include other forces such as growth, recombination, and selection.

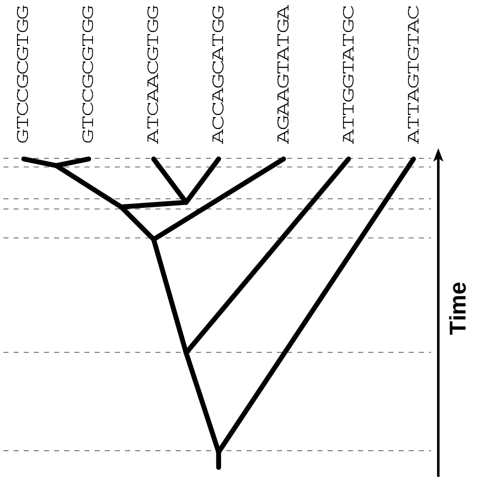

Figure 1: A coalescent tree with sampled sequences

Markov chain Monte Carlo (MCMC) integration

Construction of random genealogies (Simulation studies) is simple with the coalescent approach (e.g. the method of Slatkin and Maddison 1989). Inference of parameters is much harder, especially when we want not to lose any information in the data (Felsenstein 1992). In a likelihood framework we would like to simply integrate over all possible genealogies G and solve for the population parameters Θ at the maximum likelihood

$$\mathrm{L}(\Theta) = \int_{g \in G} \mathrm{P}(g|\Theta)\mathrm{P}(D|g)dg, \tag{1}$$

where $\mathrm{P}(D|g)$ is the likelihood of the genealogy with the sample data. This is not possible; there are too many different topologies with different branch lengths. But we can approximate by using a biased random walk through the genealogy space and then infer the parameters from the sampled genealogies correcting for the biased sampling:

$$\mathrm{L}(\Theta) = \int_{g \sim \mathrm{P}(g|\Theta_0)\mathrm{P}(D|g)} \frac{\mathrm{P}(g|\Theta)}{\mathrm{P}(g|\Theta_0)} dg \tag{2}$$

(MCMC: Hammersley and Handscomb 1964, MCMC and coalescence: Kuhner et al. 1996)

Table 1: Simulation with unequal known parameters of 100 two-locus datasets with 25 individuals in each population and 500 base pairs (bp) per locus. Std. dev. is the standard deviation.

	Population 1		Population 2	
	$4N_e\mu$	$4N_em$	$4N_e\mu$	$4N_em$
Truth	0.0500	10.00	0.0050	1.00
Mean	0.0476	8.35	0.0048	1.21
Std. dev.	0.0052	1.09	0.0005	0.15

Two population exchange migrants

We will explore the details of the MCMC mechanism in a simple two population model with the parameters: $\Theta_1 = 4N_e^{(1)}\mu$, $\Theta_2 = 4N_e^{(2)}\mu$, $\mathcal{M}_1 = m_1/\mu$, $\mathcal{M}_2 = m_2/\mu$ (we need to scale by the unknown mutation rate μ of our data).

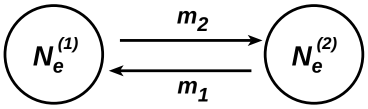

Figure 2: Two population model with population sizes $N_e^{(1)}$, $N_e^{(2)}$, and migration rates m_1, m_2.

- Assumptions: Population have constant size and exist forever, migration rate is constant through time, and the genetic markers are neutral.
- We can jointly estimate migration rates and population sizes
- Example of a simulation study (Table 1), where I generated 100 single locus data sets and then analyzed them with the program MIGRATE (Beerli 1997).
- Problems: perhaps not a natural situation; how long do we need to run the genealogy sampler?

Migration matrix model

- Assumptions: same as with 2 populations
- Simulation studies with (a) 4 sampled populations and (b) with 3 sampled population and one population where we don't have data.

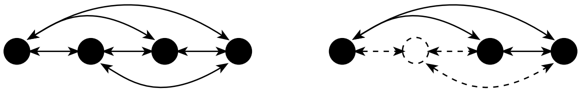

Figure 3: Population structure used in simulations.

- Problems: how many genealogies to sample? Number of parameters increases quadratically.

Comparison with F_{ST}

Simulation studies can show that the ML-estimator delivers better result than F_{ST}, and results are still accurate when population sizes and/or migration rates are unequal (Table 1).

Hypothesis testing using likelihood ratios

The maximum likelihood framework makes it easy to test hypotheses. I expect that these tests will supersede standard test based on F_{ST}. I will show a few examples and hope that I am able to have a version of MIGRATE finished in March so that everybody can experiment with their own data in the "data section".

$$H_0 : \hat{N}_e = N_e^{(x)}$$

$$\text{Test-statistic:} \qquad -2\log\left(\frac{L(\Theta_x)}{L(\hat{\Theta})}\right) \leq \chi^2_{df,\alpha}$$

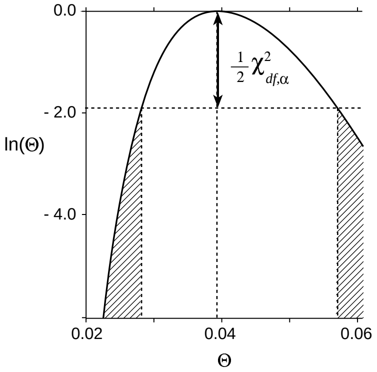

Figure 4: Likelihood ratio test: dashed areas are outside of the 95% confidence limit. Θ is $4N_e\mu$; $df = 1$, $\alpha = 0.05$

Data type and mutation rate

We have mutation models for infinite allele model, microsatellite stepwise mutation model (Valdez and Slatkin 1993, Di Rienzo et al. 1994), and finite sites sequence model (e.g. Swofford et al. 1996).

What's the effect of the data type to the estimate of migration rates? The data type is not that important, for the quality of the migration rate estimates, but the variance of the estimates is dependent on the number of unlinked loci (Fig. 5) having independent coalescent trees and the variability in the data, the more segregating sites or polymorphic loci are present the better the estimates of the migration rates.

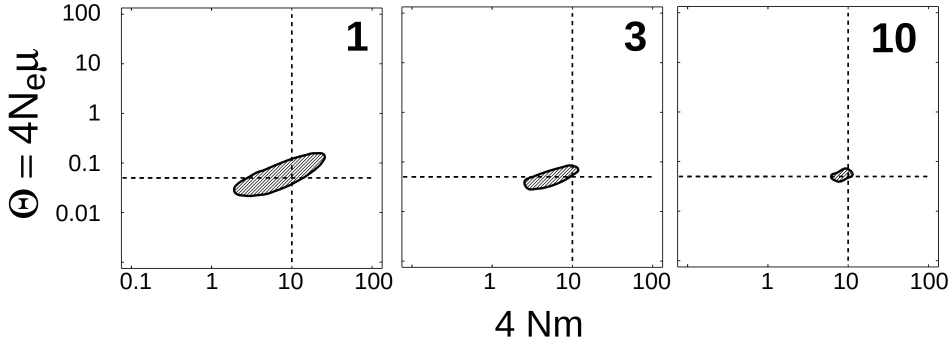

Figure 5: Variance of parameter estimates: the dashed area is the 95% confidence area, the numbers 1, 3, and 10 are the numbers of sampled loci

Mutation rate is not constant: incorporation of the variance of the mutation rate is possible by assuming that it follows a Gamma distribution (Fig. 6) and estimating the shape parameter α of this distribution jointly with the population parameters by integrating over all mutation rates x

$$\mathrm{L}(\Theta,\mathcal{M},\alpha)=\prod_{l}\int_{0}^{\infty}\frac{e^{-\alpha x/\Theta_1}x^{\alpha-1}}{\Gamma(\alpha)\left(\frac{\Theta_1}{\alpha}\right)^{\alpha}}\mathrm{L}(x,\Theta_l,\mathcal{M}_l)dx,$$

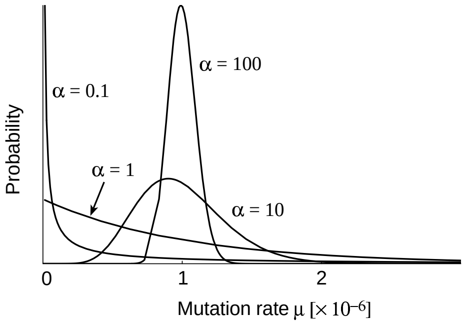

Figure 6: Gamma distributed mutation rates, with different shape parameter α and the same mean

Summary

- Coalescence theory enables us to estimate population parameters by including sample data and taking the possible histories of the populations into account.
- Expansion of the coalescence model to any migration model is possible.
- Maximum likelihood ratio test of arbitrary hypotheses.
- Multi-locus enzyme electrophoretic data and microsatellite markers delivers good migration rate estimates compared to mtDNA sequence data, because the quality of the result is dependent on the number of loci and the variability in the data.
- The assumption that the mutation rate over loci is constant is obviously wrong for electrophoretic markers and microsatellites and taking the variation of the mutation rate into account should improve the estimates of population parameters.

Bibliography

Citations with a ⋆ are recommended to read and/or introductory, citations with a • are rather difficult.

BEERLI, P., 1997, MIGRATE DOCUMENTATION version 0.3. Distributed over the Internet: http://evolution.genetics.washington.edu/lamarc.html.

DI RIENZO A., A. C. PETERSON, J. C. GARZA, A. M. VALDES, M. SLATKIN, and N. B . FREIMER, 1994, Mutational processes of simple-sequence repeat loci in human populations. Genetics **91** (8): 3166–3170.

⋆ DONNELLY, P. and S. TAVARÉ, 1997, *Progress in population genetics and human evolution*. IMA volumes in mathematics and its applications **87**, Springer, New York.

FELSENSTEIN, J., 1973, Maximum likelihood estimation of evolutionary trees from continuous characters. American Journal of Human Genetics **25**: 471–492.

FELSENSTEIN, J., 1988, Phylogenies from molecular sequences: inference and reliability. Annual Review of Genetics **22**: 521–565.

FELSENSTEIN, J., 1992, Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genetics Research **59**: 139–147.

• GRIFFITHS, R. and S. TAVARÉ, 1994, Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B Biol Sci **344** (1310): 403–10, Department of Mathematics, Monash University, Clayton, Victoria, Australia.

HAMMERSLEY, J. and D. HANDSCOMB, 1964, *Monte Carlo Methods*. Methuen and Co., London.

⋆ HUDSON, R. R., 1990, Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, vol. 7, pp. 1–44.

• 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*, edited by J. Gani and E. Hannan, pp. 27–43, Applied Probability Trust, London.

⋆ KUHNER, M., J. YAMATO, and J. FELSENSTEIN, 1995, Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140** (4): 1421–30, Department of Genetics, University of Washington, Seattle 98195-7360, USA.

METROPOLIS, N., A. ROSENBLUTH, M. ROSENBLUTH, A. TELLER, and E. TELLER, 1953, Equation of state calculations by fast computing machines. Journal of Chemical Physics **21**: 1087–1092.

NATH, H. and R. GRIFFITHS, 1993, The coalescent in two colonies with symmetric migration. Journal of Mathematical Biology **31** (8): 841–51.

NATH, H. and R. GRIFFITHS, 1996, Estimation in an Island Model Using Simulation. Theoretical Population Biology **50**: 227–253.

• NOTOHARA, M., 1990, The coalescent and the genealogical process in geographically structured population. Journal of Mathematical Biology **29** (1): 59–75.

SLATKIN, M., 1991, Inbreeding coefficients and coalescence times. Genetical Research **58** (2): 167–75, Department of Integrative Biology, University of California, Berkeley 94720.

⋆ SLATKIN, M. and W. MADDISON, 1989, A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics **123** (3): 603–613, Department of Zoology, University of California, Berkeley 94720.

⋆ SWOFFORD, D., G. OLSEN, P. WADDELL, and D. HILLIS, 1996, Phylogenetic Inference. In *Molecular Systematics*, edited by D. Hillis, C. Moritz, and B. Mable, pp. 407–514, Sinauer Associates, Sunderland, Massachusetts.

• Takahata, N., 1988, The coalescent in two partially isolated diffusion populations. Genetical Research **52** (3): 213–22.

• Takahata, N. and M. Slatkin, 1990, Genealogy of neutral genes in two partially isolated populations. Theoretical Population Biology **38** (3): 331–50, National Institute of Genetics, Mishima, Japan.

Valdez A. M. and M. Slatkin, 1993, Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133** (3): 737–749, Department of Zoology, University of California, Berkeley 94720.

Wakeley, J. and J. Hey, 1997, Estimating ancestral population parameters. Genetics **145** (3): 847–855.

Software, with emphasis on using the coalescent

[this list is certainly not complete]

- LAMARC package [**L**ikelihood **A**nalysis with **M**etropolis **A**lgorithm using **R**andom **C**oalenscence. Three programs are currently available: Coalesce, Fluctuate, and Migrate. C-source code and binaries for Windows, Mac, LINUX, DUNIX, NEXTSTEP.
  Website at `evolution.genetics.washington.edu/lamarc.html`

- MISAT estimates the effective population size of a single population using microsatellite data and can also test if the one-step model or a multi-step model is appropriate. Binaries for Macintosh and Windows.
  Website at `http://mw511.biol.berkeley.edu/software.html`

- Sites is a computer program for the analysis of comparative DNA sequence data (Hey and Wakeley, 1997. A coalescent estimator of the population recombination rate. Genetics 145: 833-846) . C source code and binaries for DOS and Macintosh.
  Website at `http://heylab.rutgers.edu`

- Upblue is a least square estimator for population size (Fu, Y. X., 1994. An phylogenetic estimator of effective population size or mutation rate. Genetics 136:685-692). Fortran program or use the website directly to calculate results
  `http://www.hgc.sph.uth.tmc.edu/fu/`

- Calculation of 4Nm using the method of Slatkin and Maddison (1989), you need to calculate the minimal mumber of migration events on the genealogy either by hand or using MacClade (Maddison and Maddison 1992, Sinauer). Pascal source code.
  Website at `http://mw511.biol.berkeley.edu/software.html`

- Several programs for the estimation of population size, exponential growth, recombination rate, migration rate, time of the last common ancestor. Contact Bob Griffiths (email: ...) for more information.