Inference of demographic parameters using importance sampling on coalescence history

Raphaël Leblois, M. de Iorio, R. Griffiths, François Rousset

Inference of demographic parameters using importance sampling on coalescence history

Raphaël Leblois and François Rousset

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

Université Montpellier 2, CNRS, Institut des Sciences de l'Evolution, France

Unité Origine, Structure et Evolution de la Biodiversité, Museum National d'Histoire Naturelle, Paris, France

State of our art

We developed and evaluated the performance of maximum likelihood (ML) analysis of allele frequency data in a linear array of populations under isolation by distance (IBD) and in a model of isolation with migration (IM) between two populations.

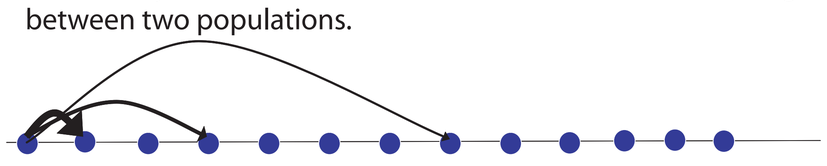

Fig. 1 : schematic representation of the linear IBD model

Inference in a linear habitat under IBD

The application on a linear IBD is the only algorithm that has been published so far (in Rousset & Leblois 2007 MBE). It has been validated using simulated and real data set analyses.

Under this IBD model, the parameters are a mutation rate and either a dispersal rate in a stepping stone model or a dispersal rate and a scale parameter in a geometric dispersal model.

An approximate procedure known as maximum product of approximate conditional (PAC) likelihood is found (i) to perform as well as ML (Fig. 2) and (ii) to be 10-1000x faster than ML, allowing the consideration of much larger habitat and more complex demographic models.

Linear IBD simulation results

Mis-specification biases may occur because the importance sampling algorithm is formally defined in term of mutation and migration rates scaled by the total size of the population, and this size may differ widely in the statistical model and in reality.

As expected, ML generally performs well when the statistical model is correctly specified. Otherwise, mutation rate estimates are much closer to mutation rate scaled by number of demes in the statistical model than scaled by number of demes in reality when mutation probability is high and dispersal is most limited. This mis-specification bias actually has practical benefits because it allows relatively unbiased estimation of the mutation rate scaled by deme size which could even provide estimates of mutation probability under the relatively mild assumption that deme sizes are known.

Migration rate estimates show roughly similar trends, but they may not always be easily interpreted as low-bias estimates of dispersal rate under any scaling.

Estimation of the dispersal scale parameter is much less precise and is also affected by mis-specification of the number of demes. However, the different biases compensate each other in such a way that good estimation of the so-called neighborhood size (or more precisely the product of population density and mean-squared parent–offspring dispersal distance : $4D\sigma^2$) is achieved.

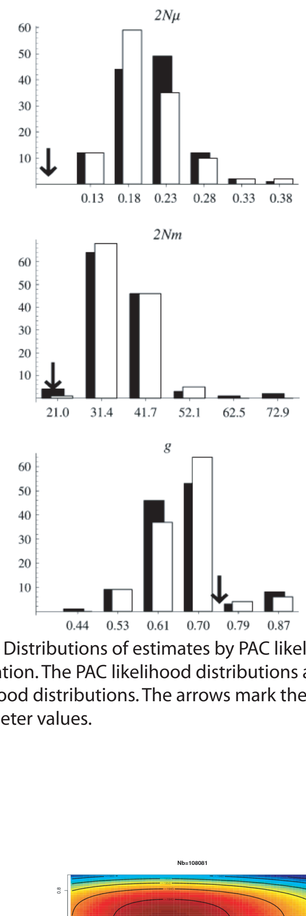

Fig. 2 - Distributions of estimates by PAC likelihood and IS estimation. The PAC likelihood distributions are laid over the likelihood distributions. The arrows mark the position of the parameter values.

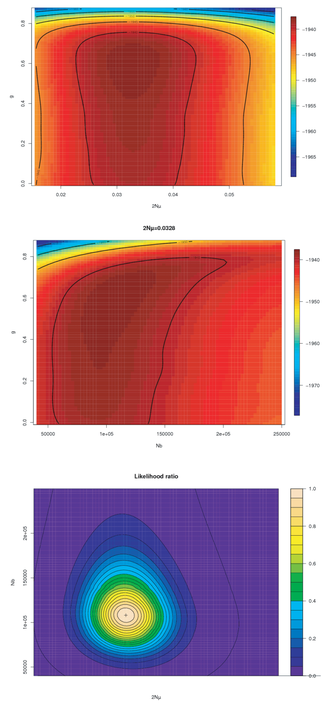

Fig. 5 : graphical output from Migraine on the damselfly example

Real data set analysis in a linear habitat :
the damselfly data set from Watts et al. 2007 Mol. Ecol.

In this application, we found results congruent with our simulation tests (Fig. 5) : estimation of the neighborhood size $4D\sigma^2$ seems to be relatively precise, as well as estimation of the population size scaled by mutation rate. The estimation of the dispersal scale parameter seems to be muc harder.

The damselfly example also alows us to compare our indirect ML genetic estimate with direct demographic estimate (from a CMR study) and another indirect genetic estimate, using the regression methods of Rousset (1997), of the same quantity $4D\sigma^2$. The different estimates, reported in Table 1, show good congruence between the 3 methods.

Implementation in the software MIGRAINE

The Algorithms used in this study are implemented in MIGRAINE, a free program for likelihood analyses of genetic data, with a focus on spatially structured populations. The demographic models currently implemented in this program are :
(1) a simple model of isolation by distance in a linear habitat, as described in Rousset & Leblois (2007), which includes the infinite island model as a special case.;
(2) an isolationwith migration (IM) model with two populations. Note that this IM algorithm is still under development but will soon be availlable.

At the heart of the program is the importance sampling algorithm defined by de Iorio & Griffths (2004). Some approximate procedures ("PAC-likelihood") are also available and have been shown to be at least as efficient than exact maximum likelihood for time constant models.

MIGRAINE is currently designed for allelic type data sets only : a K-alleles mutation model is implemented for all demographic models and a strict stepwise mutation model (SMM) is implemented to some extent for models with one or two populations (not yet published). However, we plan to adapt MIGRAINE for sequence data in a "near" future.

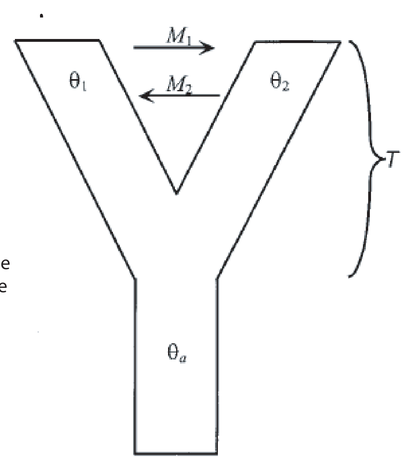

Fig. 3 : schematic representation of the IM model with 2 populations

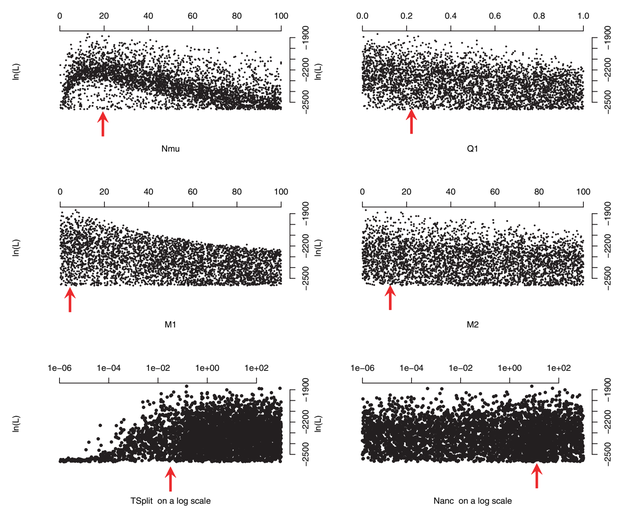

Fig. 4 : projection of likelihood estimates along each parameter using MIGRAINE under the IM model. The simulated data set analysed here was generated using parameter values represented by the red arrows placed on each graphic.

Preliminary results under the IM model

We implemented a simple IM model with two populations with potentialy different sizes and asymmetric migration rates. This model has 6 parameters which are the total population size scaled by mutation rate, the relative populations sizes, the two migration rates scaled by total population size, the time of divergence scaled by total population size and the ancestral population size scaled by mutation rate (Fig. 3).

Our first results on simulated data set are quite interesting because they show that :
(1) Our algorithm is able to infer all 6 parameters meaning that it will be able do detect asymetric population sizes and migration rates, event if the precision is not equal for all parameters.
(2) graphical output (likelihood points as shown on Fig. 4 and soon likelhood and likelihood ratio plots) can help a lot in understanding what information is available in the data. One illustration on Fig. 4 is that microsatellite data will not have much information on large divergence time and ancestral populations when the population split is not very recent. It is thus quite easy to understand why some parameters are more or less well estimated depending on the marker used and the time of divergence of the populations studied.

Table 1 : direct and indirect estimates of $4D\sigma^2$ using different methods on the damselfly data set

Direct estimate	Indirect estimates	
Capture-Mark-Recapture	Rousset 1997's Regression	MIGRAINE Maximum Lilelihood
277 894	179 058 (e estimator) 242 816 (a estimator) 95% CI : [66 015; 392 866]	108 081 [65 000;185 000]

Few other demographic models such as single population with size fluctuations in time, n-population with constant migration and the possibility to consider multiple samples in time are almost ready to be tested. We are looking for students or post-docs to develop those models. Do not hesitate to contact us if you are potentialy interested. Migraine will thus be regularly updated to consider more demographic and mutational models so

Check the Migraine web page regularly to find new model implementations and Download Migraine at :

http://kimura.univ-montp2.fr/~rousset/Migraine.htm

References :

Rousset F, Leblois R. 2007. Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model misspecification. Molecular Biology and Evolution 24: 2730–2745

De Iorio M, Griffiths RC. 2004. Importance Sampling on Coalescent Histories. Advances in Applied Probabilities 36: 417-434.

DeIorio M, Griffiths R, Leblois R, Rousset F. 2005. Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. Theoretical Population Biology 68: 41-53.

Rousset F. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145:1219–1228.

Watts P.C., F. Rousset, I.J. Saccheri, R. Leblois, S.J. Kemp, D.J. Thompson. 2007. Compatibility of genetic and demographic estimates of 'neighbourhood size' in insect populations: analysis of Coenagrion mercuriale (Odonata: Zygoptera) using an improved estimator of genetic divergence. Molecular Ecology 16, 737–751.

R. Leblois venue to SMBE 2008 was partially supported by a grant from the ANR Biodiversity : Endemism in New Caledonia : A phylogenetic and population-level study of its origins (MNHN, OSEB)