Molecular Ecology Resources (2008) doi: 10.1111/j.1755-0998.2008.02417.x

COMPUTER PROGRAMS

IBDSim: a computer program to simulate genotypic data under isolation by distance

RAPHAËL LEBLOIS*, ARNAUD ESTOUP† and FRANÇOIS ROUSSET‡
*Muséum National d'Histoire Naturelle, UMR 5202 CNRS/MNHN 'Origine, Structure et Evolution de la Biodiversité', Paris, France, †INRA, UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro), Campus International de Baillarguet CS 30 016 F-34988, Montferrier-sur-Lez cedex, France, ‡Université Montpellier 2, CNRS, Institut des sciences de l'Évolution, Montpellier, France

Abstract

IBDSim is a package for the simulation of genotypic data under isolation by distance. It is based on a backward 'generation by generation' coalescent algorithm allowing the consideration of various isolation by distance models with discrete subpopulations as well as continuous populations. Many dispersal distributions can be considered as well as heterogeneities in space and time of the demographic parameters. Typical applications of our program include (i) the study of the effect of various sampling, mutational and demographic factors on the pattern of genetic variation; and (ii) the production of test data sets to assess the influence of these factors on inferential methods available to analyse genotypic data.

Keywords: coalescence, dispersal distribution, isolation by distance, microsatellite DNA, simulation, spatial and temporal heterogeneities

Received 7 September 2008; revision accepted 8 September 2008

Introduction

IBDSim (version 1.0) is a computer package for the simulation of genotypic data under isolation by distance (IBD) models for a set of independent loci. IBDSim is based on the backward-in-time coalescent approach, which allows the generation of large data sets considering complex demographic scenarios. Contrary to the 'ms' and 'SimCoal2' software that also use a coalescent approach (Hudson 2002; Laval & Excoffier 2004), IBDSim is especially designed for simulation under isolation-by-distance models, and hence, can easily consider many different dispersal distributions under either a model with discrete subpopulations or a large continuous population.

Theoretical analysis of isolation-by-distance models was pioneered by Wright (1946) and Malécot (1948) and many authors have since used identity by descent to describe how the genetic structure develops under isolation by distance (Maruyama 1972; Sawyer 1977; Slatkin 1993; Rousset 1997). However, no analytical treatments of identity by descent, coalescence times or coalescence probability have been carried out for more than two genes, strongly limiting the potential of analytical studies for larger samples. Typical applications of our program include the study of the effect of sampling, mutational and demographic factors on the pattern of genetic variation and the production of test data sets to assess the influence of these factors on any inferential methods available to analyse genotypic data based on independent loci. The IBDSim program (executable and sources) as well as its user manual are freely available at http://kimura.univ-montp2.fr/~rousset/IBDSim.html.

Correspondence: Raphael Leblois, Unité Origine, Structure et Evolution de la Biodiversité, UMR CNRS/MNHN 5202, Département systématique et Evolution, Case Postale 39, Batiment de Phanérogamie, Muséum National d'Histoire Naturelle, 16 rue Buffon, 75005 Paris, France. Fax: +33 (0)1 40 79 33 42; E-mail: leblois@mnhn.fr

Principle of simulation algorithm

Because of the complexity of the models considered, we used an exact algorithm for which coalescence and migration events are considered generation by generation rather than the large-N approximations of the n-coalescent theory (Kingman 1982). Such algorithm leads to less efficient simulations in terms of computation time but is much more flexible when complex demographic and dispersal features

are considered. The generation-by-generation algorithm has been detailed in Leblois *et al*. (2003, 2004, 2006) and we just summarize below the main ideas underlying the global algorithm.

Demographic models and life cycle

The demographic models that can be considered are (i) the demic lattice model with each lattice node corresponding to a panmictic subpopulation of size N individuals, and (ii) the 'continuous' lattice model with each lattice node corresponding to one individual. For all simulations, the life cycle is divided into five steps: (i) each individual gives birth to a great number of gametes, and dies; (ii) gametes undergo the effect of mutations; (iii) gametes disperse; (iv) diploid individuals are formed, if necessary, by considering Hardy–Weinberg equilibrium within each deme, and (v) competition brings back the number of adults in each deme to N, with $N = 1$ for the continuous model and $N > 1$ for the demic model.

Generation-by-generation coalescent algorithm

In our algorithm, coalescence and migration events are considered 'generation by generation' until the common ancestor of the sample has been found. At each generation, the coordinates of the parents of our sampled genes are randomly assessed using $b_{dx,dy}$, the backward dispersal function (detailed in the next section). Once the position of the parents on the lattice is known, the coalescence events occurring at this generation are assessed. A coalescence event occurs if genes are both on the same lattice node and if they originate from the same parental gene. Multiple coalescences are allowed and the probability for a coalescence of k genes in a given parental gene is $1/2N^{k-1}$. Note that, since multiple coalescent events are taken into account, it allows to build an exact coalescent tree even for small values of N.

Backward and forward dispersal distributions

We used the 'backward' dispersal distribution because the position of the parental gene is determined knowing the position of its descendant gene. This 'backward' function is computed using $f_{dx,dy}$, the forward dispersal density function describing where descendants go, and assuming that dispersal is independent in each direction, so that $f_{dx,dy} = f_{dx} \times f_{dy}$. In the simplest case, considering that density is homogeneous in space, backward dispersal functions are equal to forward dispersal functions, so that $b_{dx,dy} = f_{dx,dy} = f_{dx} \times f_{dy}$. When density is not homogeneous in space, each lattice node has a backward distribution that depends on the density of each surrounding node (computation of the backward distribution in such cases is detailed in Leblois *et al*. 2004 and in the user manual).

Because biologically realistic dispersal functions often have a high kurtosis, we used different families of forward dispersal distributions for which suitable choice of their parameter values allows high kurtosis and high migration rates.

The first distributions are truncated variants of the discrete Pareto, or Zeta, distribution (see, e.g. Patil & Joshi 1968) with the probability of moving k steps in one direction being of the form:

$$f_k = f_{-k} = \frac{M}{k^n}, \qquad \text{(eqn 1)}$$

with parameters M and n, controlling the total dispersal rate and the kurtosis, respectively.

The second family of dispersal distributions is obtained as mixtures of convolutions of stepping-stone steps and is a convenient way to model discrete distributions with various forms (Chesson & Lee 2005). Parameterization of the Sichel mixture distribution is not trivial but details on each parameter and formulas to compute various moments of the distribution as well as its kernel are given in Chesson & Lee (2005). Examples that achieve some given second moment (σ) and kurtosis are given in Watts *et al*. (2007).

For convenience, we also considered geometric dispersal distributions for which the probability of moving k steps in one direction is:

$$f_k = f_{-k} = \frac{m}{2}(1-g)g^{k-1} \qquad \text{(eqn 2)}$$

with m controlling the total emigration rate and g the shape of the distribution. Note that (i) geometric distributions cannot be used to achieve high kurtosis with large migration rates; (ii) the stepping-stone dispersal is the limit of the geometric distribution with $g \to 0$.

Mutation models

Four theoretical mutation models are implemented in our program: (i) the infinite allele model (IAM, Kimura & Crow 1964); (ii) the K-allele model (KAM, Crow & Kimura 1970); (iii) the strict stepwise-mutation model (SMM, Ohta & Kimura 1973); and (iv) the generalized stepwise model (GSM, e.g. Estoup *et al*. 2002). Allele size constraints are included in our program for the SMM and GSM by imposing reflecting boundaries to the allele size range (e.g. Estoup *et al*. 2002). Finally, interlocus variability in the mutation rate is optionally modelled by drawing single locus mutation rate values in a gamma distribution with parameters (shape, scale) being $(2, \mu/2)$ where μ is the mean mutation rate specified by the user.

Heterogeneities in time and space

Demographic heterogeneity in time can be simulated with our program in various ways. Up to three discrete

changes can occur from present to the most recent common ancestor of each tree, each of this temporal changes being characterized by new demographic settings: lattice size, density of individuals on the lattice, demic structure and dispersal distribution. Demographic heterogeneity in space can be simulated by considering that a rectangular zone of the lattice have its specific density. Such settings allowed us to simulate, among other possibilities, temporal changes in dispersal, temporal changes in density with constant lattice size, temporal changes in lattice size with constant density, and presence of high- or low-density zones within the lattice (see Leblois *et al*. 2004, 2006 for examples of simulations under such complex demographic scenarios).

Input and output files

IBDSim reads one generic text file that allows one to control all options of IBDSim. A detailed explanation of the setting file is given in the user manual. Depending on the options chosen, IBDSim can generate different types of output files including all simulated data sets in three different formats read by GenePop version 4 (Rousset 2008), Migrate (Beerli & Felsenstein 2001) and Migraine (Rousset & Leblois 2007) as well as different summary statistic files with records, for each simulated data file and over all multilocus runs, of various genetic statistics (e.g. observed and expected heterozygosities, number of alleles, variance in allelic size, F_{IS}, coalescence times, effective dispersal events). Detailed information on each file generated by IBDSim is provided in the user manual.

Acknowledgements

We thank S. Piry for valuable code assistance; V. Ravigné, R. Vitalis and F. Halkett for constructive comments on the manuscript. This is publication ISEM–2008-096.

References

Beerli P, 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.

Chesson P, Lee CT (2005) Families of discrete kernels for modeling dispersal. *Theoretical Population Biology*, **67**, 241–256.

Crow JF, Kimura M (1970) *An Introduction to Population Genetics Theory*. Harper & Row, New York.

Estoup A, Jarne P, Cornuet JM (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology*, **11**, 1591–1604.

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

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

Kingman JFC (1982) The coalescent. *Stochastic Processes and their Applications*, **13**, 235–248.

Laval G, Excoffier L (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*, **20**, 2485–2487.

Leblois R, Estoup A, Rousset F (2003) Influence of mutational and sampling factors on the estimation of demographic parameters in a continuous population under isolation by distance. *Molecular Biology and Evolution*, **20**, 491–502.

Leblois R, Rousset F, Estoup A (2004) Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population from microsatellite data. *Genetics*, **166**, 1081–1092.

Leblois R, Estoup A, Streiff R (2006) Habitat contraction and reduction in population size: does isolation by distance matter? *Molecular Ecology*, **15**, 3601–3615.

Malécot G (1948) *Les Mathématiques de L'hérédité*. Masson, Paris.

Maruyama T (1972) Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. *Genetics*, **70**, 639–651.

Ohta T, 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.

Patil GP, Joshi SW (1968) *A Dictionary and Bibliography of Discrete Distributions*. Oliver & Boyd, Edinburgh, UK.

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

Rousset F (2008) GenePop'007: a complete re-implementation of the GenePop software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

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

Sawyer S (1977) Asymptotic properties of the equilibrium probability of identity in a geographically structured population. *Advances in Applied Probabilities*, **9**, 268–282.

Slatkin M (1993) Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, **47**, 264–279.

Watts PC, Rousset F, Saccheri IJ, Leblois R, Kemp SJ, Thompson DJ (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.

Wright S (1946) Isolation by distance under diverse systems of mating. *Genetics*, **31**, 39–59.