

PROGRAM NOTE

METASIM 1.0: an individual-based environment for simulating population genetics of complex population dynamics

ALLAN E. STRAND

*Department of Biology, College of Charleston, Charleston, SC 29424, USA***Abstract**

METASIM provides a flexible environment in which to perform individual-based population genetic simulations. A wide range of landscape-level dynamics, population structures, and within-population demographics can be represented using the framework implemented in this software. In addition, temporal variation in all demographic characteristics can be simulated, both deterministically and stochastically. Such simulations can be used to produce null distributions of genotypes under realistic conditions. These genotypic data can then be used by a variety of analytical programs to develop null expectations of any population genetic statistic estimated from genotypic data.

Keywords: metapopulation, migration model, null distribution

Received 07 December 2001; revision received 1 February 2002; accepted 13 February 2002

Population-genetic markers can provide valuable insights into the demography of natural populations. These insights include estimation of population sizes, migration rates, and mating systems, each of which may be of great interest to ecologists, evolutionary biologists and conservation biologists (Milligan *et al.* 1994). Unfortunately, many population genetic summary statistics make necessary, but restrictive, assumptions with respect to numbers of populations, population structure, mutation model, and effective population size. The effects of violations of these assumptions are not always clear. As a result, the statistical behaviour of population-genetic summary statistics (for example, F_{ST}) are ill-defined in many natural systems (Whitlock & McCauley 1999).

One solution to this problem is to simulate the system of interest and characterize the distribution of summary statistics under various demographic scenarios. Because of continuing advancement in computer technology, characterizing distributions using individual-based simulations has become a viable approach. These simulations represent natural conditions in an explicitly stochastic fashion by treating individuals as autonomous units that move through the environment, experience demographic transitions, and interact with each other based upon computer-generated random numbers. Individual-based models are now used

both in ecological (Greene & Stamps 2001) and evolutionary contexts (Balloux *et al.* 2000); however, there is an absence of models that realistically link ecology and population genetics. Here, I present a model that combines realistic demography, including temporal variation in vital rates, with neutral evolution at multilocus genotypes.

The initial objective for developing METASIM was to produce null expectations of population genetic statistics under realistic evolutionary scenarios. In particular, I wanted to include realistic demography such as overlapping generations as well as metapopulation dynamics. Empirical data would then be compared to these null expectations to test whether these data could be explained by the postulated scenario. For example, the null distribution of a genetic estimator of apparent gene flow could be simulated in a case where no migration is actually occurring. Actual estimates based upon genetics could then be compared to this distribution to test whether gene flow is occurring. Additionally, this software could be used to estimate the power of a particular technique in inferring demographic parameters under realistic conditions, and to provide an environment to compare alternative approaches to infer the same parameter.

The METASIM library and application are implemented in C++. The object-orientation of this language provides a natural representation of individual-based models. Furthermore, C++ extends easily to include new population

Correspondence: Allan E. Strand. E-mail: stranda@cofc.edu

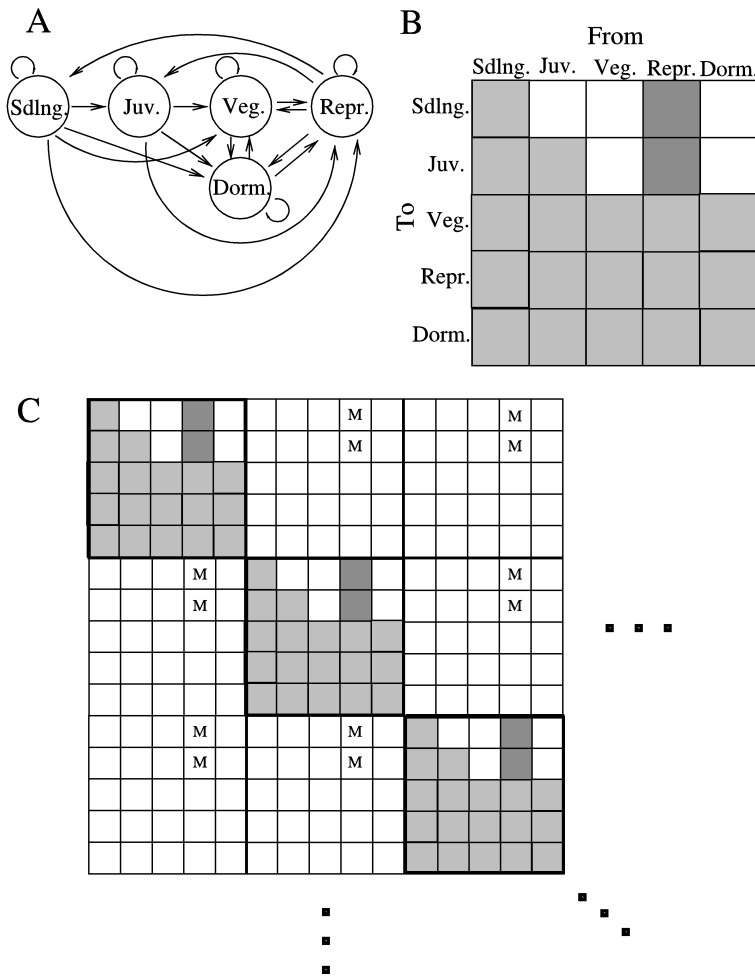


Fig. 1 Example of within-habitat and landscape models for METASIM simulations. Panel A is a life-cycle graph that illustrates the transitions that can occur among different life-stages of *G. pneumonanthe*. Seedlings, juveniles, vegetative adults, reproductive adults and dormant individuals are denoted by 'Sdng.', 'Juv.', 'Veg.', 'Repr.', and 'Dorm.', respectively. Panel B represents these transitions in matrix form. The light grey elements are growth and survivorship and the dark grey elements refer to reproduction. Panel C illustrates a portion of the entire landscape. The matrices on the diagonal refer to within-population demographies indicated in panels A and B. The off-diagonal elements that contain M quantify the amount of migration occurring among populations. In the example simulations, $M = 0$. In this overview, growth and survivorship matrices are combined with reproduction matrices, in the manner typical in matrix population models (Caswell 1989). The male contribution matrices are not included.

or genetic characteristics. For example, a new genetic mutation model could readily be added to the simulation. The METASIM library and applications have been compiled and run on the LINUX, Mac OS X®, FREEBSD, and Microsoft WINDOWS® operating systems. Because the Gnu C++ compiler (<http://www.gnu.org/software/gcc>) on which METASIM was developed has been ported to many operating systems, it should be possible to port METASIM to any of those environments. Currently, I provide source code as well as binaries for LINUX, Mac OS X®, and Microsoft WINDOWS®. The binaries and entire source code are freely available at <http://www.cofc.edu/~stranda/>.

The Metasim library implements an array of habitats in a landscape. Each of the habitats supports a population of individuals and possesses unique demographies, carrying capacities (possibly zero), and extinction probabilities. Individuals move through their life-cycle and through space based upon random numbers chosen from user-specified distributions. The moments of these distributions are entered in a form similar to Lefkovich matrices (Lefkovich 1965), allowing for arbitrarily complex life-histories. The carrying capacity prevents populations

from outgrowing computational resources. Extinction rates provide a means to implement extinction due to environmental stochasticity. Three separate habitat matrices describe population characteristics. The first defines rates of growth and survival, the second defines maternal production of offspring and the third, male contribution to those offspring. Each habitat matrix forms diagonal sub-matrices of three landscape-level matrices. The off-diagonal elements of the landscape matrices specify interactions among populations. This structure is extremely flexible. In theory each life-history stage in each habitat can contribute migrants to any other habitat's life-history categories. Figure 1 provides an example overview of the relationship between within-habitat and landscape processes.

Individuals are the primary discrete object in METASIM. Each individual has a life-history stage (this stage includes the population in which the individual is located), birth date, and multilocus genotype. In each time-period, the new life-history stage of an existing individual is determined by choosing a random number from the multinomial distribution specified by the columns of the growth and survival matrices described above. The appropriate 'from'

column is chosen based upon the current life-history state of the organism.

New individuals are born into the system by sequentially choosing all individuals from each reproductive class. The number of offspring in each demographic class produced by each individual is determined by choosing a number at random from Poisson distributions. Means of these Poisson distributions are entered by the user in habitat reproduction matrices to describe within-population reproduction and in the off-diagonal elements of the landscape reproduction matrices to describe production of offspring that migrate. The source of male gametes are determined by the third set of matrices; elements in these matrices specify the probability that a sperm comes from a particular stage-class.

Although METASIM can be used to simulate metapopulation dynamics alone, the real strength of the program is its ability to include multilocus genotypes for each individual. These genotypes can be a combination of different types of maternally inherited haploid and biparentally inherited diploid loci. All loci are unlinked, although maternal inheritance effectively links haploid loci. Each locus has a unique mutation rate that is held constant across habitats. Three types of genetic loci can be implemented. The first mutates under an infinite allele model (Kimura & Crow 1964). This model may best simulate allozyme data. The second model is a strict ladder or stepwise mutation model (Kimura & Ohta 1978) that may be more appropriate for simulating microsatellite data. The third type of locus is a DNA sequence of user-selectable length. When a mutation occurs in a particular sequence, all bases are equally likely to change; furthermore, there is no substitution bias.

Variation in demographic characteristics among habitats is modeled by randomly selecting alternative habitat matrices from a list provided by the user. The probability that a matrix will be chosen for any particular habitat is also user-selectable.

Temporal variation in a landscape can be achieved in two ways. First, all three landscape matrices (survival, reproduction, and male function) can change over time in a deterministic fashion. This provides a mechanism to examine nonequilibrium dynamics in a population system. For example, a system of populations could exchange migrants at high rates for the initial time-periods of a simulation followed by a series of time periods where no migrants were exchanged. This would simulate a single population fragmenting into multiple, isolated populations. An additional approach to temporal variation chooses sets of demographic matrices from a user-defined list. As in the spatial variation, the probability that a certain set of matrices is chosen each time period is user selectable. This feature can be used to simulate stochastic variation in any demographic characteristic.

METASIM does not perform analyses; it is intended as a simulation engine only. To facilitate analysis of simulation

results, a bundled application, METATRANS, is provided. METATRANS reads files in the METASIM format, subsamples individuals, and converts these files into a number of formats used by popular analytical programs (GDA, ARLEQUIN 2.0, and BIOSYS). In addition, a format suitable for input to a set of R language (Ihaka & Gentleman 1996) routines included in the METASIM package is provided. These routines implement a nested ANOVA approach to estimating F_{ST} (Weir & Cockerham 1984). In the following example, I used two stage-transition models for *Gentiana pneumonanthe* (Oostermeijer *et al.* 1996). These two matrices represent average matrices estimated in haymeadows and heathlands (Oostermeijer *et al.* 1996; Table 3). The simulation consisted of a single population that fragmented into six isolated populations after 1×10^4 years of random mating. I ran the simulation for 2×10^4 years post-fragmentation. Twenty iterations of the entire 3×10^4 years history were performed for the simulation. Fragmentation of populations in this manner occur as the result of climate changes such as those following the Pleistocene (Strand *et al.* 1996).

Each individual possessed a six locus genotype. Loci one and two were 100 nucleotide (nt) haploid sequences. Loci three and four were diploid and evolving under a strict stepwise model, and loci five and six were diploid 100 nt sequences. Loci one, three, and five had allele-wide mutation rates of 5×10^{-4} . The remaining loci had mutation rates of 5×10^{-6} . Initially, a single allele was present at each locus. I fixed the carrying capacity of the single population at 6×10^4 for the initial 1×10^4 years. During the remainder of the simulation, individual populations had carrying capacities set at 1×10^4 . To represent the demography within each habitat, every year either the haymeadows or heathlands matrices were chosen with probabilities 0.01 and 0.99, respectively. These probabilities were chosen so that it would be extremely unlikely for a population to go extinct during these exemplar simulations; the intrinsic rate of increase in the haymeadow was 0.92 and the rate of increase in the heathlands was 1.16.

I calculated θ (Weir & Cockerham 1984), for each landscape every 500 years. For each locus, I based these calculations on the frequency of haplotypes or alleles and I ignored the potential information contained in the allele states. For each calculation, I chose 30 individuals from each of the six populations. Therefore, temporal variation observed in θ results from the combined effects of evolutionary processes and simulated population sampling.

Trajectories of θ as a function of time are presented in Fig. 2. A large portion of the variation occurred among populations for loci with high mutation rates. Little variation in this result was observed among simulation runs. Also, the estimates of θ reached equilibrium approximately 10 000 years following isolation. Conversely, a large amount of stochasticity was observed in estimates of θ when mutation rates were at lower levels. For loci two and

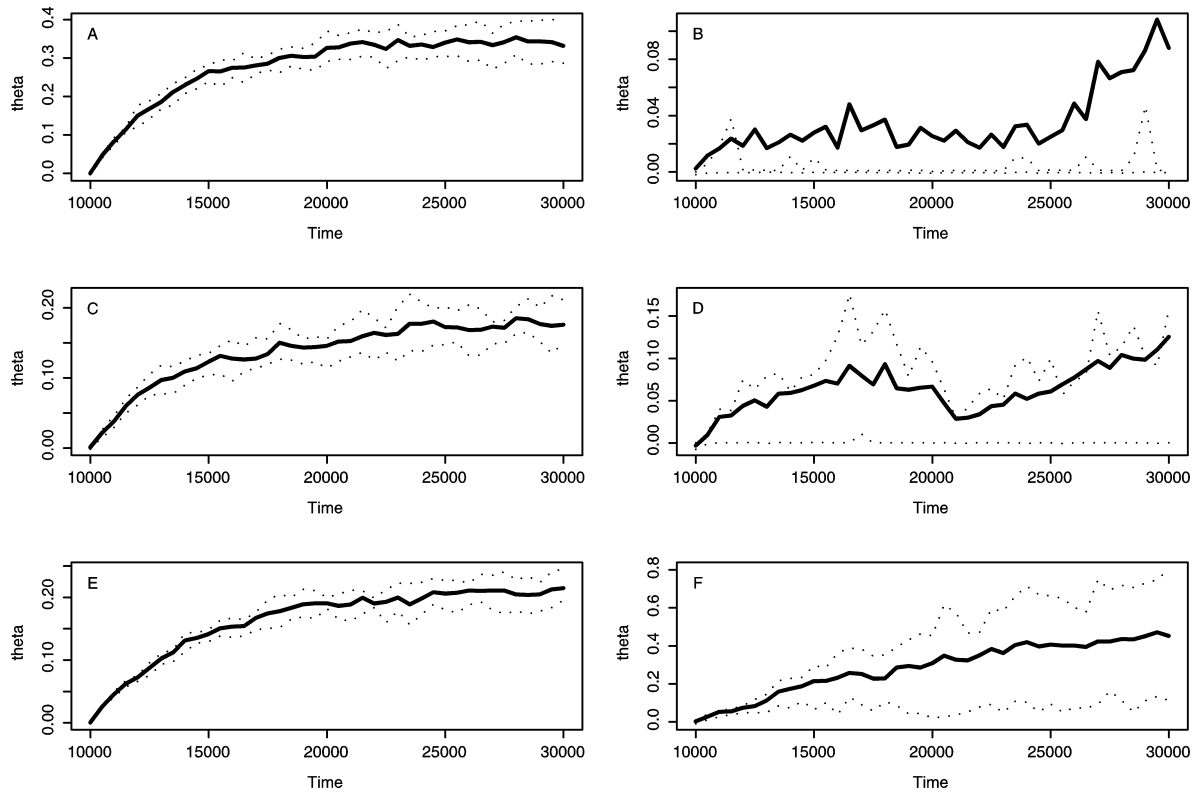


Fig. 2 Trajectory of θ (Weir & Cockerham 1984) over 20 000 years following complete isolation. Solid lines refer to means and lower and upper dotted lines refer to 25th and 75th percentiles, respectively. Means and percentiles were calculated over 20 iterations. Panels A-F correspond to loci 1–6 described in text, respectively.

four θ never rose significantly above zero, even after 20 000 years. Locus 6 did increase over time, although in 25% of the simulations θ remained near zero. Furthermore, it appears that even after 20 000 years, estimates of θ have not reached equilibrium in any of the loci with low mutation rates.

The results in Fig. 2 demonstrate that under realistic conditions, interactions among demography, inheritance, mutation model, and mutation rate can strongly affect the behaviour of population genetic summary statistics. Because of the complexity of these interactions, it is likely that simulations will provide the only avenue with which to investigate them. METASIM implements a flexible environment in which to perform these simulations.

Acknowledgements

This work was supported in part by a grant from the United States Fish and Wildlife Service. In addition, I would like to thank M. Cain, P. Rosel, and M. Zatzoff for helpful comments.

References

- Balloux F, Brunner H, Lugon-Moulin N, Hausser J, Goudet J (2000) Microsatellites can be misleading: An empirical and simulation study. *Evolution*, **54**, 1414–1422.
- Caswell H (1989) *Matrix Population Models. Construction, Analysis and Interpretation*. Sinauer, Sunderland, Massachusetts.
- Greene C, Stamps JA (2001) Habitat selection at low population densities. *Ecology*, **82**, 2091–2100.
- Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.
- Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. *Genetics*, **49**, 725–738.
- Kimura M, Ohta T (1978) Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proceedings of the National Academy of Sciences of the USA*, **75**, 2868–2872.
- Lefkovich LP (1965) The study of population growth in organisms grouped by stages. *Biometrics*, **21**, 1–18.
- Milligan BG, Leebens-Mack J, Strand AE (1994) Conservation genetics: beyond the maintenance of marker diversity. *Molecular Ecology*, **3**, 423–435.
- Oostermeijer J, Brugman M, De Boer E, Den Nijs H (1996) Temporal and spatial variation in the demography of *Gentiana pneumonanthe*, a rare perennial herb. *Journal of Ecology*, **84**, 153–166.
- Strand AE, Milligan BG, Pruitt CM (1996) Are populations islands? Analysis of Chloroplast DNA Variation in *Aquilegia*. *Evolution*, **50**, 1822–1829.
- Weir BS, Cockerham CC (1984) Estimating F -statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.
- Whitlock M, McCauley D (1999) Indirect measures of gene flow and migration: $F_{ST} \neq 1/(Nm + 1)$. *Heredity*, **82**, 117–125.