SelSim: a program to simulate population genetic data with natural selection and recombination

Chris C. A. Spencer∗ and Graham Coop

Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, UK

Received on April 22, 2004; revised on June 22, 2004; accepted on July 6, 2004
Advance Access publication July 22, 2004

ABSTRACT

Summary: SelSim is a program for Monte Carlo simulation of DNA polymorphism data for a recombining region within which a single bi-allelic site has experienced natural selection. SelSim allows simulation from either a fully stochastic model of, or deterministic approximations to, natural selection within a coalescent framework. A number of different mutation models are available for simulating surrounding neutral variation. The package enables a detailed exploration of the effects of different models and strengths of selection on patterns of diversity. This provides a tool for the statistical analysis of both empirical data and methods designed to detect natural selection.

Availability: http://www.stats.ox.ac.uk/mathgen/software.html
Contact: spencer@stats.ox.ac.uk
Supplementary information: http://www.stats.ox.ac.uk/mathgen/software.html

INTRODUCTION

For many population genetic models, analytical results are unavailable for quantities of interest (e.g. distributions of test statistics); therefore, Monte Carlo methods are often used (e.g. Hudson, 2002; Posada and Wiuf, 2003). The SelSim package is a tool for the simulation of population genetic data within which a single, bi-allelic locus has experienced natural selection on a background of an otherwise neutral, recombining DNA sequence for a constant effective population size, $N_e$. SelSim utilizes the coalescent (Kingman, 1982), which is a stochastic model for describing the underlying ancestral relationship between members of a random sample of DNA sequences drawn from a randomly mating population.

Incorporating natural selection into the coalescent is problematic because the relationship between ancestors and descendants depends on the allelic type at the selected site (Kaplan et al., 1988; Neuhauser and Krone, 1997). However, by first simulating the population frequency of the selected allele through time (termed the trajectory) a genealogical history may be generated conditional upon this trajectory (Fig. 1). Most past approaches have either simulated the trajectory forward in time or used deterministic approximations. Both approaches have their drawbacks, particularly when selection is weak; forward simulation can be inefficient because it is not possible to condition upon the present day population frequency, while deterministic approximations can distort the effect of selection (e.g. see Barton and Etheridge, 2004). In contrast, SelSim uses a backwards in time stochastic model, by exploiting reversibility arguments, to simulate trajectories back from the current day population frequency (Griffiths, 2003). Slatkin (2001) uses a similar method in order to simulate natural selection in a variable population size for a non-recombining region. However in the model implemented here the population size is constant, and the trajectory is correctly time-reversed meaning that there is no need for weighting or rejection of trajectories.

SelSim is the first package to incorporate both fully stochastic models and deterministic approximations to natural selection, allowing simulation at all strengths of selection for a wide range of fitness models. Possible applications include investigating the consequences of assuming selective neutrality when natural selection has occurred; testing the statistical power of methods designed to detect or estimate the strength of natural selection; or fitting non-neutral models to empirical data.

THE MODEL

The statistical properties of the trajectory of an allele backwards in time from a current frequency $x$ back to the time of its introduction into the population, are identical to those of an allele forward in time starting again from $x$ and proceeding towards loss (Watterson, 1977). Thus, it is possible to model the trajectory backward in time by simulating from a Moran Model forward in time conditional on the eventual loss of the allele from the population (Coop and Griffiths, 2004; Griffiths, 2003). SelSim’s stochastic trajectories permit the use of a general diploid model for the fitness relationship between alleles; the homozygote for the derived allele

∗To whom correspondence should be addressed.
Fig. 1. A caricature of the coalescent with natural selection. The trajectory of the selected allele through time, \( x(t) \), is shown as a grey line. A coalescent history (without recombination) with \( n \) sample members with the derived allele and \( m \) of the ancestral allele is superimposed on this trajectory. The current day, \( t = 0 \), is marked by a dashed line. The time at which the selected mutation arose in the population is marked by a dot on the genealogy.

has relative fitness \((1 + s)\), the heterozygote \((1 + sh)\) and the homozygote for the ancestral allele has relative fitness 1. It is possible to simulate positive directional and balancing (complete overdominant) selection in both the stochastic and deterministic case.

Given the trajectory, the ancestral process is similar to a two-deme structured coalescent model (e.g. see Nordborg, 2001) where the population size at time \( t \) is \( x(t) \) for lineages linked to the derived allele and \( 1 - x(t) \) for lineages linked to the ancestral allele (Kaplan et al., 1988). Recombination allows lineages to switch allelic types and is analogous to moving between the two demes backward in time (Hudson and Kaplan, 1988). A computationally useful feature of the coalescent is that, for neutral mutations, the ancestral process can be separated from the mutation process so that neutral mutations can be placed on the ancestry specifying the pattern of mutations in the sample of DNA sequences.

FUNCTIONALITY

The selected site may be chosen to be polymorphic or fixed in the population at the time of sampling. If the derived allele is fixed, then the user may choose the time in the past when fixation occurred. If the derived allele is segregating, a current day population frequency \( x(0) \) and sample configuration (the number of derived alleles, \( n \), and ancestral alleles, \( m \) ) may be set by the user or generated automatically by one of the following options:

1. Specify \( n \) and \( m \) and draw \( x(0) \) from the density of the population frequency given the selection coefficient and the sample configuration (Griffiths, 2003).
2. Specify \( x(0) \) and binomially draw \( n \) and \( m \) conditionally.
3. Draw \( x(0) \) from the density of the population frequency conditional on the selection coefficient and on the allele segregating in the population. Then draw \( n \) and \( m \) as in (2).

SelSim can simulate either a homogeneous recombination rate across the region or use an arbitrary recombination map. Four different mutation schemes are available in two separate executables:

**SelSim (Single nucleotide polymorphisms)**

1. Infinitely-many-sites: The user can specify a population scaled mutation rate \( \theta = 4N_e \mu \) for the sequence, where \( \mu \) is the per generation rate of mutation. Otherwise, the user can either fix the number or location of SNPs.
2. Finite sites: An arbitrary nucleotide mutation rate matrix can be specified.

**SelSimM (Microsatellites)**

3. A flexible single-step microsatellite model is implemented allowing length-dependent mutation rate and directional bias (Sainudiin et al., 2004).
4. Infinitely-many-alleles: Under this model a new allele type is generated by each mutation at each locus.

For mutation models 2–4 SelSim allows different parameters relating to mutation to be set at different loci. For each simulation, the location of polymorphic sites and the DNA sequence haplotypes are output along with the time in the genealogy underlying each polymorphic site. It is also possible to output the location of recombination events and statistics of the genealogies underlying each locus. For 25 derived and 25 ancestral sequences of 1000 sites with a population scaled recombination rate of 10 and a population scaled selection coefficient of 200, the deterministic model takes 51 s to perform 1000 runs while the stochastic model (with a population size of 1000) takes 54 s on a 2.4 GHz machine.

SelSim takes as input a control file which specifies all available options. An on-line version of SelSim is available, in reduced form, allowing graphical representation of a single realization from the model. Both command line and on-line versions are written in C++. The command line version is available precompiled under Windows XP, Linux and Macintosh, with source code available from the authors on request.
REFERENCES


