Mlcoalsim: Multilocus Coalescent Simulations

Sebastian E. Ramos-Onsins [1,2] and Thomas Mitchell-Olds [1,3]

[1]Max-Planck Institute for Chemical Ecology, Hans-Knöll Str. 8, D-07745 Jena, Germany.
[2]Present address: Departament de Genètica, Universitat de Barcelona, Diagonal 645, Barcelona, Spain.
[3]Present address: Department of Biology, Duke University, Durham, NC 27708, USA.

Abstract: Coalescent theory is a powerful tool for population geneticists as well as molecular biologists interested in understanding the patterns and levels of DNA variation. Using coalescent Monte Carlo simulations it is possible to obtain the empirical distributions for a number of statistics across a wide range of evolutionary models; these distributions can be used to test evolutionary hypotheses using experimental data. The *mlcoalsim* application presented here (based on a version of the *ms* program, Hudson, 2002) adds important new features to improve methodology (uncertainty and conditional methods for mutation and recombination), models (including strong positive selection, finite sites and heterogeneity in mutation and recombination rates) and analyses (calculating a number of statistics used in population genetics and *P*-values for observed data). One of the most important features of *mlcoalsim* is the analysis of multilocus data in linked and independent regions. In summary, *mlcoalsim* is an integrated software application aimed at researchers interested in molecular evolution. *mlcoalsim* is written in ANSI C and is available at: http://www.ub.es/softevol/mlcoalsim.

Keywords: Neutrality tests, Rejection algorithm, Population Genetics, Multilocus analyses, Coalescent simulations

Introduction

Statistical inference of molecular population data under different evolutionary models typically employs a coalescent framework (Kingman, 1982a,b; Hudson, 1990; Donnelly and Tavaré, 1995; Nordborg, 2001). Hudson's *ms* (Hudson, 2002) application enabled a large number of population geneticists and molecular biologists to examine data under different evolutionary models. In recent years, a number of coalescent programs focused on the generation of genetic data have been published (e.g. *SimCoal*, Excoffier et al. 2000; Laval and Excoffier, 2004; *SelSim*, Spencer and Coop, 2004; *CoaSim*, Mailund et al 2005; *FastCoal*, Marjoram and Wall, 2005). Nevertheless, multilocus data obtained by high throughput techniques (e.g. the Drosophila Polymorphisms Sequencing Project, as well as smaller projects such as those described by Akey et al. 2004; Schmid et al. 2005) are not easily analyzed using available software. Here we describe the *mlcoalsim* software application which, unlike other available tools, allows the generation of simulated genetic data and the calculation of descriptive statistics for a large number of loci under different evolutionary models, as well as obtaining *P*-values of observed data.

Program Overview

mlcoalsim enables researchers to compare single and multilocus data with several common evolutionary models. It is an integrated application that not only constructs coalescent trees and sequences but also calculates a number of summary statistics that are useful for the examination of evolutionary hypotheses. This program is designed to generate within-species genetic data; that is, the level of nucleotide variation should not be too high—a maximum of approximately 5%—in order to avoid important errors (a more sophisticated substitution model should be used). For the same reason, the level of divergence from an outgroup species should be no greater than 10–15%.

Multilocus analyses

One of the main features of *mlcoalsim* is the generation of DNA samples and calculation of a number of statistical tests for a set of multiple loci with variable levels of intragenic recombination. There are

Correspondence: Sebastian E. Ramos-Onsins, Departament de Genètica, Universitat de Barcelona, Diagonal 645, Barcelona 08028, Spain. Tel: 34 934035304; Email: sramosonsins@ub.edu

two options for multilocus analysis: using independent (unlinked) loci and using a single long region separated into several fragments. The first option (independent loci) allows the independent analysis of each locus and the calculation of summary statistics (the average and variance for all loci of each statistic). This option is useful for contrasting data with demographic models that would affect the entire genome. For such an analysis, a correction factor for population size depending on the chromosomal location of each locus (e.g. autosomal, sexual) is needed. The second option (linked loci) generates samples for an entire linked region and calculates statistics for specified fragments within this region or for a sliding window analysis. The "linked" option is useful in evolutionary processes that affect only specific regions, such as a selective sweep in a recombining region.

Uncertainty in mutation and recombination rates

Mutation and recombination rates are critical parameters which are usually unknown. In order to consider the uncertainty of these two parameters, *mlcoalsim* can sample the rates from a distribution (uniform and gamma distributions are used) instead of using a fixed value. In addition, *mlcoalsim* can generate samples by fixing the observed values, the number of segregating sites, the minimum number of recombination events and (optionally) the number of haplotypes. This last option is obtained using the rejection method 2 of Tavaré (Tavaré et al. 1997). Posterior distributions for the population mutation and for the population recombination parameter are recorded.

Heterogeneity in mutation and recombination rate across the sequence

mlcoalsim is also able to take into account differences in the mutation and in recombination rates across the studied region. Heterogeneity is modelled with a gamma distribution, modelling from extreme hotspots regions (i.e. in case using heterogeneity for the recombination rate, only few position are enabled to recombine while others can not) to uniform values for all positions. Furthermore, it is possible to fix the average number of invariant positions (position that can not mutate) for the studied region.

Evolutionary models

mlcoalsim includes the following evolutionary models: the neutral stationary panmictic model, the finite island model, models with changing population sizes over time, refugia models and deterministic positive selection (not all of these models can be used simultaneously). *mlcoalsim* allows the use of neutral and positive selection models for different independent loci, and changing population size also can be used with a finite island model.

Statistics

A number of statistics and related tests used in population genetics are displayed in the output (Table 1). The statistics incorporated in this program describe the level and patterns of diversity for a given sample.

Different statistics that estimate the level of variation are included (θ, Watterson, 1975, π, Tajima, 1983, and θ_H, Fay and Wu, 2000) for the entire sample. Although these estimates are calculated using different approaches, the values should be equal under the assumption of a neutral stationary panmictic model. The average levels of variation within and among populations are also estimated (π_w and π_b, Hudson et al. 1992), as well as the average differentiation among populations with the *Fst* statistic (e.g. Hudson et al. 1992).

A description of the patterns of diversity is obtained using two main classes of statistics (Ramos-Onsins and Rozas, 2002): Class I statistics, which use the mutation frequency information, and Class II statistics, which use information from the haplotype distribution. Class I includes Tajima's *D* test (TD, Tajima, 1989), Fu and Li's tests (FD*, FF*, FD, FF, Fu and Li, 1993), Fay and Wu's *H* test (Fay and Wu, 2000), *R2* (Ramos-Onsins and Rozas, 2002) and weighted statistics for a multilocus approach such as *D/Dmin* (Schaeffer, 2002) and *H/Hmin* (Schmid et al. 2005). Class II includes the number of haplotypes *Kw* (Strobeck, 1987) and the haplotype diversity *Hw* (Depaulis and Veuille, 1998), both weighted by the number of samples for a better multilocus comparison, *Fs* (Fu, 1997), the statistics *B* and *Q*, (Wall, 1999) which count differences in haplotype structure at adjacent positions, the *ZA* statistic (Rozas et al. 2001) as a measure of linkage disequilibrium at adjacent positions, *maxhap* (Depaulis et al. 2003) and *maxhap1*

Table 1. List of the main statistics included in mlcoalsim.

Name	Statistic	Citation
TD	Tajima's *D* test	Tajima, 1989
Fs	Fu's *Fs* test	Fu, 1997
FD*	Fu and Li's *D** test	Fu and Li, 1993
FF*	Fu and Li's *F** test	Fu and Li, 1993
FD	Fu and Li's *D* test	Fu and Li, 1993
FF	Fu and Li's *F* test	Fu and Li, 1993
H	Fay and Wu's *H* test	Fay and Wu, 2000
B	Wall's *B* test	Wall, 1999
Q	Wall's *Q* test	Wall, 1999
ZA	*ZA*	Rozas et al. 2001
Fst	F_{st}	Hudson et al. 1992
Kw	No. haplotypes/*n*	Strobeck, 1987
Hw	Haplotype diversity/*n*	Depaulis and Veuille, 1998
R2	*R2* test	Ramos-Onsins and Rozas, 2002
S	No. of biallelic mutations	
thetaWatt	θ	Watterson, 1975
thetaTaj	π	Tajima, 1983
thetaFW	θ_H	Fay and Wu, 2000
pi_w	π within populations	e.g. Hudson et al. 1992
pi_b	π among populations	e.g. Hudson et al. 1992
D/Dmin	*D/Dmin*	Schaeffer, 2002; Schmid et al. 2005
H/Hmin	H/Hmin	Schmid et al. 2005
maxhap	No. lines in most common haplotype/*n*	Depaulis et al. 2003
maxhap1	maxhap excepting one biallelic mutation	Hudson et al 1994; Rozas et al. 2001
Rm	*Rm*	Hudson and Kaplan, 1985

n is the number of sequence lines.
See text and *mlcoalsim* documentation for a brief description of statistics.

(simplified from Hudson et al. 1994), which counts the number of lines with the most common haplotype (i.e. *maxhap*) but allowing a single segregating site within the largest "haplotype" group (Rozas et al. 2001). Finally, the minimum number of recombination events, *Rm* (Hudson and Kaplan, 1985), is also calculated.

Multilocus analyses generate a comprehensive output with the calculated statistics with their average and variance. Only biallelic positions are considered for the analyses given that tri- or tetra-allelic positions are rare in within-species samples.

Other technical features

The generation of random deviates from uniform, binomial, Poisson, and gamma distributions and the determining of roots for complex functions are based on Lanczos (1964); Atkinson (1979); Cheng and Feast (1979); Fishman (1979); Ridders (1979); Press et al (1992) and Press and Teukolsky (1992). The Rm function was obtained and modified from Wall's code (Wall, 2000). The gamma function was partially obtained from Grassly, Adachi and Rambaut code (Grassly et al. 1997).

Acknowledgements

We would like to thank everyone who contributed to improving and debugging this application program, particularly those working in the labs of M. Aguadé, W. Stephan and T. Mitchell-Olds. Thanks to J. Rozas for his help and to Y. Kim for helping with the selective model. This work is partially supported by "Distinció per la Promoció de la Recerca Universitària" awarded by the Autonomous Government of Catalonia, grant BFU200402253 from the Spanish Ministry of Education and Science awarded to M. Aguadé and by the Max-Planck Society, Germany.

References

Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.J., Shriver, M.D., Nickerson, D.A. and Krugyak, L. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol*., 2(10), e286.

Atkinson, A.C. 1979. The computer generation of poisson random variables. *Appl. Statist*., 28:1, 29–35.

Cheng, R.C. and Feast, G.M. 1979. Some simple gamma variate generators. *Appl. Statist*., 28:3, 290–295.

Depaulis, F., Mousset, S. and Veuille. M. 2003. Power of neutrality tests to detect bottlenecks and hitchhiking. *J. Mol. Evol*., 57 Suppl 1, S190–200.

Depaulis, F. and Veuille, M. 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Mol. Biol. Evol.,* 15(12):1788–1790.

Donnelly, P. and Tavaré, S. 1995. Coalescent and genealogical structure under neutrality. *Ann. Rev. Genet*., 29:401–421.

Excoffier, L., Novembre, J. and Schneider, S. 2000. SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *J. Hered*., 91(6): 506–509.

Fay, J.C. and Wu., C.I. 2000. Hitchhiking under positive Darwinian selection. *Genetics,* 155(3):1405–1413.

Fishman, G.S. 1979. Sampling from the binomial distribution on a computer. *J. Am. Statist. Ass*., 74:366, 418–423.

Fu, Y.X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics, 147(2): 915–925.

Fu, Y.X. and Li W.H. 1993. Statistical tests of neutrality of mutations. Genetics, 133(3):693–709.

Grassly, N.C., Adachi, J. and Rambaut A. 1997. PSeq-Gen: an application for the Monte Carlo simulation of protein sequence evolution along phylogenetic trees. *Comput Appl Biosci*., 13 (5):559–560.

Hudson, R.R. 1990. Gene genealogies and the coalescent process. In D. Futuyama and J. Antonovics (Eds.), *Oxford Surveys in Evolutionnary Biology*, Volume 7, pp. 1–45. Oxford: Oxford University Press.

Hudson, R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics,* 18(2):337–338.

Hudson, R.R., Bailey, K., Skarecky, D., Kwiatowsky, J. and Ayala, F. 1994. Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. *Genetics,* 136:1329–1340.

Hudson, R.R. and Kaplan, N.L. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics,* 111(1):147–164.

Hudson, R.R., Slatkin, M. and Maddison W.P. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics,* 132(2):583–589.

Kingman, J.F.C. 1982. The coalescent. Stochast. *Proc. Appl*., 13, 235–248.

Kingman, J.F.C. 1982. On the genealogy of large populations. *J. Appl. Prob*., 19A, 27–43.

Lanczos, C. 1964. A precision approximation of the gamma function. *J. SIAM,* 1:86–96.

Laval, G. and 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(15):2485–2487.

Mailund, T., Schierup, M., Pedersen, C.N.S., Mechlenborg, P.J.M., Madsen, J.N. and Scauser, L. 2005. Coasim: A flexible environment for simulating genetic data under coalescent models. BMC *Bioinformatics,* 6:252.

Marjoram, P. and Wall, J.D. 2006. Fast "coalescent" simulation. BMC *Genetics,* 7:16.

Nordborg M. 2001. Coalescent theory. In D. Balding, M. Bishop, and C. Cannings (Eds.), *Handbook of Statistical Genetics*, pp. 179–212. Chichester: John Wiley and Chichester Sons.

Press, W.H. and Teukolsky, S.A. 1992. Portable random number generators: 6(5):522. Computers in Physics 6:522–524.

Press, W.H., Teukolsky, S.A. Vetterling, W.T. and Flannery, B.P. 1992. Numerical Recipes in C. The Art of Scientific Computing. Cambridge University Press.

Ramos-Onsins, S.E. and Rozas, J. 2002. Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol.,* 19(12):2092–2100.

Ridders, C.J.F. 1979. A new algorithm for computing a single root of a real continuous function. IEEE Transactions on Circuits and Systems 26:11, 979–980.

Rozas, J., Gullaud, M., Blandin, G. and Aguade, M. 2001. DNA variation at the rp49 gene region of Drosophila simulans: evolutionary inferences from an unusual haplotype structure. *Genetics,* 158(3):1147–1155.

Schaeffer, S. 2002. Molecular population genetics of sequence length diversity in the Adh region of Drosophila pseudoobscura. *Genet. Res.,* 80:163–175.

Schmid, K.J., Ramos-Onsins, S.E., Ringys-Beckstein, H., Weisshar, B. and Mitchell-Olds T. 2005. A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics,* 169:1601–1615.

Spencer, C.C.A. and Coop, G. 2004. *SelSim*: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics,* 20(18):3673–3675.

Strobeck, C. 1987. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics,* 117:149–153.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics,* 105(2):437–460.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics,* 123(3):585–595.

Tavaré, S., Balding, D.J., Griffiths, R.C. and Donnelly, P. 1997. Inferring coalescence times from DNA sequence data. *Genetics,* 145(2): 505–518.

Wall, J.D. 1999. Recombination and the power of statistical tests of neutrality. *Genet. Res*., 74:65–79.

Wall, J.D. 2000. A comparison of estimators of the population recombination rate. *Mol. Biol. Evol*., 17:156–163.

Watterson, G. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.,* 7:256–276.