Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw355
Advance Access Publication Date: 16 June 2016
Applications Note

OXFORD

Genetics and population analysis

ARGON: fast, whole-genome simulation of the discrete time Wright-fisher process

Pier Francesco Palamara[1,2,*]

[1]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA and [2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on January 15, 2016; revised on May 6, 2016; accepted on May 31, 2016

Abstract

Motivation: Simulation under the coalescent model is ubiquitous in the analysis of genetic data. The rapid growth of real data sets from multiple human populations led to increasing interest in simulating very large sample sizes at whole-chromosome scales. When the sample size is large, the coalescent model becomes an increasingly inaccurate approximation of the discrete time Wright-Fisher model (DTWF). Analytical and computational treatment of the DTWF, however, is generally harder.
Results: We present a simulator (ARGON) for the DTWF process that scales up to hundreds of thousands of samples and whole-chromosome lengths, with a time/memory performance comparable or superior to currently available methods for coalescent simulation. The simulator supports arbitrary demographic history, migration, Newick tree output, variable mutation/recombination rates and gene conversion, and efficiently outputs pairwise identical-by-descent sharing data.
Availability: ARGON (version 0.1) is written in Java, open source, and freely available at https://github.com/pierpal/ARGON.
Contact: ppalama@hsph.harvard.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The coalescent (Kingman, 1982) can be constructed as an approximation of the discrete time Wright-Fisher process (DTWF, Fisher *et al.*, 1922; Wright, 1931), and leads to simplified analytical and computational treatment. Simulators based on the coalescent process (e.g. Hudson, 2002) have been extensively adopted in computational methods. The coalescent approximation, however, relies on the assumption that the sample size is small compared with the effective population size, and violations of this assumption may result in substantial distortions of key genealogical properties (Bhaskar *et al.*, 2014; Wakeley and Takahashi, 2003).

Until very recently, coalescent simulators did not scale up to long chromosomes and the very large sample sizes of modern day data sets, which now comprise hundreds of thousands of individuals. The recently developed simulators COSI2 (Shlyakhter *et al.*, 2014) and SCRM (Staab *et al.*, 2015) enable fast simulation under approximate coalescent models. The COSI algorithm uses a standard backwards-in-time approach, while SCRM adopts a 'sequential' approach (McVean and Cardin, 2005; Wiuf and Hein, 2000). Both can be also used to simulate large sample sizes and chromosome-long regions under the 'exact' coalescent process with reasonable time and memory requirements.

Here, we present ARGON, an efficient simulator of the DTWF process that scales up to very large chromosomes, and hundreds of thoudsands of samples. The simulator offers substantially improved performance compared with recent DTWF simulators, e.g. GENOME (Liang *et al.*, 2007), and is comparable or superior to current coalescent simulators in terms of speed and memory usage.

2 Approach

ARGON proceeds backwards in time one generation at a time, occasionally sampling coalescent and recombination events subject to

population structure and migration. Each individual is represented as a list of regions for which not all samples have found a common ancestor. Crossover and non-crossover recombination events are sampled in genetic space and rounded to the closest physiscal base pair position based on the desired uniform or variable recombination rate. Whenever a coalescence occurs, regions within individuals are annotated with links to descendant nodes in the ancestral recombination graph (ARG).

When compared with other DTWF implementations (e.g. GENOME), ARGON offers substantially improved speed and memory usage (see Supplementary Materials). In ARGON, large regions are represented as intervals with arbitrary boundary values, and hash map data structures are extensively used to take advantage of sparsity, avoiding explicit representation of all individuals. ARGON can run in approximate mode, so that recombinations breakpoints are rounded to blocks of a user-specified genetic length, as implemented in the GENOME simulator. This reduces the granularity of the recombination process, improving speed and memory usage, at the cost of slightly inflated correlation of nearby markers.

ARGON can efficiently output a list of identical-by-descent (IBD) segments, which are delimited by the occurrence of recombination events that change the most recent common ancestor for pairs of samples. In ARGON, these events are detected by visiting internal ARG nodes, while previous approaches to output simulated IBD sharing data (e.g. Palamara *et al.*, 2012) required comparing recent ancestry for all pairs of individual at each marginal tree.

3 Results

3.1 Accuracy for small sample sizes

When $n \ll N_e$, the coalescent is a good approximation of the DTWF. We performed extensive testing for several scenarios including population size variation, migration across multiple demes, and gene conversion. We report detailed results in the Supplementary Materials. We find good agreement between ARGON and MS. We also tested the accuracy of COSI version 2.0, SCRM version 1.6.1, and MSprime version 0.1.6 (Kelleher *et al.*, 2015), a new efficient simulator for which a preliminary version was released at the time of writing. All simulators were found to be well calibrated against MS.

3.2 Deviation of the coalescent from the DTWF

The coalescent becomes a poor approximation of the DTWF process when the sample size is not substantially smaller than the effective population size (Bhaskar *et al.*, 2014; Wakeley and Takahashi, 2003). We verified that ARGON matches the theoretical prediction for the number of singletons and doubletons described in Bhaskar *et al.* (2014) for the DTWF (see Table 1). We simulated populations of effective size $N_e = 1000$ and $N_e = 20\,000$ haploid individuals, and sampled all present-day individuals. While ARGON matches the prediction of Bhaskar *et al.* (2014) in both exact and approximate mode, coalescent simulations substantially deviate from the DTWF model.

3.3 Scalability to large sample size and whole-chromosome length

We tested the run time and memory usage of ARGON and two recently developed programs that enable simulating very large sample sizes and long chromosomes: COSI2 and SCRM (see Table 2). We find that for large parameter values SCRM generally performs worse than ARGON. COSI2 is generally faster, but scales poorly for memory usage as the size of the region and the sample size grows. Additional tests, including a pre-release version of MSprime, are detailed in the Supplementary Materials.

We further tested the performance of approximate algorithms for the same set of simulation parameters (see Table 2). We compared ARGON with a minimum recombination block size of 10 μm (AR_{10}), SCRM with the '−l' flag set to 0 (SC_0), and COSI2 with the '−u' flag

Table 1. Number of singleton and doubleton alleles when $n = N_e$. We simulated a region of 1 Mb, $\mu = 2 \times 10^{-8}$ and 1 cM/Mb, where θ_s is the average simulation result and θ_t is the theoretical expectation. The ± sign introduces a standard error

	$n = N_e$	singletons %	doubletons %
MS	20 000	**−10.81±0.01**	**+4.87±0.02**
ARGON	20 000	+0.01±0.01	−0.02±0.02
$ARGON_{10}$	20 000	+0.00±0.01	−0.02±0.02
MS	1000	**−10.80±0.05**	**+4.80±0.09**
ARGON	1000	+0.03±0.05	−0.11±0.09
$ARGON_{10}$	1000	+0.02±0.05	+0.17±0.09

Errors are obtained as $100 \times (\theta_s - \theta_t)/\theta_s$, where θ_s is the average simulation result and θ_t is the theoretical expectation. The ± sign introduces a SE.

Table 2. Comparison of simulation algorithms with respect to chromosome length (Mb), exponential expansion rate (ρ), sample size (n), and ancestral population size (A)

				Time			Memory		
Mb	$10^2\rho$	$n/10^3$	A	AR	SC	CS	AR	SC	CS
100	0.0	20	20	0.48	1.42	0.12	10.1	19.6	17.3
300	0.0	20	20	2.56	†	0.59	15.2	†	41.4
500	0.0	20	20	5.35	†	†	19.8	†	†
100	0.0	1	10	0.09	0.02	0.01	6.5	0.5	0.2
100	0.0	10	10	0.12	0.27	0.02	8.5	4.9	4.4
100	0.0	100	10	1.01	6.20	52.33	12.6	48.9	6.8
100	0.0	200	10	3.05	†	†	19.4	†	†
100	0.0	300	10	10.19	†	†	24.6	†	†
100	0.347	20	10	0.39	1.37	0.07	10.5	19.6	9.0
100	0.576	20	10	0.94	2.37	†	13.0	39.1	†
100	1.151	20	10	3.97	†	†	25.0	†	†
				Time			Memory		
Mb	$10^2\rho$	$n/10^3$	A	AR_{10}	SC_0	CS_0	AR_{10}	SC_0	CS_0
100	0.0	20	20	0.12	1.23	0.15	8.2	19.6	17.3
300	0.0	20	20	0.46	†	0.89	9.7	†	41.4
500	0.0	20	20	0.79	†	†	11.0	†	†
100	0.0	1	10	0.02	0.02	0.01	3.3	0.5	0.3
100	0.0	10	10	0.04	0.28	0.03	7.9	4.9	4.4
100	0.0	100	10	0.34	6.36	*	8.0	48.9	*
100	0.0	200	10	0.66	†	*	8.4	†	*
100	0.0	300	10	1.59	†	†	17.6	†	†
100	0.347	20	10	0.09	2.51	0.07	8.2	39.1	8.9
100	0.576	20	10	0.15	†	0.13	8.6	†	9.4
100	1.151	20	10	0.22	†	†	12.0	†	†

Simulation parameters and setup are detailed in the Supplementary Materials. DTWF and coalescent algorithms: AR, ARGON; SC, SCRM; CS, COSI2; approximate algorithms: AR_{10}, ARGON with minimum recombination block of size 10 μM; SC_0, SCRM with; '−l' set to 0; CS_0, COSI2 with '−u' set to 0; † represents runs terminated due to insufficient memory (>60Gb) or a memory error; * indicates that the program took longer than 100 h to complete.

set to 0 (CS_0). We find that ARGON's speed and memory usage is substantially improved, at the cost of slightly inflated correlation for neighboring markers (see Supplementary Materials). For a constant population of size $N_e = 20\,000$, for instance, squared correlation (r^2) of markers 0 to 50 Kb apart (using 1 cM/Mb) was increased by ~27% for AR_{10}, but remained unchanged for markers at a larger distance. For CS_0 and SC_0, r^2 was decreased by ~5% for markers at a larger distance. AR_{10} is faster than SC_0, and approximately as fast as CS_0, and uses less memory than both simulators for the large values of the test parameters. Comparison to other simulators and additional tests are detailed in the Supplementary Materials.

Acknowledgements

We thank Alkes L. Price and John Wakeley for useful discussions and comments on an early draft.

Funding

This work was supported by the National Institutes of Health (NIH) grant (R01 MH101244).

Conflict of Interest: none declared.

References

Bhaskar,A. *et al.* (2014) Distortion of genealogical properties when the sample is very large. *Proc. Natl. Acad. Sci. USA*, **111**, 2385–2390.

Fisher,R.A. *et al.* (1922) On the dominance ratio. *Proc. Royal Soc. Edinburgh*, **42**, 321–341.

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

Kelleher,J. *et al.* (2015) Efficient coalescent simulation and genealogical analysis for large sample sizes. *bioRxiv*, 033118.

Kingman,J.F.C. (1982) The coalescent. *Stochastic Process. Appl.*, **13**, 235–248.

Liang,L. *et al.* (2007) GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics*, **23**, 1565–1567.

McVean,G.A.T. and Cardin,N.J. (2005) Approximating the coalescent with recombination. *Philos. Trans. Royal Soc B Biol. Sci.*, **360**, 1387–1393.

Palamara,P.F. *et al.* (2012) Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet.*, **91**, 809–822.

Shlyakhter,I. *et al.* (2014) Cosi2: An efficient simulator of exact and approximate coalescent with selection. *Bioinformatics*, **30**, 3427–3429.

Staab,P.R. *et al.* (2015) scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, **31**, 1680–1682.

Wakeley,J. and Takahashi,T. (2003) Gene genealogies when the sample size exceeds the effective size of the population. *Mol. Biol. Evol.*, **20**, 208–213.

Wiuf,C. and Hein,J. (2000) The coalescent with gene conversion. *Genetics*, **155**, 451–462.

Wright,S. (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97.