Journal of Theoretical Biology 437 (2018) 51–57

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

Evaluating genetic drift in time-series evolutionary analysis

Nuno R. Nené[a], Ville Mustonen[b,c], Christopher J. R. Illingworth[a,*]

[a] Department of Genetics, University of Cambridge, Cambridge, UK
[b] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
[c] Department of Biosciences, Department of Computer Science, Institute of Biotechnology, University of Helsinki, Helsinki 00014, Finland

ARTICLE INFO

Article history:
Received 12 July 2016
Revised 20 June 2017
Accepted 18 September 2017
Available online 25 September 2017

Keywords:
Genetic drift
Time-resolved genome sequence data
Wright–Fisher model
Experimental evolution

ABSTRACT

The Wright–Fisher model is the most popular population model for describing the behaviour of evolutionary systems with a finite population size. Approximations have commonly been used but the model itself has rarely been tested against time-resolved genomic data. Here, we evaluate the extent to which it can be inferred as the correct model under a likelihood framework. Given genome-wide data from an evolutionary experiment, we validate the Wright–Fisher drift model as the better option for describing evolutionary trajectories in a finite population. This was found by evaluating its performance against a Gaussian model of allele frequency propagation. However, we note a range of circumstances under which standard Wright–Fisher drift cannot be correctly identified.

1. Introduction

Rapid advances in high-throughput methodologies have enabled the collection of rich time-series from experimental evolution studies. These typically address the effects of environmental conditions on adaptation stemming from *de novo* mutations (Barrick and Lenski, 2013), initial variance induced by a genetic cross (Bergström et al., 2014; Culleton et al., 2005; Mancera et al., 2008) or simply from the standing variation characterizing a polymorphic starting population (Schlötterer et al., 2014). Sequencing the emerging populations during these types of experiments allows for identification of molecular aspects behind the species' reproductive success.

Despite advances in the field, a challenge remains regarding the optimal approach for identifying loci under selection given time-resolved genomic data. Due to linkage disequilibrium, selection at a single locus can lead to changes in allele frequencies across multiple loci (Hill and Robertson, 1966), confounding single-locus approaches to the inference of selection (Illingworth and Mustonen, 2011). Further, in smaller populations, genetic drift may have a significant impact upon allele frequencies, such that the influence of selection must be distinguished from stochastic effects, arising from both propagation and sampling (Charlesworth, 2009; Jónás et al., 2016; Jorde and Ryman, 2007).

A variety of methods have been proposed for inferring selection in time-series under genetic drift, utilising the Wright–Fisher drift model for forward propagation (Ewens, 2012), approximations to the Wright–Fisher model (Feder et al., 2014; Lacerda and Seoighe, 2014; Tataru et al., 2015; Terhorst et al., 2015; Topa et al., 2015; Waxman, 2011), its diffusion limit (Bollback et al., 2008) and respective spectral decomposition approaches (Song and Steinrücken, 2012; Steinrücken et al., 2014), or effective simulation methods (Foll et al., 2015; Malaspinas, 2016). Recently, an accurate beta approximation has also be shown to model important features at the absorbing boundaries which, otherwise, would not be easily attainable (Tataru et al., 2015) (see also Tataru et al. (2016) for an extensive review of other methods). However, while the Wright–Fisher model has become the standard approach to representing genetic drift, it is built upon certain modelling assumptions, including the replacement of the entire population in successive generations. As such, other models may in some respects provide a better fit to the dynamics observed in evolutionary experiments (Der et al., 2011). Experimental demonstrations intended to validate the Wright–Fisher model have suffered from limitations in the extent of data available for analysis (Buri, 1956; Der et al., 2011).

Here, we evaluate the extent to which a Wright–Fisher model of genetic drift can be inferred from data pertaining to evolutionary trajectories, contrasting it with a model of Gaussian diffusion. The Gaussian model at first sight differs greatly from the Wright–Fisher model, lacking frequency-dependent variance, albeit we note that, when compounded with the effect of finite sampling, frequency-dependent variance does arise in the Gaussian

* Corresponding author.
E-mail address: cjri2@cam.ac.uk (C. J. R. Illingworth).

https://doi.org/10.1016/j.jtbi.2017.09.021

model. A further contrast is noted in the computational efficiency of the algorithms; the Gaussian model is analytically solvable, allowing for rapid evaluation, whereas the Wright–Fisher model is more computationally intensive. We test the extent to which a model of drift is identifiable from simulated allele frequency data and a large dataset from evolutionary experiments conducted in *Drosophila melanogaster* (Franssen et al., 2015; Orozco-terWengel et al., 2012). We note that correct inference of a Wright–Fisher model is not always possible from simulated Wright–Fisher data, with various parameters influencing model identifiability. However, data from evolutionary experiments shows evidence in favour of a Wright–Fisher drift model under a likelihood-based inference approach.

2. Results

The potential to correctly identify a model of drift was evaluated using a Hidden Markov Model with an independent emission component, based on a version of the Kalman filter (Barber, 2012; Fischer et al., 2014). In general terms, we represented the frequency of an allele as a probability distribution, propagated at each generation, and observed via a finite sequencing process. Our emission model thus represents a form of uncertainty equivalent to that arising from evolutionary experiments that have used the Pool-Seq paradigm (Kofler et al., 2012). Given Gaussian and Wright–Fisher models of propagation, their relative fit to the data was evaluated using a compound log-likelihood difference, with optimal parameters identified by a standard non-linear optimization technique.

In order to test our ability to infer correct parameters from simulated data, given the combination of the drift model with an emission component, we tested our model against 2 batches of simulations covering several population sizes and variances for the Wright–Fisher and the Gaussian model respectively. Fig. 1, shows that accurate parameter inference was achieved under each drift model. At large population sizes (or smaller variances), the expected rate of change in an allele frequency declines, so that a longer period of observation, represented by T, the trajectory length, was required to estimate N (or σ_G) to a high level of accuracy. Given 300 generations of data, accurate estimates of N or σ_G were obtained from all simulated populations (see Supporting Text for consideration of the effect of the number of trajectories on inferred parameters).

Given sufficient data generated from a pure Wright–Fisher or Gaussian model of drift, correct identification of the drift model could be achieved. However, a threshold time, sometimes of 300 generations or more, was required for this to be achieved (Fig. 2). We tested a diverse set of simulated data with several representative parameters of typical *E&R* experiments (Kofler and Schlötterer, 2014): sequencing depth, sampling period, initial allele frequency, experimental duration and population size. The underlying population size of the system, N, was a critical factor in determining the threshold for identification; at higher N, the change via drift may be insufficient for model discrimination. Further factors influenced this value; for example, trajectories starting at lower frequencies were more informative of the drift model due to increased frequency dependence, reflected, for example, in the derivative of the characteristic variance. At frequency values closer to the boundaries, $q(t) = 0$ and $q(t) = 1$, the importance of higher-order moments characterizing the Wright–Fisher model are also a strong contributing factor. An increased depth and frequency of sampling increased the extent of information available for inference; each improved the ability for model discrimination (see Fig. 2 and additional results in Supporting Text).

While the simulations discussed above consider systems in which drift is the only force driving evolution, in a biological system, other factors affect allele frequency change. Selection, mutation, and linkage disequilibrium each influence the shape of the expected distribution of allele frequencies with time, potentially affecting the identifiability of a model of drift

Natural selection acting upon a population induces changes in allele frequency over time. As such, including selection in our simulations led to an increased allele frequency variance in our simulation data. Subsequent inference of N under a neutral assumption led to underestimates of N proportionate to the number of loci at which selection acted. However, the correct inference of a Wright–Fisher drift model in each case was not compromised (see Supporting Information).

The rate of mutation in experimental systems relevant to our work, of close to $\mu \approx 10^{-9}$ (Li and Stephan, 2006), has an influence on allele frequencies much smaller than the effect of genetic drift. To explore the theoretical effect of mutation, simulations were conducted with much higher rates of mutation. From simulated data, population sizes were over-estimated if the starting frequency was 0.1 and $\mu N = 0.1$ or 0.5, and under-estimated if $\mu N = 1$ or 10 (see also Supplementary Information). At low frequencies, the influence of mutation led to incorrect model identification; the Gaussian distribution describes with greater flexibility the sample paths generated by the balance between drift, which pushes trajectories towards either of the absorbing boundaries, and mutation, which drives the frequency spectrum away from a frequency of 0 or 1. Where μN is sufficiently high, drift is overcome by the tendency of mutation to push frequencies to $q(t) = 0.5$. Considering simulations with a starting frequency of 0.5, consistent overestimates of N were obtained to compensate for the effect of mutation keeping the allele frequency close to a constant value. However, in these cases, the Wright–Fisher model was correctly identified in comparison to the Gaussian drift model.

The presence of linkage disequilibrium between loci may act as a confounding factor for selection identification. Yet, for model identifiability without selection, hitch-hiking effects should only have a significant impact if the number of founding haplotypes is reduced or if the size of genomes is small (Franssen et al., 2015; Terhorst et al., 2015). Under these conditions, a random bias in allele frequency change may be observed, leading to possible incorrect model identification. For the simulated genomes under a neutral coalescent model employed here (see Methods), propagation with linkage, even for a low number of founding haplotypes, did not lead to incorrect drift model identification. Population sizes for these datasets were slightly over-estimated (see Supplementary Information).

Applying the model to experimental genomic data (Franssen et al., 2015), an improved fit was not seen for the Wright–Fisher model across all statistical measures considered (see Supporting Text, where the error in the estimated compound variance is evaluated). However, a clear result in favour of this model was seen via a likelihood calculation. Estimated population sizes calculated under the Wright–Fisher model are shown in Fig. 3 (A). Consistent with the identification of selection in the data (Franssen et al., 2015), these estimates are lower than the reported consensus size of 1000. Further calculations were performed to evaluate models of drift over the subset of loci in all chromosomes that did not reach fixation. This was intended to verify whether the improved performance of the Wright–Fisher model arose from the natural inclusion of fixation events in this drift model; a more artificial approach was required in the case of the Gaussian drift model. While average likelihood differences for this dataset were reduced, the tendency across chromosomes observed in Fig. 3 was not altered.

In the results of Fig. 3, differences between the estimates obtained were observed for different replica datasets. As noted in supplementary Fig. F.14, the differences between initial distributions is minimal, likely excluding this as an explanation for the