Bacmeta: simulation for genomic evolution in bacterial metapopulations

Aleksi Sipola [1,2], Pekka Marttinen [2] and Jukka Corander [1,3,4]

[1]Department of Mathematics and Statistics, University of Helsinki, Helsinki, 00014, Finland;
[2]Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, 00076 Aalto, Finland;
[3]Pathogen Genomics, Wellcome Trust Sanger Institute, Cambridge, CB10 1SA, UK;
[4]Department of Biostatistics, University of Oslo, Oslo, 0317, Norway.

Abstract:

The advent of genomic data from densely sampled bacterial populations has created a need for flexible simulators by which models and hypotheses can be efficiently investigated in the light of empirical observations. Bacmeta provides fast stochastic simulation of neutral evolution within a large collection of interconnected bacterial populations with completely adjustable connectivity network. Stochastic events of mutations, recombinations, insertions/deletions, migrations and microepidemics can be simulated in discrete non-overlapping generations with a Wright-Fisher model that operates on explicit sequence data of any desired genome length. Each model component, including locus, bacterial strain, population, and ultimately the whole metapopulation, is efficiently simulated using C++ objects, and detailed metadata from each level of the simulation can be acquired. The software can be executed in a cluster environment using simple textual input files, enabling, e.g., large-scale simulations and likelihood-free inference. Bacmeta is implemented with C++ for Linux, Mac and Windows. It is available at https://bitbucket.org/aleksisipola/bacmeta under the BSD 3-clause license.

Contact:
aleksi.sipola@helsinki.fi,
jukka.corander@medisin.uio.no
Supplementary information: Supplementary data are available online at bioRxiv.

1 Introduction

Simulation models can be used for prediction, parameter estimation, and for validating methods used in population genomics (Hoban *et al.*, 2012). Most general-purpose simulators are tailored mainly for eukaryotes (e.g., Arenas and Posada, 2014). However, many studies on evolutionary processes in bacteria have emerged recently, using simulation software tailored for their specific purposes (Fraser *et al.*, 2007; Friedman *et al.*, 2013; Marttinen *et al.*, 2015; Niehus *et al.*, 2015; Numminen *et al.*, 2016). Simulators can be divided into two categories (Hoban *et al.*, 2012): coalescent simulation starts with the present-day population and simulates backwards in time, coalescing individuals until the most recent common ancestor is found, while forward simulators maintain a population of individuals and simulate forward in time by sampling the next generation from the current one. In general, coalescent simulators are faster, by only considering the ancestors of the current individuals, but forward simulation allows greater flexibility to define the model. This makes the latter particularly attractive for bacteria, where recombination shuffles genetic material between genomes in a complex manner that depends, for example, on the genetic and physical distance between the donor and recipient strains. Furthermore, recombination may cause different parts of the genome to have completely distinct population histories (Feil *et al.*, 2001; Mostowy *et al.*, 2017), undermining the assumption of a single coalescent. The recently published general-purpose simulators tailored for bacteria have all been based on the coalescent approach (Brown *et al.*, 2016; De Maio and Wilson, 2017). Hence, there is a need for an efficient general-purpose forward simulator for bacterial population genomics.

Bacmeta provides an efficient C++ implementation of a finite metapopulation Wright-Fisher model with explicit genome sequences evolving for each strain present in the metapopulation. Use of shared pointers of C++ and compact object representations result in low memory and runtime requirements. The model allows multiple arbitrarily connected populations, each with thousands of bacteria, for which the genome sequences are subjected to evolutionary events over discrete non-overlapping generations. Bacmeta implements a large variety of different event types governed by user-defined parameters using simple textual input files, which provides a convenient framework for large-scale simulations, integration with other software and likelihood-free inference. For example, Bacmeta could be used for testing methods for inferring recombination (Croucher *et al.*, 2014; Didelot and Wilson, 2015; Mostowy *et al.*, 2017), since every past evolutionary event can be stored to provide the ground-truth. Another potential application is likelihood-free inference for model parameters, as in Marttinen *et al.* (2015); Numminen *et al.* (2016); De Maio and Wilson (2017), based on the Approximate Bayesian Computation inference framework (Beaumont *et al.*, 2002; Lintusaari *et al.*, 2017).

2 Features

The main input parameters for Bacmeta are defined in a plain ascii text file. Optional migration parameters can be given in a second input file. Examples of these are given in Tables S1 and S2. The simulation of each evolutionary event, including reproduction, is executed at each generation for a desired number of iterations. The events that can be simulated are displayed in Figure 1A and B.

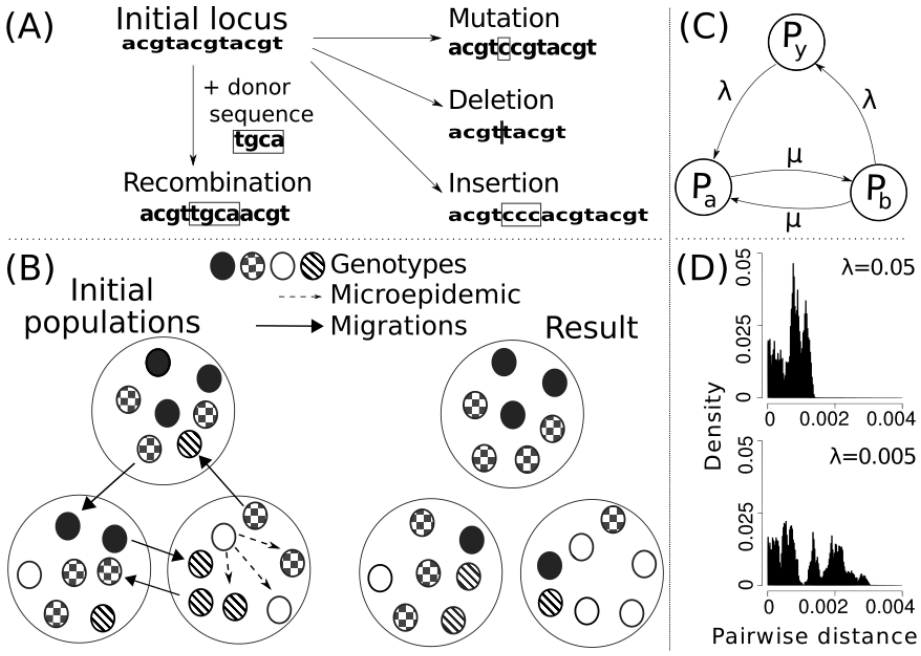

Figure 1: (A) Illustrations of the evolutionary events and (B) illustrations of population dynamic events, excluding random sampling between generations. (C) Example case: The migration connectedness of metapopulation as a network graph, where P_y is the observed population and edge weight λ represents the inward and outward migration rates of P_y and $\mu = 0.01$. (D) Effect of low versus high value of λ on pairwise distances in population P_y. Computed from 10 observed simulations per λ-value.

The order of the events can be fixed or random. Each generation ends in the selection of bacteria for seeding the next generation by random sampling with replacement. The count of each event type per generation is modeled as a Poisson process with a user-defined rate parameter. For migration and microepidemic events we use the parametrization introduced by Numminen *et al.* (2016). For mutations and recombinations we use the same approach as Marttinen *et al.* (2015), except that mutations are generated under an explicit

mutation model with separate user-defined weights for all nucleotide pairs in ACGT. For insertion/deletion sizes we use the model presented by Benner *et al.* (1993), with rate defined in relation to mutations. Note that this also allows for more coarse-grained summaries of the produced data, for example as an infinite alleles type integer labeling for each locus, useful for certain types of genomic analyses. Haplotypes can be flexibly represented as genomic islands of any desired length and number, such that separate genomic regions or secondary chromosomes can be imitated. Outputs from the simulator include synthetic DNA sequences, pairwise distance measures and several different summaries, e.g., counts of the different events.

3 Example case

For an illustrative example of the functionality and performance of Bacmeta, we considered the effect of inter-population connectedness via migration parameter setup. We simulated 10 occurrences of the cases: *i*) low connectedness of observed population P_y, and case *ii*) high connectedness of observed population P_y, each for 20 000 generations. We used a metapopulation consisting of three populations, with migration routes as illustrated in Figure 1C and the migration rate input file displayed in Table S2. General simulation parameters were as given in Table S1 and followed values of recombinogenic bacteria. Figure 1D shows results. As expected in this setup, the higher connectivity led to markedly lower pairwise distances, due to the reduced divergence between the populations. Runtimes for these simulations were between 35-50 seconds on a single core of Intel Core i5-7200U CPU @ 2.50GHz.

Funding

This work has been funded by the Academy of Finland (COIN Centre of Excellence and grants 286607 and 294015 to PM) and ERC (grant 742158 to JC).

References

Arenas, M. and Posada, D. (2014). Simulation of genome-wide evolution under heterogeneous substitution models and complex multispecies coalescent histories. *Molecular Biology and Evolution*, **31**(5), 1295.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, **162**(4), 2025–2035.

Benner, S. A., Cohen, M. A., and Gonnet, G. H. (1993). Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *Journal of Molecular Biology*, **229**(4), 1065 – 1082.

Brown, T., Didelot, X., Wilson, D. J., and De Maio, N. (2016). Simbac: simulation of whole bacterial genomes with homologous recombination. *Microbial genomics*, **2**(1).

Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., Parkhill, J., and Harris, S. R. (2014). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using gubbins. *Nucleic acids research*, page gku1196.

De Maio, N. and Wilson, D. J. (2017). The bacterial sequential markov coalescent. *Genetics*, **206**(1), 333–343.

Didelot, X. and Wilson, D. J. (2015). Clonalframeml: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol*, **11**(2), e1004041.

Feil, E. J., Holmes, E. C., Bessen, D. E., Chan, M.-S., Day, N. P., Enright, M. C., Goldstein, R., Hood, D. W., Kalia, A., Moore, C. E., *et al.* (2001). Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proceedings of the National Academy of Sciences*, **98**(1), 182–187.

Fraser, C., Hanage, W. P., and Spratt, B. G. (2007). Recombination and the nature of bacterial speciation. *Science*, **315**(5811), 476–480.

Friedman, J., Alm, E. J., and Shapiro, B. J. (2013). Sympatric speciation: when is it possible in bacteria. *PLoS ONE*, **8**(1), e53539.

Hoban, S., Bertorelle, G., and Gaggiotti, O. E. (2012). Computer simulations: tools for population and evolutionary genetics. *Nature Reviews Genetics*, **13**(2), 110–122.

Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2017). Fundamentals and recent developments in approximate bayesian computation. *Systematic biology*, **66**(1), e66–e82.

Marttinen, P., Croucher, N. J., Gutmann, M. U., Corander, J., and Hanage, W. P. (2015). Recombination produces coherent bacterial species clusters in both core and accessory genomes. *Microbial Genomics*, **1**(5).

Mostowy, R., Croucher, N. J., Andam, C. P., Corander, J., Hanage, W. P., and Marttinen, P. (2017). Efficient inference of recent and ancestral recombination within bacterial populations. *Molecular biology and evolution*, **34**(5), 1167–1182.

Niehus, R., Mitri, S., Fletcher, A. G., and Foster, K. R. (2015). Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nature Communications*, **6**.

Numminen, E., Gutmann, M., Shubin, M., Marttinen, P., Meric, G., van Schaik, W., Coque, T. M., Baquero, F., Willems, R. J., Sheppard, S. K., *et al.* (2016). The impact of host metapopulation structure on the population genetics of colonizing bacteria. *Journal of theoretical biology*, **396**, 53–62.