SimMLST: simulation of multi-locus sequence typing data under a neutral model

Xavier Didelot1,∗, Daniel Lawson2 and Daniel Falush3
1Department of Statistics, University of Warwick, Coventry, 2Department of Mathematics, University of Bristol, Bristol, UK and 3Department of Microbiology, University College Cork, Cork, Ireland

Received on February 17, 2009; revised on March 9, 2009; accepted on March 11, 2009
Advance Access publication March 13, 2009

1 INTRODUCTION

Multi-locus sequence typing (MLST) was introduced by Maiden et al. (1998) as a method of characterization of bacterial isolates from a given species. It relies on the sequencing of several housekeeping gene fragments of 400–500 bp each to determine the type of an isolate. MLST was originally proposed for the typing of isolates of Neisseria meningitidis (Maiden et al., 1998), but has since been applied to over 50 000 isolates from over 50 different species (Maiden, 2006; Urwin and Maiden, 2003). MLST results can easily be shared and compared between laboratories (Urwin and Maiden, 2003), and are routinely made available on the http://web.mpiib-berlin.mpg.de/mlst/ web site hosted by the Max Planck Institute.

The ability to simulate MLST under a neutral model is useful to make interpretations about sampled datasets, for example, to infer the values of evolutionary parameters (Fearnhead et al., 2005; Fraser et al., 2005), to analyze the role played by selection (Buckee et al., 2008), to apply approximate Bayesian computing methods (Marjoram et al., 2003; Wilson et al., 2009) or to test methods of genealogical inference (Didelot and Falush, 2007; Falush et al., 2006; Turner et al., 2007). For this last task, it is necessary to simulate the clonal genealogy (Guttman, 1997) that gave rise to the data as well as the data itself.

∗To whom correspondence should be addressed.

Several methods of MLST simulations have been described before. Both Fraser et al. (2005) and Falush et al. (2006) used a forward in-time approach which required to simulate a whole population (rather than just a sample) and to wait until equilibrium is reached. Fearnhead et al. (2005) used the backward in time (coalescent) simulation program MS (Hudson, 2002) to generate a single large genetic region which contained the MLST loci at a large (10 kb) distance from one another. Here, we present a more efficient method to simulate MLST data.

2 MODEL

The basic model we assume is the coalescent with gene conversion (Wright and Heinit, 2000). This model is similar to the popular coalescent with recombination (Hudson, 1983), but assumes that when two cells recombine, the resulting genome is identical to that of the receiver except for a (small) contiguous fragment which comes from the donor. Since this is how recombination takes place in bacteria (be it through conjugation, transformation or transduction), the coalescent with gene conversion is the appropriate model to simulate MLST data. We follow Wright and Heinit (2000) in assuming that a gene conversion event is equally likely to be initiated at any point on the genome and that the length of an import is geometrically distributed with parameter \( \delta \), which is consistent with empirical evidence (Falush et al., 2001; Fearnhead et al., 2005; Jolley et al., 2005).

3 ALGORITHM

Wright and Heinit (2000) proposed an algorithm to simulate a single locus under the coalescent with gene conversion. Here we extend their algorithm in three respects. First we perform multi-locus simulation by using the result from Didelot and Falush (2007, Equation 4) that the starting point for a recombination is \( \delta \) times more likely to be at the beginning of a locus than it is to be at a site within a locus. Second we only simulate the ancestral material (Hudson, 1983, 2002) of each lineage for efficiency, that is the positions which are ancestral to at least one individual in the sample. We use rejection sampling to ignore any recombination event that does not split the ancestral material of a lineage into two non-empty subsets. Third, we jointly simulate the clonal genealogy (Guttman, 1997) with the data. The clonal genealogy is obtained by tracing the lineage that is the recipient at each recombination event. Correct simulation of
Fig. 1. Genealogical history for a simulated sample of three isolates and two loci. Each locus is represented by a box, in which the ancestral material is in gray. The clonal genealogy is in bold.

the clonal genealogy requires to allow it not to carry any ancestral material, unlike other lineages as described above.

Figure 1 illustrates the working of our algorithm for a sample of three isolates and two genes. The ancestry of the sample is traced back in time until all isolates find a common ancestor at all sites. The clonal genealogy is represented in bold on Figure 1. The second isolate is the result of a recombination in which a fragment of the second gene was imported, and the first gene of the first isolate was imported from above the clonal root. Recombination allows different gene fragments to have different genealogies: although isolates 2 and 3 are the most closely related in the first gene, isolates 1 and 2 are the most closely related in the inserted fragment of the second gene.

Our algorithm is compatible with any population dynamics model by simple rescaling of the timescale in the ancestry graph (Griffiths and Tavare, 1994) and we allow specification of any piecewise exponential or constant dynamics through the use of program arguments similar to those of MS (Hudson, 2002). Data are then generated by adding mutations as a Poisson process on the ancestry graph. We use the mutational model of Jukes and Cantor (1969) by default, but our program can be used in conjunction with seq-gen (Rambaut and Grass, 1997) to simulate a wide range of other models.

4 CONCLUSION

SimMLST jointly simulates MLST data and the clonal relationships between isolates. It outputs the MLST data in the flexible extended Multi-Fasta Alignment (XMFA) format, and the clonal genealogy in the Newick format. SimMLST can also be used in conjunction with the graph-drawing package DOT (Gansner et al., 1993) to represent the full genealogical history of a sample (as shown in Fig. 1).

SimMLST is more efficient than previous methods because it only simulates the recombination events that had an impact on the data, not those that fall out of the sequenced regions or out of the ancestral material of a lineage. It is therefore optimal in the size of the ancestral graphs that it generate to simulate the data. Like all coalescent-based methods, the time and memory requirements of SimMLST increase much faster than linearly with the overall recombination rate \( \rho \). Yet it can support values of \( \rho \) up to several thousands, which is more than recorded in the MLST of any bacterial species (Didelot and Falush, 2007; Fearnhead et al., 2005; Fraser et al., 2005; Jolley et al., 2005).

The high efficiency of SimMLST is useful in order to infer evolutionary parameters from simulations, which typically requires to generate thousands of datasets with a wide range of parameters. It is also required to generate larger datasets (in the number of sequenced sites) than MLST, where \( \rho \) will be higher. Assuming a per-site recombination rate similar to that observed in MLST data for Neisseria meningitidis, SimMLST can generate datasets up to a few hundreds of kilo base pair. An approximation to the coalescent with gene-conversion process would however be required to simulate whole genomes for frequently recombining species.

ACKNOWLEDGEMENTS

The authors thank the editor and two anonymous referees for their insightful comments.

Funding: Wellcome Trust.
Conflict of Interest: none declared.

REFERENCES

X. Didelot et al.


