Bioinformatics Advance Access published March 13, 2009

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

Xavier Didelot [1,*], Daniel Lawson [2] and Daniel Falush [3]

[1]Department of Statistics, University of Warwick, UK
[2]Department of Mathematics, University of Bristol, UK
[3]Department of Microbiology, University College Cork, Ireland

Associate Editor: Prof. Martin Bishop

ABSTRACT

Summary: Multi-locus sequence typing (MLST) is a widely used method of characterization of bacterial isolates. It has been applied to over 50,000 isolates in over 50 different species. Here we present a coalescent method to jointly simulate MLST data and the clonal genealogy that gave rise to the sample.

Availability and Implementation: SimMLST was implemented in C++ and Qt4 for the graphical user interface. It is distributed under the terms of the GNU General Public License. Source code and binaries for Windows and Linux are available from http://go.warwick.ac.uk/SimMLST. A user guide and a technical description of the algorithm are provided with the program.

Contact: X.Didelot@warwick.ac.uk

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 400-500bp 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 (Urwin and Maiden, 2003; Maiden, 2006). MLST results can easily be shared and compared between laboratories (Urwin and Maiden, 2003), and are routinely made available on the http://pubMLST.org/ website hosted by the University of Oxford, the http://www.mlst.net/ website hosted by Imperial College and the http://web.mpiib-berlin.mpg.de/mlst/ website 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 (Fraser *et al.*, 2005; Fearnhead *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 (Falush *et al.*, 2006; Didelot and Falush, 2007; 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 (10kbp) 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 (Wiuf and Hein, 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 Wiuf and Hein (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 δ, which is consistent with empirical evidence (Falush *et al.*, 2001; Fearnhead *et al.*, 2005; Jolley *et al.*, 2005).

3 ALGORITHM

Wiuf and Hein (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) (Eq. 4) that the starting point for a recombination is δ 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

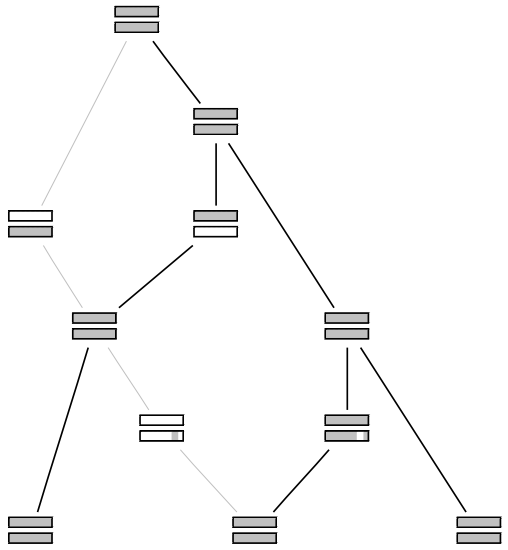

Fig. 1. Genealogical history for a simulated sample of three isolates and two loci. Each locus is represented by a box, in which the ancestral material is in grey. The clonal genealogy is in bold.

clonal genealogy is obtained by tracing the lineage that is the recipient at each recombination event. Correct simulation of the clonal genealogy requires to allow it not to carry any ancestral material, unlike other lineages as described above.

Fig. 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 Fig. 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 is 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 ρ. Yet it can support values of ρ up to several thousands, which is more than recorded in the MLST of any bacterial species (Fearnhead *et al.*, 2005; Jolley *et al.*, 2005; Fraser *et al.*, 2005; Didelot and Falush, 2007).

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 ρ will be higher. Assuming a per-site recombination rate similar to that observed in MLST data for *N. meningitidis*, SimMLST can generate datasets up to a few hundreds of Kbp. An approximation to the coalescent with gene-conversion process would however be required to simulate whole genomes for frequently recombining species.

ACKNOWLEDGEMENT

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

Funding: This work was funded by a grant from the Wellcome Trust.

Conflict of Interest: None declared.

REFERENCES

Buckee, C. O., Jolley, K. A., Recker, M., Penman, B., Kriz, P., Gupta, S., and Maiden, M. C. J. (2008). Role of selection in the emergence of lineages and the evolution of virulence in Neisseria meningitidis. *Proceedings of the National Academy of Sciences*, **105**(39), 15082–15087.

Didelot, X. and Falush, D. (2007). Inference of Bacterial Microevolution Using Multilocus Sequence Data. *Genetics*, **175**(3), 1251–1266.

Falush, D., Kraft, C., Taylor, N. S., Correa, P., Fox, J. G., Achtman, M., and Suerbaum, S. (2001). Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proc Natl Acad Sci U S A*, **98**(26), 15056–15061.

Falush, D., Torpdahl, M., Didelot, X., Conrad, D., Wilson, D., and Achtman, M. (2006). Mismatch induced speciation in Salmonella: model and data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **361**(1475), 2045–2053.

Fearnhead, P., Smith, N. G., Barrigas, M., Fox, A., and French, N. (2005). Analysis of recombination in *Campylobacter jejuni* from MLST population data. *J Mol Evol*, **61**(3), 333–340.

Fraser, C., Hanage, W. P., and Spratt, B. G. (2005). Neutral microepidemic evolution of bacterial pathogens. *Proc Natl Acad Sci U S A*, **102**(6), 1968–1973.

Gansner, E., Koutsofios, E., North, S., and Vo, K. (1993). A technique for drawing directed graphs. *IEEE Transactions on Software Engineering*, **19**(3), 214–230.

Griffiths, R. and Tavare, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **344**(1310), 403–410.

Guttman, D. (1997). Recombination and clonality in natural populations of Escherichia coli. *Trends in Ecology & Evolution*, **12**(1), 16–22.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, **23**, 183–201.

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

Jolley, K. A., Wilson, D. J., Kriz, P., Mcvean, G., and Maiden, M. C. J. (2005). The influence of mutation, recombination, population history, and selection on patterns of

genetic diversity in *Neisseria meningitidis*. *Mol Biol Evol*, **22**(3), 562–569.

Jukes, T. H. and Cantor, C. R. (1969). *Evolution of protein molecules*, pages 21–132. H. N. Munro, ed., Mammalian Protein Metabolism.

Maiden, M. (2006). Multilocus Sequence Typing of Bacteria. *Annu Rev Microbiol*, **60**, 561–88.

Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M., and Spratt, B. G. (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *PNAS*, **95**(6), 3140–3145.

Marjoram, P., Molitor, J., Plagnol, V., and Tavare, S. (2003). Markov chain monte carlo without likelihoods. *Proc Natl Acad Sci U S A*, **100**(26), 15324–15328.

Rambaut, A. and Grass, N. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, **13**(3), 235–238.

Turner, K., Hanage, W., Fraser, C., Connor, T., and Spratt, B. (2007). Assessing the reliability of eburst using simulated populations with known ancestry. *BMC Microbiology*, **7**(1), 30.

Urwin, R. and Maiden, M. (2003). Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol*, **11**, 479–87.

Wilson, D. J., Gabriel, E., Leatherbarrow, A. J., Cheesbrough, J., Gee, S., Bolton, E., Fox, A., Hart, C. A., Diggle, P. J., and Fearnhead, P. (2009). Rapid Evolution and the Importance of Recombination to the Gastroenteric Pathogen Campylobacter jejuni. *Mol Biol Evol*, **26**(2), 385–397.

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