Bacmeta - Model manual

Contents

Bacmeta provides simulations of a forward-time, finite metapopulation Wright-Fisher, neutral evolution model with explicit genome sequences. The simulations begin with single genotype and progress in forward time with mutations, recombinations, indels, migrations and microepidemics occurring stochastically in the populations across discrete generations.

This manual describes how the events are simulated and provides references to previous work on which the model is based. Relevant parameters for each event are shown as the parameter key in brackets, e.g. [$NLOC$].

See InputOutputManual.pdf for further information about the parameters.

Developer welcomes all forms of feedback, help inquiries, developing suggestions and general correspondence about the program via: aleksi.sipola [at] helsinki.fi

Initial stage

The model assumes metapopulation consisting of $n_{pops} = [NPOP]$ populations. These populations each have same amount of bacteria, $n_{bacs} = [NBAC]$. Each bacterium has $n_{loci} = [NLOC]$ loci of length $L = [LOLE]$. The bases for initial sequences are sampled based on probabilities defined by parameters $[PROA]$, $[PROT], [PROG]$ and $[PROC]$.

Mutations

Mutations are modeled as a Poisson process with rate $\lambda_{mutations}$, as in Marttinen *et al.* (2015), but with the distinction of implementing explicit sequences instead of infinite alleles model.

The amount of mutations in generation is sampled for each population with rate:

$$\lambda_{mutations} = \theta n_{bacs} n_{loci} L = [MUTR] \cdot [NBAC] \cdot [NLOC] \cdot [LOLE]$$

For each mutation, the site is randomly sampled from a random bacterium. The outcome of a mutation is sampled based on probabilities in a mutation matrix that is defined by parameters $[ATOT], [ATOG][ATOC], [TTOA]$...

In this example matrix, each row represents initial base at the site, columns represent the mutation outcome and the elements show corresponding parameter key and example value:

	A	**T**	**G**	**C**
A	0	ATOT=0.7	ATOG=0.2	ATOC=0.1
T	TTOA=0.3	0	TTOG=0.4	TTOC=0.3
G	GTOA=0.1	GTOT=0.2	0	GTOC=0.6
C	CTOA=0.4	CTOT=0.3	CTOG=0.1	0

Recombinations

Recombinations are also modelled as Poisson process, with rate defined in relation to mutations:

$$\begin{aligned}\lambda_{recombinations} &= \rho n_{bacs} n_{loci} L \\ &= s\theta n_{bacs} n_{loci} L \\ &= [RECR] \cdot [MUTR] \cdot [NBAC] \cdot [NLOC] \cdot [LOLE]\end{aligned}$$

where s is the ratio of recombinations per mutations, i.e. $s = \rho/\theta$.

The recombination lengths can be set either to full locus or to be sampled from geometric distribution with mean $[RECL]$. The model assumes the loci to have a base distance greater than recombination lengths, so recombinations only affect one locus at a time.

The model allows for recombination acceptance testing based on similarity of donor and recipient sequences. Here, the probability of a recombination succeeding is defined as:

$$P(Accept\ recombination) = 10^{-hA/l} = 10^{-h[RECA]/l}$$

where h is the Hamming distance between the sequences, l is the length of recombining sequences and A is parameter $[RECA]$. Setting $[RECA] = 0$ results in acceptance of all recombinations, effectively disabling this similarity testing. These features of the recombination modelling also follow Marttinen *et al.* (2015).

In this testing, indels are considered similarly to base differences: When sequences are aligned with their true corresponding sites (i.e. non-matching insertion or deletion position are filled with "NULL" symbols), every non-matching position between sequences increases the Hamming distance by one. Insertions are compared by aligning their starting positions and concatenating the shorter insertion to same length with "NULL"-symbols.

Indels

Insertions and deletions are modelled also as Poisson process, with rates defined in relation to mutations:

$$\lambda_{insertion} = \iota\lambda_{mutations} = [INSR] \cdot [MUTR] \cdot [NBAC] \cdot [NLOC] \cdot [LOLE]$$
$$\lambda_{deletion} = \delta\lambda_{mutations} = [DELR] \cdot [MUTR] \cdot [NBAC] \cdot [NLOC] \cdot [LOLE]$$

The length of each indel is modelled based on the model shown in Benner *et al.* (1993). Here each indel has length sampled from Zipf distribution with parameter $a_{insertion} = [INSL]$ and $a_{deletion} = [DELL]$. A maximum length for indels can be set with $[INDM]$.

The bases for insertions are sampled with the probabilities that are used for initial sequences, i.e. $[PROA], [PROT], [PROG]$ and $[PROC]$.

Migrations

Migrations are modelled along Numminen *et al.* (2016), so that in one migration event m random bacteria from population i replace m bacteria from population j. In each generation, migration event for each of the routes "from population i to population j" happens with probability $[MIGP]$. The amount of migrating bacteria, m, is sampled for each migration event from Poisson distribution with rate:

$$\tau n_{bacs} = \tau \cdot [NBAC]$$

where τ is either global rate $[MIGR]$ or a specific rate defined for the route in question. The route specific non-homogeneous migration rates are set using input file "migration.input" and are given as adjacency matrix for directed graph, so that edge weights represent migration rates for each "from population i to population j"-route . See **"InputOutputManual.pdf"** for furthers instructions.

Microepidemics

Microepidemics are also modelled along Numminen *et al.* (2016), so that in a microepidemic event k copies of one random bacteria replace k random bacteria in the same population. The amount of microepidemics in generation is modelled for each population as Poisson process with rate

$$\omega n_{bacs} = [MICS] \cdot [NBAC]$$

The size k of microepidemic event is modelled as Poisson process with rate:

$$\gamma = [MICA]$$

Random selection and event order

In end of each generation, new population is sampled from current population using random selection with replacement. This can be toggled of with parameter $[RSEL]$.

The default order of events is mutations, indels, recombinations, migration and then microepidemics. The order can be set to be shuffled for each generation with the input parameter $[RORD]$.

References

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.

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).

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.