Using ModifyXML.py to write input files for transmission tree reconstruction in BEAST (BEASTLIER)

Matthew Hall

1 What you will need

- An alignment of sequence data from a pathogen outbreak. The current version of BEASTLIER assumes that one or more sequences are available from every clinical case, and is designed for datasets of this sort. However, known clinical cases for which genetic data is unavailable can be given a noninformative sequence consisting entirely of the nucleotide code N. All sequences must be given a sampling date; noninformative sequences can be given any date on which the relevant host was known to be infected.
- A CSV file of epidemiological information on each host. It must contain columns with heading "Host_ID" and "End_date"; the former should list a unique identifier for each host and the latter the time at which it became noninfectious. (Estimation of these dates is currently unimplemented but is likely to follow.) End times can be expressed as dates or as numbers. Known potential hosts that were never infected can be included here and should be given "NA" as an end date. The file can optionally include "latitude" and "longitude" columns giving spatial cooardinates; if these are absent then the model will not have a geographical component and hosts will have been assumed to be freely-mixing.
- A second CSV file giving information on the sequences. This should have a row for each sequence (including noninformative sequences) columns "Taxon_ID", "Host_ID", and "Exam_date". The host ID should be the same as that given in the first CSV file, the taxon ID the sequence name in the alignment file, and the exam date the date of collection in the same format as the end dates in the first CSV file. A single host may be the source of multiple sequences with the same or different dates, which should be given in separate rows in this file.
- An up-to-date copy of BEAST (1), version 1.8.4 or later.
- A Python installation. This script will work with no additional libraries installed, but lxml is recommended as it formats the XML output for better human readability.

2 BEAUTi

The first step is to generate a BEAST XML file using BEAUTi; tutorials can be found on the BEAST website. It is important to specify your desired substitution and molecular clock models as the BEASTLIER script will not change these. On the other hand, anything BEAUTi adds regarding tip dates or the tree prior will be overwritten, so the "Tips" and "Trees" tabs can be ignored.

3 Running the script

The script can be run on the command line ("`python ModifyXML.py`"), and takes five compulsory arguments, in order:

- The name of the XML file produced by BEAUTi

- The name of the CSV file of host data
- The name of the CSV file of taxon data
- A file name for the modified XML output
- A string that forms the root of the file names for BEAST output. For example, if this is "`output_beast`", then BEAST will write the parameter log to "`output_beast.log.txt`", the phylogenetic tree log to "`output_beast.trees.txt`", and the transmission tree log to "`output_beast.net.txt`".

Numerous optional arguments can also be supplied, which should be given before the compulsory five. If none are, then the script will prompt the user to specify the prior distributions for infectious and (if desired) latent periods. In the absence of other optional arguments, the XML will be written as follows:

- No spatial component will be present in the analysis; the force of infection that each infected host exerts on each susceptible is identical.
- The starting phylogeny and transmission tree are to be randomly generated by BEAST.
- The MCMC chain length and sampling frequency are as specified in the original BEAUTi file.
- The prior on the time of the index infection is the improper uniform distribution.
- Dates in the CSV file will be parsed as decimal numbers.

The `examples` folder of the BEASTLIER GitHub repository contains input and output files for ModifyXML. The call was:

```
python ModifyXML.py -k l -i ng 10 100 1 1 -l g 200 0.01 -- simulation_X_1_slow_constant.xml epidata_X_1.csv taxondata_X_1.csv simulation_X_1_slow_WCC.xml simulation_X_1_slow_WCC
```

The full list of possible optional arguments are as follows:

- `-d` or `--dateFormat` takes one argument and gives the format for the dates appearing in the two CSV files. See the documentation for the Python datetime library for information. If absent, dates are assumed to be decimal numbers, in units of days.
- `-k` or `--kernel` takes one argument which specifies the type of spatial kernel that will be used. This is a function $f(d)$ of the spatial distance between two hosts such that the force of infection that an infectious host applies to a susceptible a distance d away is $r \times f(d)$ where r is a base transmission rate. There are four choices:
  - "`e`", exponential, one parameter α, $f(d) = e^{-\alpha d}$
  - "`p`", power law, one parameter α, $f(d) = d^{-\alpha}$
  - "`g`", Gaussian, one parameter α, $f(d) = e^{-\alpha d^2}$
  - "`l`", logistic, two parameters α and r_0, $f(d) = \frac{1}{1+\left(\frac{d}{r_0}\right)^{\alpha}}$
  - "`x`" specifies that no geographical component is to be included and is equivalent to omitting the `-k` option.

  The script configures the XML such that all kernel parameters are estimated by the MCMC, and have an exponential prior distribution with a mean of 1. This behaviour can be changed by hand-modifying the XML.
- `-i` or `--infectiousPeriods` takes a varying number of arguments and specifies either the prior distribution of the infectious period of the infection, or, if infectious periods are assumed to be drawn from an unknown Normal distribution, the parameters of the Normal-Gamma hyperprior that specifies prior beliefs about the distribution of the mean and precision of that distribution. The script assumes the same distribution for the infectious periods of all hosts, but separate distributions can be used on different sets of hosts by modifying the XML by hand; see "Advanced XML modification" below. The first argument to `-i` specifies the type of prior and the remaining arguments its parameters. Choices are:

  - "`ng`", normal-gamma. This takes four additional arguments for the parameters μ_0, κ_0, α_0 and β_0. The prior assumption is that infectious periods are drawn from a normal distribution with unknown mean μ and precision (1 over variance) τ. The prior distribution of τ is a gamma distribution with shape α_0 and rate β_0, and given τ, the prior distribution of μ is a normal distribution with mean μ_0 and precision $\tau\kappa_0$.
  - "`n`", normal. This takes two additional arguments for the mean and variance of the distribution, in that order.
  - "`l`", lognormal. This takes two additional arguments for the mean and variance of the distribution (on the log scale), in that order.
  - "`g`", gamma. This takes two additional arguments for the shape and scale parameters of the distribution, in that order.
  - "`e`", exponential. This takes one additional argument for the mean of the distribution.

- `-l` or `--latentPeriods` also takes a varying number of arguments and specifies whether latent periods are to be included in the model, and if they are, what the prior distribution of their length should be. Note that as currently configured, BEASTLIER assumes that all latent periods are equal. The script assumes the same distribution for the infectious periods of all hosts, but separate distributions can be used on different sets of hosts by modifying the XML by hand; see "Advanced XML modification" below. The first argument to `-i` specifies the type of prior and the remaining arguments its parameters. Choices are:
  - "`n`", normal. This takes two additional arguments for the mean and variance of the distribution, in that order.
  - "`l`", lognormal. This takes two additional arguments for the mean and variance of the distribution (on the log scale), in that order.
  - "`g`", gamma. This takes two additional arguments for the shape and scale parameters of the distribution, in that order.
  - "`e`", exponential. This takes one additional argument for the mean of the distribution.
  - "`x`", no latent periods. Hosts will be assumed to become infectious immediately.

  It is recommended that a latent period be included and that a strongly informative prior be placed on its length.

- `-s` or `--startingPTree` takes one argument which should be the path of a file containing a phylogeny in Newick format to be used as the starting phylogenetic tree.

- `-t` or `--startingTTree` takes one argument which should be the path of a file in CSV format which specifies the starting transmission tree; the first column should list each host (by the IDs specified in the compulsory CSV files) and the second its infector. The index host in this tree should have "Start" in the second column. The file is assumed to have a header line. Note that there is no guarantee that a specified starting transmission tree exists as a partition of a specified (or indeed random) starting phylogeny and if both `-s` and `-t` are given then BEAST may not start unless the combination is chosen carefully.

- `-f` or `--fixedPT` takes no further arguments and instructs BEAST to run on a fixed phylogeny (which should be specified with `-s`). The genetic data is irrelevant.

- `-g` or `--fixedTT` takes no further arguments and instructs BEAST to fix the transmission tree. Unless `-f` is also specified the MCMC will still propose new phylogenies from amongst those that will support the starting transmission tree. `-t` is essential and `-s` is recommended as a random starting phylogeny may not be compatible with the provided transmission tree.

- `-c` or `--chainLength` takes one argument that specifies a new MCMC chain length to replace the one given in the input XML file. If absent, the original length will be kept.

- `-e` or `--sampleEvery` takes one argument that specifies a new MCMC sampling frequency (for all log files, as well as the screen log) to replace the ones given in the input XML file. If absent, the original frequencies will be kept.
- `-x` or `--indexDatePrior` takes two arguments that specify the mean and standard deviation of the normal prior distribution of the date of the index infection. If this argument is not specified this is given an improper uniform prior. The first argument is the mean and should be given in the date format specified by `-d` or in days if that is absent. The second argument is the standard deviation and should be given in days.

A note on the within-host model. This script inserts the same within-host demographic model that was used in the Hall et al. paper; such that the effective populations size within each obeys the same logistic function. The priors on its parameters are also the same as in the paper. This can be modified by hand (see "Advanced XML modification"). As mentioned in the paper, however, peculiar behaviour may result if the prior distribution is not specified such that the effective population size at the time of transmission is considerably smaller than that at the end of infection.

4 Advanced XML modification

This section is designed for those who understand BEAST XML enough to perform modifications by hand, but please feel free to contact the author for help with making custom input files.

Changing distributions

The within-host population model, which appears in the `<demographicModel>` sub-element of `<caseToCaseTransmissionLikelihood>`, can be exchanged for another simple coalescent model be exchanging the `<logisticGrowthN0>` element for another. Possibilities are `<exponentialGrowth>`, `<logisticGrowth>` and `<expansion>`; `<constantSize>` will work but is not recommended as it implies no transmission bottleneck. See the full BEAST XML documentation for the required sub-elements of these models. Priors on the values of model parameters should be inserted in the `<prior>` block and MCMC proposals to change them in the `<operators>` block.

Categories

Instead of assuming a single prior distribution or hyperprior for the length of infectious or latent periods, the set of hosts can be subdivided into categories prior to the analysis, such that all hosts in a given category have the same prior distribution. This is done by providing separate `<infectiousPeriodPrior>` and `<latentPeriods>` elements as children to the `<categoryOutbreak>` element, whose sub-elements specify the prior distributions and have `id` strings. The `infectiousCategory` and `latentCategory` attributes of individual `categoryCase` elements map to the distribution IDs, specifying which category each host (case) belongs to.

Prior distribution for the identity of the index host

The output of ModifyXML assumes that we have no *a priori* knowledge about the most likely index host of the outbreak. However, the optional `indexPriorWeight` attribute of each `categoryCase` element allows the user to specify a prior weight. If absent, this value is assumed to be 1. The prior distribution will be a discrete distribution whereby each host has probability equal to its weight over the sum of all the weights. Weights of zero are acceptable, but may cause poor mixing and my instinct is that they should be used sparingly.