Package 'BORIS'

June 24, 2019

Type Package

Title Bayesian Outbreak Reconstruction Inference and Simulation

Version 1.0.0.9002

Date 2019-06-24

Author Simon Firestone <simon.firestone@unimelb.edu.au> with contributions from Max S.Y. Lau, Saritha Kodikara, Haydar Demirhan, Yoko Hayama, Takehisa Yamamoto, Toshiyuki Tsutsui, Gavin Gibson, Glenn Marion, George Streftaris, Lewis Stone, Mark Stevenson

Maintainer Simon Firestone <simon.firestone@unimelb.edu.au>

Description Bayesian reconstruction of outbreak transmission trees integrating epidemiological and genomic data.

License GPL (>= 2) | file LICENSE

Imports ape, coda, grDevices, Rcpp (>= 1.0.0), utils

LinkingTo Rcpp, BH

SystemRequirements C++11

NeedsCompilation yes

LazyData true

RoxygenNote 6.1.1

Encoding UTF-8

Suggests knitr, rmarkdown

VignetteBuilder knitr

Depends R (>= 3.5.0)

Archs i386, x64

R topics documented:

BORIS *Bayesian Outbreak Reconstruction Inference and Simulation (BORIS)*

Description

Outbreak reconstruction using epidemiological and genomic data based on the Susceptible Exposed Infectious Recovered (SEIR) systematic Bayesian Markov Chain Monte Carlo (MCMC) algorithm described in Firestone et. al. (under peer review), which extends the model of Lau et. al. (2015). Infers the transmission network, genomes and unobserved parameters at all important time points in the outbreak. Includes functions for forwards simulation of outbreaks using a related algorithm, both for model verification and for Bayesian real-time predictions.

Details

Key function are compiled from C++ source code, using the Boost header library (requiring 'BH' and 'Rcpp') and work upon inputs from the inputPath and dump extensive outputs in the outputPath, and include:

- `sim` for simulating outbreak data
- `infer` for infering the transmission network and unobserved parameters.

For an example of usage see the package vignette.

To install from source, requires Rtools to be installed as described in the installation vignette and also here.

Author(s)

Simon M. Firestone <simon.firestone@unimelb.edu.au>, Max S.Y. Lau, Saritha Kodikara and Haydar Demirhan, with contributions from Glenn Marion, George Streftaris, Gavin Gibson, Yoko Hayama, Takehisa Yamamoto, Toshi Tsutsui, Lewis Stone and Mark A. Stevenson

Maintainer: Simon Firestone <simon.firestone@unimelb.edu.au>

References

Firestone, S.M., Hayama, Y., Lau, M.S., Yamamoto, T., Nishi, T., Bradhurst, R.A., Demirhan, H., Stevenson, M.A., Tsutsui, T., under peer review. Bayesian transmission network reconstruction for foot-and-mouth disease outbreaks incorporating farm-level covariates. Proceedings of the Royal Society B: Biological Sciences.

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

==========

Acknowledgements:

The research that led to this package development was supported by:

1. An Australian Research Council Discovery Early Career Researcher Award (DE160100477).

2. The Japanese Ministry of Agriculture, Forestry and Fisheries (Management Technologies for the Risk of Introduction of Livestock Infectious Diseases and Their Wildlife-borne Spread in Japan, FY2018-2022).

3. Heriot-Watt University and the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS).

4. A consultancy agreement undertaken for the New Zealand Ministry of Primary Industries.

5. The University of Melbourne's High Performance Computing system SPARTAN (Lafayette et al., 2016).

==========

See Also

infer sim

`inf1` *Example* `infected_source_current.csv` *output from* `infer()`.

Description

This data set provides an example of infected source output from the file `infected_source_current.csv` from a single run of `infer()`, seeded with 1. Used as an example dataset in the vignette.

Usage

```
data(inf1)
```

Format

A data frame with 10000 inferred observations (1 row per iteration) of the infected source for each of the 100 individuals (1 per column):

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

`inf2` *Example* `infected_source_current.csv` *output from* `infer()`.

Description

This data set provides an example of infected source output from the file `infected_source_current.csv` from a single run of `infer()`, seeded with 102. Used as an example dataset in the vignette.

Usage

```
data(inf1)
```

Format

A data frame with 10000 inferred observations (1 row per iteration) of the infected source for each of the 100 individuals (1 per column):

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

`infer` *Implement SEIR MCMC inference based on epidemiological and genomic data*

Description

`infer` implements the SEIR MCMC model to infer the transmission network and key epidemiological and phylogenetic parameters.

Usage

```
infer(covariates = NULL,
      moves.inputs = NULL,
      parsAux = NULL,
      keyInits = NULL,
      priors = NULL,
      scalingFactors = NULL,
      seed = NULL,
      accTable = NULL,
      t.sample = NULL,
      inputPath = NULL,
      outputPath = NULL,
      dnaPath = NULL,
      debug = NULL)
```

Arguments

covariates	A data.frame including farm covariate data in the format detailed in infer.epi.input.
moves.inputs	A data.frame including animal movement/contact-tracing data in the format detailed in infer.moves.input.
parsAux	A data.frame including the parameters for MCMC implementation in the format detailed in infer.param.aux.
keyInits	A data.frame including key initial values for parameters. For definition of elements see infer.param.key.
priors	A data.frame including prior settings for parameters. For definition of elements see infer.param.priors.
scalingFactors	A data.frame including scaling factors for paramaters, also known as operators or proposal distances. For definition of elements see infer.param.sf.
seed	An integer seed for for the random number generator.
accTable	A data.frame including known the sources of infection for each individual for simulated outbreaks, or 9999 for each unknown source.
t.sample	A data.frame of the timing of sampling that led to genomic data being available for each individual. If not sampled, then enter the unassigned_value.
inputPath	The path to the folder where the input files will be created.
outputPath	The path to the folder where the output files will be created.
dnaPath	The path to sequence data. This full path must point to a directory containing a single .fasta file.
debug	A logical that produces verbose outputs for debugging only if $=1$.

Details

infer implements the Susceptible Exposed Infectious Recovered (SEIR) systematic Bayesian Markov Chain Monte Carlo (MCMC) inference as described in Firestone et. al. (under peer review) which is an extension of that originally presented by Lau et. al. (2015), incorporating additional terms representing individual covariates (risk categories) in the epidemiological data.

The infer function calls compiled C++ source code and works upon inputs from the inputPath and dumps extensive outputs in the outputPath. NOTE: infer deletes all files in the inputPath and outputPath directories prior to simulation.

Value

This function outputs six files at `outputPath` that are explained below:

`con_seq_current.csv`
: The grand master sequence, Gm, output each `n_output_gm` iterations (one line per sequence) as a comma-separated file with the nucleotides A,G,T,C, represented by 1,2,3,4, respectively.

`infected_source_current.csv`
: The inferred sources of each individual output each `n_output_source` iterations (one line per iteration) as a comma-separated file.

`parameters_current.log`
: The inferred parameters at each step of the MCMC chain (one line per iteration), with the first 10 discarded so that it can be opened directly in the program Tracer.

`seqs_current.csv`
: The inferred sequence for each individual, output each `n_output_gm` iterations (one line per sequence), as a comma-separated file with the nucleotides A,G,T,C, represented by 1,2,3,4, respectively. The inferred sequences for the first individual can be found in the rows: `seq(1,n_iterations,by=n_output_gm)`.

`t_e_current.csv`
: Inferred timing of exposure for each infected individual (one line per iteration) as a comma-separated file.

`t_i_current.csv`
: Inferred timing of onset of infectiousness for each infected individual (one line per iteration) as a comma-separated file.

Author(s)

Simon Firestone, Max S.Y. Lau, Haydar Demirhan

Maintainer: Simon Firestone <simon.firestone@unimelb.edu.au>

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

Examples

```
data(infer.epi.input)
data(infer.moves.input)

n<-nrow(infer.epi.input)


 # The auxilliary parameters
 pars.aux <- data.frame('n' = 100,
                        'kernel_type' = 'power_law',
                        'coord_type' = 'longlat',
                        't_max' = 100,
```

```
                          'unassigned_time' =  9e+6,
                          'processes' = 1,
                          'n_seq' = 5,
                          'n_base' = 7667,
                          'n_iterations' = 1e5,
                          'n_frequ' = 10,
                          'n_output_source' = 1,
                          'n_output_gm' = 1000,
                          'n_cout' = 10,
                          'opt_latgamma' = 1,
                          'opt_k80' = 1,
                          'opt_betaij' = 1,
                          'opt_ti_update' = 1,
                          'opt_movt' = 0,
                          stringsAsFactors = FALSE)

    # Initial values for the key parameters
     para.key.inits <- data.frame('alpha' = 2e-4,
                                  'beta' = 0.1,
                                  'lat_mu' = 5,
                                  'lat_var' = 1,
                                  'c' = 10,
                                  'd' = 3,
                                  'k_1' = 3,
                                  'mu_1' = 3e-05,
                                  'mu_2' = 1e-06,
                                  'p_ber' = 0.2,
                                  'phi_inf1' = 1,
                                  'phi_inf2' = 1,
                                  'rho_susc1' = 1,
                                  'rho_susc2' = 1,
                                  'nu_inf' = 0.2,
                                  'tau_susc'= 0.1,
                                  'beta_m'= 0.5)

    # The prior information
    para.priors <- data.frame('t_range' = 7,
                              't_back' = 21,
                              't_bound_hi' = 10,
                              'rate_exp_prior' = 0.001,
                              'ind_n_base_part' = 0,
                              'n_base_part' = 1000,
                              'alpha_hi' = 0.1,
                              'beta_hi' = 50,
                              'mu_lat_hi' = 50,
                              'var_lat_lo' = 0.1,
                              'var_lat_hi' = 50,
                              'c_hi' = 100,
                              'd_hi' = 100,
                              'k_1_hi' = 100,
                              'mu_1_hi' = 0.1,
                              'mu_2_hi' = 0.1,
                              'p_ber_hi' = 1.0,
                              'phi_inf1_hi' = 500,
                              'phi_inf2_hi' = 500,
                              'rho_susc1_hi' = 500,
                              'rho_susc2_hi' = 500,
```

```
                          'nu_inf_lo' = 0,
                          'nu_inf_hi' = 1,
                          'tau_susc_lo' = 0,
                          'tau_susc_hi' = 1,
                          'beta_m_hi' = 5,
                          'trace_window' = 20)

  #scaling factors (also known as operators or proposal distances)
  para.sf <- data.frame('alpha_sf' = 0.001,
                        'beta_sf' = 0.5,
                        'lat_mu_sf' = 1.25,
                        'lat_var_sf' = 1.75,
                        'c_sf' = 1.25,
                        'd_sf' = 0.75,
                        'k_1_sf' = 1,
                        'mu_1_sf' = 2.5e-5,
                        'mu_2_sf' = 2.5e-6,
                        'p_ber_sf' = 0.02,
                        'phi_inf1_sf' = 1.75,
                        'phi_inf2_sf' = 1.5,
                        'rho_susc1_sf' = 1,
                        'rho_susc2_sf' = 1.25,
                        'nu_inf_sf' = 0.25,
                        'tau_susc_sf' = 0.25,
                        'beta_m_sf' = 1)



## Not run:
  infer.out<-infer(covariates = infer.epi.input,
                   moves.inputs = infer.moves.input,
                   parsAux = pars.aux,
                   keyInits = para.key.inits,
                   priors = para.priors,
                   scalingFactors = para.sf,
                   seed = 1,
                   accTable = data.frame(rep(9999, n)),
                   t.sample = infer.epi.input$t_s,
                   inputPath = "./inputs",
                   outputPath = "./outputs",
                   dnaPath = "./gen_inputs")

## End(Not run)

# See package vignette for a complete example.
```

`infer.epi.input` *Epidemiological (covariate) input data on farms in the population affected by the outbreak to be reconstructed.*

Description

This data set provides an example of the inputs on the population of farms housing individual animals at risk for which an outbreak will be reconstructed and parameters inferred, along with corresponding genomic data.

Usage

```
data(infer.epi.input)
```

Format

A data frame with 100 observations (1 per farm) on the following 11 variables:

k A unique identifier for each farm. Starts at zero given C++ indexing system.

coor_x x coordinate of the farm's centroid, e.g., Longitude in decimal degrees using WGS84 datum.

coor_y y coordinate of the farm's centroid, e.g., Latitude in decimal degrees using WGS84 datum.

t_o Day of onset of first clinical signs observed in this individual (or in animals on this farm), since some arbitrary origin (Day 0). Can be left as the unassigned value (default `9e06`), for uninfected farms or farms where clinical onset was not observed.

t_s Day of sampling that led to diagnosis.

t_r Day removed, recovered or culled, if known or the unassigned value (default `9e06`) if unknown.

ftype A 3-level categorical variable used to represent farm type, here, in terms of the predominant susceptible species held, i.e. ftype = 0 if predominant species is cattle, ftype = 1 if predominant species is pigs, ftype = 2 if predominant species is sheep/other. This variable is later converted into an indicator variable.

herdn Number of susceptible animals held on the farm at the start of the outbreak.

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

`infer.moves.input` *Movement (contract-tracing) input data between farms in the population affected by the outbreak to be reconstructed.*

Description

This data set provides an example of inputs on the movements of animals between farms housing individual animals at risk in which an outbreak will be reconstructed and parameters inferred, along with corresponding genomic data.

Usage

```
data(infer.moves.input)
```

Format

A data frame with 100 observations (1 per movement) on the following 3 variables:

from_k the unique identifier of the source farm for the animals moved. Starts at zero given C++ indexing system.

to_k the unique identifier of the destination farm for the animals moved. Starts at zero given C++ indexing system.

t_m day that the movement occurred.

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

---

infer.param.aux *Auxillary input parameters for configuring the MCMC used by the* infer *function.*

Description

This data set is an example of the additional parameter inputs that configure the Bayesian SEIR MCMC that reconstructs the outbreak and infers parameters and missing data.

Usage

```
data(infer.param.aux)
```

Format

A data frame with 1 observations of the following 18 variables:

n Number of individuals/farms, presently limited to 500 considering RAM requirements.

kernel_type Kernel type (string variable) that can take the values "power_law" / "exponential" / "cauchy" / "gaussian".

coord_type Coordinate system: decimal degrees (i.e. latitudes and longitudes) or cartesian (projected coordinate reference system) can take the values "longlat" / "cartesian"

t_max Upper limit of observation period.

unassigned_time An abritary extreme value e.g. 9e+10, to indicate that an event does not happen e.g. no infection.

processes Number of parallel processes (if run on a single computer $=1$; only set to ≥ 1 if on a SLURM cluster and submitted with a task array for parallel computation of multiple chains)

n_seq Initial number of sequences expected per indivual unit at risk. Defines memory allocated, defaults to 5, updated if more are needed.

n_base Number of bases (nucleotides) in length for the sequences to be simulated with the Kimura (1980) model. Presently limited to 10,000. For longer sequences, it is suggested to use partial sequences of SNPs only.

n_iterations Number of iterations to run the MCMC. Typically 100,000s to millions of iterations are required to reach convergence of multiple chains, which should be checked. An additional 10 iterations are included, so that the first 10 can be discarded from the output file `parameters_current.log` and it can then be opened directly in the program Tracer.

n_frequ Frequency of updating exposure times.

n_output_source Frequency of outputing updated sources for each known infected farm.

n_output_gm Frequency of outputting an update of the grand master sequence (Gm) and the inferred sequences of each individual.

n_cout Frequency of updating the console output.

opt_latgamma Implementation option: 1 = assume a Gamma distribution for the latent period (0 = assume a Gaussian distribution).

opt_k80 Implementation option: 1 = reformulated K80 DNA substitution model to match original 1980 paper (0 = original version in Lau et al. (2015) based on a secondary reference).

opt_betaij Implementation option: 1 = farm-level covariates incorporated into beta, i.e. β_{ij} Lau modified model from Firestone et al. (2019b) (0 = originally implemented model from Lau et al. (2015).

opt_ti_update Implementation option: 1 = update timing of inferred onset of infectivity (0 = in Lau original implementation based on simulated data, see Supporting Information in Lau et al. (2015)).

opt_movt Implementation option under development: 1 = animal movements/contact-tracing incorporated into likelihood estimation with inferred parameter β_m (0 = originally implemented model from Lau et al. (2015).

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

Examples

```
pars.aux <- data.frame('n' = 100,
                       'kernel_type' = 'power_law',
                       'coord_type' = 'longlat',
                       't_max' = 100,
                       'unassigned_time' =  9e+6,
                       'processes' = 1,
                       'n_seq' = 5,
                       'n_base' = 7667,
                       'n_iterations' = 1e5,
                       'n_frequ' = 10,
                       'n_output_source' = 1,
                       'n_output_gm' = 1000,
                       'n_cout' = 10,
```

```
'opt_latgamma' = 1,
'opt_k80' = 1,
'opt_betaij' = 1,
'opt_ti_update' = 1,
'opt_movt' = 0,
stringsAsFactors = FALSE)
```

`infer.param.key` *Initialising values for key parameters inferred with the* `infer` *function.*

Description

This data set is an example of the key initialising parameter inputs used in the Bayesian SEIR MCMC that reconstructs the outbreak and infers parameters and missing data.

Usage

```
data(infer.param.key)
```

Format

A data frame with 1 observation of the following 17 variables:

alpha Initialising value for the background (primary) transmission rate of infection

beta Initialising value for the secondary transmission rate

lat_mu Initialising value for the mean of the duration of the farm-level latent period.

lat_var Initialising value for the variance of the duration of the farm-level latent period.

c Initialising value for the scale parameter of Weibull distribution representing the mean infectious period.

d Initialising value for the shape parameter of Weibull distribution representing the mean infectious period. $=1.0$ for an Exponential distribution.

k_1 Initialising value for the spatial transmission kernel shape parameter.

mu_1 Initialising value for the rate of transition mutations based on Kimura's 2 parameter nucleotide substitution model (Kimura, 1980).

mu_2 Initialising value for the rate of transversion mutations based on Kimura's 2 parameter nucleotide substitution model (Kimura, 1980).

p_ber Initialising value for the probability that a nucleotide base of each of the primary (seeding) sequences has of differing from the base at the corresponding site in the sequence of the universal master sequence [for details see (Lau et al., 2015)].

phi_inf1 Initialising value for the multiplicative effect of predominant species on premises-level infectivity compared to a reference of ftype0 farms, for ftype1 farms.

phi_inf2 Initialising value for the multiplicative effect of predominant species on premises-level infectivity compared to a reference of ftype0 farms, for ftype2 farms.

rho_susc1 Initialising value for the multiplicative effect of predominant species on premises-level susceptibility compared to a reference of ftype0 farms, for ftype1 farms.

rho_susc2 Initialising value for the multiplicative effect of predominant species on premises-level susceptibility compared to a reference of ftype0 farms, for ftype2 farms.

nu_inf Initialising value for the effect (power) of number of animals on premises-level infectivity for farms.

tau_susc Initialising value for the effect (power) of number of animals on premises-level susceptibility for farms.

beta_m Initialising value for the secondary transmission rate by contact-related transmission/animal movement (under development).

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

Examples

```
para.key.inits <- data.frame('alpha' = 2e-4,
                             'beta' = 0.1,
                             'lat_mu' = 5,
                             'lat_var' = 1,
                             'c' = 10,
                             'd' = 3,
                             'k_1' = 3,
                             'mu_1' = 3e-05,
                             'mu_2' = 1e-06,
                             'p_ber' = 0.2,
                             'phi_inf1' = 1,
                             'phi_inf2' = 1,
                             'rho_susc1' = 1,
                             'rho_susc2' = 1,
                             'nu_inf' = 0.2,
                             'tau_susc'= 0.1,
                             'beta_m'= 0.5)
```

---

infer.param.priors *The prior information for parameters inferred with the* infer *function.*

Description

This data set is an example of the prior parameterisation used in the Bayesian SEIR MCMC that reconstructs the outbreak and infers parameters and missing data.

Usage

```
data(infer.param.priors)
```

Format

A data frame with 1 observation of the following 27 variables:

t_range Used for update of the time of onset of infectiousness (t_i) if opt_ti_update=1, proposed values of t_i range between time of onset of clinical signs (t_o) +− range (for details see paper Supporting Information in Lau et al., 2015).

t_back Maximum assumed length for the latent period.

t_bound_hi Used when proposing initial sources, t_i of infectee must be within t_bound_hi days of proposed t_i of j (essentially the upper bound of the generation interval).

rate_exp_prior The rate of exposure to be used as the mode of a vague prior.

ind_n_base_part Logical representing whether the genomic sequence data is partial (=1) or complete (=0).

n_base_part The partial sequence length, used if ind_n_base_part=1.

alpha_hi Lower/upper bound for prior of inferred parameter.

beta_hi Lower/upper bound for prior of inferred parameter.

mu_lat_hi Lower/upper bound for prior of inferred parameter.

var_lat_lo Lower/upper bound for prior of inferred parameter.

var_lat_hi Lower/upper bound for prior of inferred parameter.

c_hi Lower/upper bound for prior of inferred parameter.

d_hi Lower/upper bound for prior of inferred parameter.

k_1_hi Lower/upper bound for prior of inferred parameter.

mu_1_hi Lower/upper bound for prior of inferred parameter.

mu_2_hi Lower/upper bound for prior of inferred parameter.

p_ber_hi Lower/upper bound for prior of inferred parameter.

phi_inf1_hi Lower/upper bound for prior of inferred parameter.

phi_inf2_hi Lower/upper bound for prior of inferred parameter.

rho_susc1_hi Lower/upper bound for prior of inferred parameter.

rho_susc2_hi Lower/upper bound for prior of inferred parameter.

nu_inf_lo Lower/upper bound for prior of inferred parameter.

nu_inf_hi Lower/upper bound for prior of inferred parameter.

tau_susc_lo Lower/upper bound for prior of inferred parameter.

tau_susc_hi Lower/upper bound for prior of inferred parameter.

beta_m_hi Lower/upper bound for prior of inferred parameter.

trace_window Under development: The maximum time that a movement could have occurred earlier than exposure to be considered causally related to infection.

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

Examples

```
para.priors <- data.frame('t_range' = 7,
                          't_back' = 21,
                          't_bound_hi' = 10,
                          'rate_exp_prior' = 0.001,
                          'ind_n_base_part' = 0,
                          'n_base_part' = 1000,
                          'alpha_hi' = 0.1,
                          'beta_hi' = 50,
                          'mu_lat_hi' = 50,
                          'var_lat_lo' = 0.1,
                          'var_lat_hi' = 50,
                          'c_hi' = 100,
                          'd_hi' = 100,
                          'k_1_hi' = 100,
                          'mu_1_hi' = 0.1,
                          'mu_2_hi' = 0.1,
                          'p_ber_hi' = 1.0,
                          'phi_inf1_hi' = 500,
                          'phi_inf2_hi' = 500,
                          'rho_susc1_hi' = 500,
                          'rho_susc2_hi' = 500,
                          'nu_inf_lo' = 0,
                          'nu_inf_hi' = 1,
                          'tau_susc_lo' = 0,
                          'tau_susc_hi' = 1,
                          'beta_m_hi' = 5,
                          'trace_window' = 20)
```

`infer.param.sf` *Scaling factors for key parameters inferred with the* `infer` *function.*

Description

This data set is an example of the scaling factor inputs used in the Bayesian SEIR MCMC that reconstructs the outbreak and infers parameters and missing data.

Usage

```
data(infer.param.sf)
```

Format

A data frame with 1 observation of the following 17 variables:

alpha_sf Scaling factor for the proposal (i.e. operator) for the background (primary) transmission rate of infection

beta_sf Scaling factor for the proposal (i.e. operator) for the secondary transmission rate

lat_mu_sf Scaling factor for the proposal for the mean of the duration of the farm-level latent period.

lat_var_sf Scaling factor for the proposal for the variance of the duration of the farm-level latent period.

c_sf Scaling factor for the proposal for the scale parameter of Weibull distribution representing the mean infectious period.

d_sf Scaling factor for the proposal for the shape parameter of Weibull distribution representing the mean infectious period. $=1.0$ for an Exponential distribution.

k_1_sf Scaling factor for the proposal for the spatial transmission kernel shape parameter.

mu_1_sf Scaling factor for the proposal for the rate of transition mutations based on Kimura's 2 parameter nucleotide substitution model (Kimura, 1980).

mu_2_sf Scaling factor for the proposal for the rate of transversion mutations based on Kimura's 2 parameter nucleotide substitution model (Kimura, 1980).

p_ber_sf Scaling factor for the proposal for the probability that a nucleotide base of each of the primary (seeding) sequences has of differing from the base at the corresponding site in the sequence of the universal master sequence [for details see (Lau et al., 2015)].

phi_inf1_sf Scaling factor for the proposal for the multiplicative effect of predominant species on premises-level infectivity compared to a reference of ftype0 farms, for ftype1 farms.

phi_inf2_sf Scaling factor for the proposal for the multiplicative effect of predominant species on premises-level infectivity compared to a reference of ftype0 farms, for ftype2 farms.

rho_susc1_sf Scaling factor for the proposal for the multiplicative effect of predominant species on premises-level susceptibility compared to a reference of ftype0 farms, for ftype1 farms.

rho_susc2_sf Scaling factor for the proposal for the multiplicative effect of predominant species on premises-level susceptibility compared to a reference of ftype0 farms, for ftype2 farms.

nu_inf_sf Scaling factor for the proposal for the effect (power) of number of animals on premises-level infectivity for farms.

tau_susc_sf Scaling factor for the proposal for the effect (power) of number of animals on premises-level susceptibility for farms.

beta_m_sf Scaling factor for the proposal for the secondary transmission rate by contact-related transmission/animal movement (under development).

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

Examples

```
para.sf <- data.frame('alpha_sf' = 0.001,
                      'beta_sf' = 0.5,
                      'lat_mu_sf' = 1.25,
                      'lat_var_sf' = 1.75,
                      'c_sf' = 1.25,
                      'd_sf' = 0.75,
                      'k_1_sf' = 1,
                      'mu_1_sf' = 2.5e-5,
                      'mu_2_sf' = 2.5e-6,
                      'p_ber_sf' = 0.02,
                      'phi_inf1_sf' = 1.75,
                      'phi_inf2_sf' = 1.5,
```

```
                    'rho_susc1_sf' = 1,
                    'rho_susc2_sf' = 1.25,
                    'nu_inf_sf' = 0.25,
                    'tau_susc_sf' = 0.25,
                    'beta_m_sf' = 1)
```

`nt.seq1` *Example* `seqs_current.csv` *output from* `infer()`.

Description

This data set provides an example of inferred sequence output data from the file `seqs_current.csv` from a single run of `infer()`, seeded with 1. Used as an example dataset in the vignette.

Usage

```
data(nt.seq1)
```

Format

A data frame with inferred sequences, 1 sequence per row. See the vignette for details on how these sequences are collated across multiple iterations.

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

`nt.t1` *Example* `seqs_t_current.csv` *output from* `infer()`.

Description

This data set provides an example of inferred sequence timing output data from the file `seqs_t_current.csv` from a single run of `infer()`, seeded with 1. Used as an example dataset in the vignette.

Usage

```
data(nt.t1)
```

Format

A data frame with inferred sequence timing data. Each row contains 1 set of timings per individual, per iteration when sequences are outputted according to the input parameter `n_output_gm`. See the vignette for details on how these sequences are collated across multiple iterations.

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

`paras1` *Example* `parameters_current.log` *output from* `infer()`.

Description

This data set provides an example of parameter output from the file `parameters_current.log` from a single run of `infer()`, seeded with 1. Used as an example dataset in the vignette.

Usage

```
data(paras1)
```

Format

A data frame with 10000 inferred observations (1 per iteration) of the following 20 variables:

sample Iteration number

log_likelihood Log likelihood as estimated at this iteration

corr Number of individuals for whom the inferred source in this iteration is equal to that in the input argument `accTable` to the function `infer`

alpha Initialising value for the background (primary) transmission rate of infection

beta Initialising value for the secondary transmission rate

lat_mu Initialising value for the mean of the duration of the farm-level latent period.

lat_sd Initialising value for the SD of the duration of the farm-level latent period.

c Initialising value for the scale parameter of Weibull distribution representing the mean infectious period.

d Initialising value for the shape parameter of Weibull distribution representing the mean infectious period. $=1.0$ for an Exponential distribution.

k_1 Initialising value for the spatial transmission kernel shape parameter.

mu_1 Initialising value for the rate of transition mutations based on Kimura's 2 parameter nucleotide substitution model (Kimura, 1980).

mu_2 Initialising value for the rate of transversion mutations based on Kimura's 2 parameter nucleotide substitution model (Kimura, 1980).

p_ber Initialising value for the probability that a nucleotide base of each of the primary (seeding) sequences has of differing from the base at the corresponding site in the sequence of the universal master sequence [for details see (Lau et al., 2015)].

phi_inf1 Initialising value for the multiplicative effect of predominant species on premises-level infectivity compared to a reference of ftype0 farms, for ftype1 farms.

phi_inf2 Initialising value for the multiplicative effect of predominant species on premises-level infectivity compared to a reference of ftype0 farms, for ftype2 farms.

rho_susc1 Initialising value for the multiplicative effect of predominant species on premises-level susceptibility compared to a reference of ftype0 farms, for ftype1 farms.

rho_susc2 Initialising value for the multiplicative effect of predominant species on premises-level susceptibility compared to a reference of ftype0 farms, for ftype2 farms.

nu_inf Initialising value for the effect (power) of number of animals on premises-level infectivity for farms.

tau_susc Initialising value for the effect (power) of number of animals on premises-level susceptibility for farms.

beta_m Initialising value for the secondary transmission rate by contact-related transmission/animal movement (under development).

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

`paras2` *Example* `parameters_current.log` *output from* `infer()`.

Description

This data set provides an example of parameter output from the file `parameters_current.log` from a single run of `infer()`, seeded with 102. Used as an example dataset in the vignette.

Usage

```
data(paras2)
```

Format

A data frame with 10000 inferred observations (1 per iteration) of the following 20 variables:

sample Iteration number

log_likelihood Log likelihood as estimated at this iteration

corr Number of individuals for whom the inferred source in this iteration is equal to that in the input argument `accTable` to the function `infer`

alpha Initialising value for the background (primary) transmission rate of infection

beta Initialising value for the secondary transmission rate

lat_mu Initialising value for the mean of the duration of the farm-level latent period.

lat_sd Initialising value for the SD of the duration of the farm-level latent period.

c Initialising value for the scale parameter of Weibull distribution representing the mean infectious period.

d Initialising value for the shape parameter of Weibull distribution representing the mean infectious period. =1.0 for an Exponential distribution.

k_1 Initialising value for the spatial transmission kernel shape parameter.

mu_1 Initialising value for the rate of transition mutations based on Kimura's 2 parameter nucleotide substitution model (Kimura, 1980).

mu_2 Initialising value for the rate of transversion mutations based on Kimura's 2 parameter nucleotide substitution model (Kimura, 1980).

p_ber Initialising value for the probability that a nucleotide base of each of the primary (seeding) sequences has of differing from the base at the corresponding site in the sequence of the universal master sequence [for details see (Lau et al., 2015)].

phi_inf1 Initialising value for the multiplicative effect of predominant species on premises-level infectivity compared to a reference of ftype0 farms, for ftype1 farms.

phi_inf2 Initialising value for the multiplicative effect of predominant species on premises-level infectivity compared to a reference of ftype0 farms, for ftype2 farms.

rho_susc1 Initialising value for the multiplicative effect of predominant species on premises-level susceptibility compared to a reference of ftype0 farms, for ftype1 farms.

rho_susc2 Initialising value for the multiplicative effect of predominant species on premises-level susceptibility compared to a reference of ftype0 farms, for ftype2 farms.

nu_inf Initialising value for the effect (power) of number of animals on premises-level infectivity for farms.

tau_susc Initialising value for the effect (power) of number of animals on premises-level susceptibility for farms.

beta_m Initialising value for the secondary transmission rate by contact-related transmission/animal movement (under development).

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

`sim` *Simulate an outbreak with genomic data*

Description

`sim` implements a simulation routine based on Selke's threshold method.

Usage

```
sim(epi.inputs = NULL,
    moves.inputs = NULL,
    parsKey = NULL,
    parsAux = NULL,
    inputPath = NULL,
    outputPath = NULL,
    debug = NULL)
```

Arguments

epi.inputs	A data.frame including farm covariate data in the format detailed in sim.epi.input.
moves.inputs	A data.frame including animal movement/contact-tracing data in the format detailed in sim.moves.input.
parsKey	A data.frame including key values for parameters that drive the simulation in the format detailed in sim.param.key.
parsAux	A data.frame including important auxillary parameters for the simulation in the format detailed in sim.param.aux.
inputPath	The path to the folder where the input files will be created.
outputPath	The path to the folder where the output files will be created.
debug	A logical that produces verbose outputs for debugging only if =1.

Details

sim simulates outbreak datasets with epidemiological and genetic data based on input data following the approach described in Lau (2015) based on Sellke thresholds (Sellke, 1983).

The sim function calls compiled C++ source code and works upon inputs from the inputPath and dumps extensive outputs in the outputPath. NOTE: sim deletes all files in the inputPath and outputPath directories prior to simulation.

See the respective input data examples for the format of required inputs: sim.epi.input, sim.moves.input, sim.param.key and sim.param.aux.

Value

This function outputs the following files at outputPath that are explained below:

con_seq_current.csv	The grand master sequence, Gm, from which all index sequences are derived. 1,2,3,4 represent nucleotides A,G,T,C, respectively.
epi_sim.csv	The simulated epidemiological dataset in the same format as the epi input data, now including simulated timings for exposure ("t_e"), onset of infectiousness ("t_i") and recovery/removal ("t_r") for each individual.
infected_source.txt	The simulated infectious source of each individual, based on it's C++ index, i.e. source 0 is the first individual in the dataset (k=0).
sampled_perct.txt	The proportion of infected individuals that were sampled.
subject_xx_nt.txt	Simulated nucleotide data for individual xx. If there are multiple transmission or sampling events for this individual, then each line represents the sequence at the time-point in subject_xx_t_nt.txt
subject_xx_t_nt.txt	Simulated time-points for sequences for individual xx, corresponding to the sequences stored in lines in subject_xx_nt.txt
t_sample.txt	Simulated time-points of sampling events for individual xx. If not sampled then = unassigned value (e.g., 9e06).

Author(s)

Simon Firestone, Max S.Y. Lau, Haydar Demirhan

Maintainer: Simon Firestone <simon.firestone@unimelb.edu.au>

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., 2015. Novel Bayesian inference in epidemics - model assessment and integrating epidemiological and genetic data. Department of Actuarial Mathematics & Statistics, School of Mathematical and Computer Sciences. Heriot-Watt University.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

Sellke, T., 1983. On the asymptotic distribution of the size of a stochastic epidemic. Journal of Applied Probability 20, 390-394.

Examples

```
data(sim.epi.input)
data(sim.moves.input)

# The key simulation parameters
para.key <- data.frame('alpha' = 4e-4,
                       'beta' = 0.2,
                       'mu_1' = 2e-05,
                       'mu_2' = 2e-06,
                       'a' = 3.0,
                       'b' = 2.5,
                       'c' = 21.0,
                       'd' = 4.0,
                       'k_1' = 1.7,
                       'p_ber' = 0.1,
                       'phi_inf1' = 3,
                       'phi_inf2' = 1.5,
                       'rho_susc1' = 0.4,
                       'rho_susc2' = 2,
                       'nu_inf' = 0.2,
                       'tau_susc'= 0.1,
                       'beta_m'= 1.0)

# The auxilliary parameters
pars.aux <- data.frame('n' = 100,
                       'seed' = 2468,
                       'n_base' = 7667,
                       'n_seq' = 5,
                       't_max' = 100,
                       'unassigned_time' =  9e+6,
                       'sample_range' = 10,
                       'partial_seq_out' = 0,
                       'n_base_part' = 1000,
                       'n_index' = 1,
                       'coord_type' = 'longlat',
```

```
                          'kernel_type' = 'power_law',
                          'latent_type' = 'gamma',
                          'opt_k80' = 1,
                          'opt_betaij' = 1,
                          'opt_movt' = 0,
                          'n_mov' = 60,
                          stringsAsFactors = FALSE)


  ## Not run:
    sim.out<-sim(epi.inputs = sim.epi.input,
                 moves.inputs = sim.moves.input,
                 parsKey = para.key,
                 parsAux = pars.aux,
                 inputPath = "./inputs",
                 outputPath = "./outputs")

## End(Not run)

# See package vignette for a complete example.
```

`sim.epi.input` *Epidemiological (covariate) input data on farms in the population to be simulated.*

Description

This data set provides an example of the inputs on the population of farms housing individual animals at risk in which an outbreak will be simulated, along with corresponding genomic data.

Usage

```
data(sim.epi.input)
```

Format

A data frame with 100 observations (1 per farm) on the following 11 variables:

k A unique identifier for each farm. Starts at zero given C++ indexing system.

coor_x x coordinate of the farm's centroid, e.g., Longitude in decimal degrees using WGS84 datum.

coor_y y coordinate of the farm's centroid, e.g., Latitude in decimal degrees using WGS84 datum.

t_e days of exposure after some arbitrary origin (Day 0). Can be set as 0 for the earliest source. Multiple sources are allowed. If left as the unassigned value (default `9e06`), will be simulated for individuals that are infected.

t_i days when infectious period commences. By definition this is unobserved, so given the unassigned value (default `9e06`) and simulated for individuals/farms that are infected and become infectious before `t_max`.

t_r days when infectious period ends, i.e. recovery or removal of all infectious animals on this farm (possibly through recovery, death or culling). If left as the unassigned value (default `9e06`), will be simulated for individuals that become infectious before `t_max`.

ftype0 Binary indicator variable for a 3-level categorical variable, used here as *the reference category*, to represent farm type in terms of the predominant susceptible species held, i.e. ftype0 == 1 if predominant species is cattle, otherwise ftype0 == 0.

ftype1 Binary indicator variable used here to represent farm type in terms of the predominant susceptible species held, i.e. ftype1 == 1 if predominant species is pigs, otherwise ftype1 == 0.

ftype2 Binary indicator variable used here to represent farm type in terms of the predominant susceptible species held, i.e. ftype1 == 2 if predominant species is sheep, otherwise ftype2 == 0.

herdn Number of susceptible animals held on the farm at the start of the outbreak.

status Status of the farm at the start of the period under observation: 1 == Susceptible, 2 == Exposed, 3 == Infectious, 4 == Recovered/Removed.

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

`sim.moves.input` *Movement (contract-tracing) input data between farms in the population to be simulated.*

Description

This data set provides an example of inputs on the movements of animals between farms housing individual animals at risk in which an outbreak will be simulated.

Usage

```
data(sim.moves.input)
```

Format

A data frame with 100 observations (1 per movement) on the following 3 variables:

from_k the unique identifier of the source farm for the animals moved. Starts at zero given C++ indexing system.

to_k the unique identifier of the destination farm for the animals moved. Starts at zero given C++ indexing system.

t_m day that the movement occurred.

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

`sim.param.aux` *Auxillary input parameters for simulating outbreaks with the* `sim` *function.*

Description

This data set is an example of the additional parameter inputs that shape the outbreak that will be simulated, along with corresponding genomic data.

Usage

```
data(sim.param.aux)
```

Format

A data frame with 1 observations of the following 17 variables:

n number of individual farms at risk

seed Seed for the random number generator to enable simulation runs to be reproducible.

n_base Number of bases (nucleotides) in length for the sequences to be simulated with the Kimura (1980) model. Presently limited to 10,000. For longer sequences, it is suggested to simulate the partial sequence of SNPs only.

n_seq Initial number of sequences expected per indivual unit at risk. Defines memory allocated, updated if more are needed.

t_max Upper limit of observation period.

unassigned_time An abritary extreme value e.g. 9e+10, to indicate that an event does not happen e.g. no infection.

sample_range The maximum possible delay between infection and sampling. Realistic upper bounds for some infectious diseases with overt clinical signs in naive populations can be assumed to be between 14-21 days.

partial_seq_out Logical representing whether the genomic sequence data to be simulated is partial (=1) or complete (=0).

n_base_part The partial sequence length, used if partial_seq_out=1.

n_index The number of indexes to be simulated. For a transmission tree with multifurcation use $ge1$.

coord_type Coordinate system: decimal degrees (i.e. latitudes and longitudes) or cartesian (projected coordinate reference system) can take the values `"longlat"` / `"cartesian"`

kernel_type Kernel type (string variable) that can take the values `"power_law"` / `"exponential"` / `"cauchy"` / `"gaussian"`.

latent_type Distribution type used for simulating the latent period. Presently only implemented for "gamma"

opt_k80 Implementation option: 1 = reformulated K80 DNA substitution model to match original 1980 paper (0 = original version in Lau et al. (2015) based on a secondary reference).

opt_betaij Implementation option: 1 = farm-level covariates incorporated into beta, i.e. β_{ij} Lau modified model from Firestone et al. (2019) (0 = originally implemented model from Lau et al. (2015).

opt_movt Implementation option (under development): 1 = utilise animal movement/contact-tracing data in the simulation (0 = Lau et al. (2015) original implementation)

n_mov number of animal movements or contacts in the inputted data `sim.moves.input`

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

Examples

```
pars.aux <- data.frame('n' = 100,
                       'seed' = 2468,
                       'n_base' = 7667,
                       'n_seq' = 5,
                       't_max' = 100,
                       'unassigned_time' =  9e+6,
                       'sample_range' = 10,
                       'partial_seq_out' = 0,
                       'n_base_part' = 1000,
                       'n_index' = 1,
                       'coord_type' = 'longlat',
                       'kernel_type' = 'power_law',
                       'latent_type' = 'gamma',
                       'opt_k80' = 1,
                       'opt_betaij' = 1,
                       'opt_movt' = 0,
                       'n_mov' = 60,
                       stringsAsFactors = FALSE)
```

---

sim.param.key *Key input parameters for simulating outbreaks with the `sim` function.*

Description

This data set is an example of the key parameter inputs that shape the outbreak that will be simulated, along with corresponding genomic data.

Usage

```
data(sim.param.key)
```

Format

A data frame with 1 observation of the following 17 variables:

alpha The background (primary) transmission rate of infection

beta The secondary transmission rate

mu_1 The rate of transition mutations based on Kimura's 2 parameter nucleotide substitution model (Kimura, 1980).

mu_2 The rate of transversion mutations based on Kimura's 2 parameter nucleotide substitution model (Kimura, 1980).

a Shape parameter of Gamma distribution representing the latent period.

b Scale parameter of Gamma distribution representing the latent period.

c Scale parameter of Weibull distribution representing the mean infectious period.

d Shape parameter of Weibull distribution representing the mean infectious period. $=1.0$ for an Exponential distribution.

k_1 The spatial transmission kernel shape parameter.

p_ber The probability that a nucleotide base of each of the primary (seeding) sequences has of differing from the base at the corresponding site in the sequence of the universal master sequence [for details see (Lau et al., 2015)].

phi_inf1 The multiplicative effect of predominant species on premises-level infectivity compared to a reference of ftype0 farms, for ftype1 farms.

phi_inf2 The multiplicative effect of predominant species on premises-level infectivity compared to a reference of ftype0 farms, for ftype2 farms.

rho_susc1 The multiplicative effect of predominant species on premises-level susceptibility compared to a reference of ftype0 farms, for ftype1 farms.

rho_susc2 The multiplicative effect of predominant species on premises-level susceptibility compared to a reference of ftype0 farms, for ftype2 farms.

nu_inf The effect (power) of number of animals on premises-level infectivity for farms.

tau_susc The effect (power) of number of animals on premises-level susceptibility for farms.

beta_m the secondary transmission rate by contact-related transmission/animal movement (under development).

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

Examples

```
para.key <- data.frame('alpha' = 4e-4,
                       'beta' = 0.2,
                       'mu_1' = 2e-05,
                       'mu_2' = 2e-06,
                       'a' = 3.0,
                       'b' = 2.5,
```

```
'c' = 21.0,
'd' = 4.0,
'k_1' = 1.7,
'p_ber' = 0.1,
'phi_inf1' = 3,
'phi_inf2' = 1.5,
'rho_susc1' = 0.4,
'rho_susc2' = 2,
'nu_inf' = 0.2,
'tau_susc'= 0.1,
'beta_m'= 1.0)
```

`te1` *Example* `t_e_current.csv` *output from* `infer()`.

Description

This data set provides an example of inferred timings of exposure output from the file `t_e_current.csv` from a single run of `infer()`, seeded with 1. Used as an example dataset in the vignette.

Usage

```
data(te1)
```

Format

A data frame with 10000 inferred observations (1 row per iteration) of the inferred timing of exposure for each of the 100 individuals (1 per column):

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

`te2` *Example* `t_e_current.csv` *output from* `infer()`.

Description

This data set provides an example of inferred timings of exposure output from the file `t_e_current.csv` from a single run of `infer()`, seeded with 102. Used as an example dataset in the vignette.

Usage

```
data(te2)
```

Format

A data frame with 10000 inferred observations (1 row per iteration) of the inferred timing of exposure for each of the 100 individuals (1 per column):

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

---

ti1 *Example* `t_i_current.csv` *output from* `infer()`.

Description

This data set provides an example of inferred timings of onset of infectiousness output from the file `t_i_current.csv` from a single run of `infer()`, seeded with 1. Used as an example dataset in the vignette.

Usage

```
data(ti1)
```

Format

A data frame with 10000 inferred observations (1 row per iteration) of the inferred timing of onset of infectiousness for each of the 100 individuals (1 per column):

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

ti2 *Example* t_i_current.csv *output from* infer().

Description

This data set provides an example of inferred timings of onset of infectiousness output from the file t_i_current.csv from a single run of infer(), seeded with 102. Used as an example dataset in the vignette.

Usage

```
data(ti2)
```

Format

A data frame with 10000 inferred observations (1 row per iteration) of the inferred timing of onset of infectiousness for each of the 100 individuals (1 per column):

References

Firestone, S.M., Hayama, Y., Bradhurst, R.A., Yamamoto, T., Tsutsui, T., Stevenson, M.A., 2019. Reconstructing outbreaks: a methods comparison of transmission network models. Sci. Rep.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16, 111-120.

Lau, M.S., Marion, G., Streftaris, G., Gibson, G., 2015. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput. Biol. 11, e1004633.

Index