Bayesian Estimator of Selfing (BES)

Contents

1 Introduction

BES is a software package for estimating self-fertilization (selfing) rates and other mating system parameters from genotype data. BES estimates parameters in a Bayesian framework using Markov chain Monte Carlo (MCMC). BES contains models of pure hermaphroditism, androdioecy (hermaphrodites + males), and gynodioecy (hermaphrodites + females). Under each model, BES estimates selfing rates, mutation rates, and mating-system specific parameters. BES also contains a generic model for estimating selfing rates and mutation rates independent of a mating system. Additional non-genetic information, such as field observations of the number of females or males, is required for estimating parameters under the gynodioecious model and the androdioecious model.

BES is run as a Unix command line program. It is not a GUI program; instead you must run it in a terminal. Therefore, you might want to keep a Unix Tutorial or Unix cheat sheet handy while you work.

BES runs on Linux, Mac OS X, and Windows. BES is distributed as an extension package for the BAli-Phy inference framework. You might therefore wish to refer to the BAli-Phy Documentation as well.

BES contains a number of modules that correspond to different mating system models. Each model allows estimating a different set of parameters. The generic model and the pure hermaphrodite model without inbreeding depression can be run without modification to estimate the selfing rate and locus-specific mutation rates.

However, the gynodioecious model and the androdiecious model require additional information besides the genetic data, such as (for example) field observations on the fraction of hermaphrodites. Therefore, the user must edit these modules to add this information before attempting to run these models. This manual describes how to add information, but is not a substitute for understanding something about the structure of the model.

2 Installation

2.1 Installing BAli-Phy

Since BES is an extension package for BALi-Phy, you must first install BAli-Phy before you can use BES. To install BAli-Phy, follow the installation instructions for BAli-Phy.

2.2 Installing additional software

You should also install the following software:

- Tracer helps to visualize the results of MCMC runs.

2.3 Installing BES

First, check that the `bali-phy-pkg` command works:

```
% bali-phy-pkg help
```

Download and install the BES package:

```
% bali-phy-pkg install BES
```

Check that the package is installed:

```
% bali-phy-pkg packages
```

To see what modules were installed, run:

```
% bali-phy-pkg files BES
```

You can uninstall the package by running:

```
% bali-phy-pkg uninstall BES
```

Next, download some additional modules for particular mating systems. These files are not installed into the package directory because they must be manually modified before they are used.

- HermID.hs
- Andro.hs
- AndroID.hs
- Gyno.hs

Keep in mind that only the generic model and the pure hermaphrodite model without inbreeding depression can be used to run an analysis without any modification.

3 Running the program

3.1 Quick Start

First, check that the model loads correctly:

```
% bali-phy -M PopGen.Selfing.Generic --test --- Examples/outfile.001.70.001.phase
```

If that works, then run the MCMC using the generic model:

```
% bali-phy -M PopGen.Selfing.Generic --iter=1000 --- Examples/outfile.001.70.001.phase &
```

This should create a directory called `Generic-1/` (or `Generic-2/`, etc.) that contains the output files.

Second, load the output file `Generic-1/C1.log` using the GUI program Tracer. On Unix, you can run this from the command line as follows:

```
% tracer Generic-1/C1.log &
```

It is also possible to use a non-graphical program statreport to view the estimates of the selfing rate

```
% statreport --select="s*" Generic-1/C1.log
```

This can be useful when analyzing data in a terminal.

4 Input

Input files must be in PHASE format. Both alleles for each locus should be specified, with NA given to indicate missing data.

A PHASE file contains a 3-line header, followed by a single line for each observed individual. The header consists of

1. The number of individuals, on a line by itself.
2. The number of loci, on a line by itself.
3. A sequence of 'M's (for microsatellite) on a line by itself. The number of M's should equal the number of loci.

The line describing each individual should contain an individual name, followed by a list of integer allele names. The name and the numbers should be tab-delimited, and there should be twice the number of alleles as loci, since there are 2 alleles per locus. Integer allele names must be positive. The purpose of this is to avoid confusion, since 0 and negative numbers are often used to indicate missing data.

Here is a very small PHASE file as an illustration:

```
2
3
```

```
MMM
sample.1    23  23  2   1   NA  NA
sample.2    23  20  1   1   4   5
```

5 Output

5.1 Output directory

BAli-Phy creates a new directory to store its output files each time it is run. By default, the directory name is the name of the model file, with a number added to the end to make it unique. BAli-Phy first checks if there is already a directory called *file*-1/, and then moves on to *file*-2/, etc. until it find an unused directory name.

You can specify a different name to use instead of the model file name by using the `--name` option.

5.2 Output files

BAli-Phy writes the following output files inside the directory that it creates:

File name	Description
C1.out	General information: command line, start time, etc.
C1.err	May contain error messages.
C1.log	MCMC samples for different variables.

The `C1.log` file contains MCMC samples for the variables that are being estimated. For example, the variable `s*` indicates the fraction of uniparental individuals at the time of breeding.

Since each mating system has a unique set of parameters, the variables for each model will be described in the section for that model.

6 Analyzing output files

6.1 Using `statreport`

In order to determine estimates of parameters, you can use the program `statreport`:

```
% statreport C1.log
```

You can plot the posterior distribution for specific parameters by using the select option:

```
% statreport --select "s*" C1.log
```

Here quotes are necessary to make sure that the `*` is not interpreted by the command line shell, but is passed in to the `statreport` program unchanged.

You can also use the arguments `--mean`, `--mode`, and `--median` (the default) to exime different properties of the posterior distribution.

6.2 Using `tracer`

Tracer is a graphical program for exploring posterior distributions.

Load the `C1.log` file using `tracer`. On Unix, if `tracer` is in your `PATH`, you can do this by typing

```
% tracer C1.log &
```

7 Mating system models

The stochastic process that generates the genetic data is described in terms of mating system parameters Ψ and the effective scaled mutation rates Θ_l^* for each locus l. The set of variables Ψ is specific for each mating system.

Some quantites are of interest for each mating system

- s^* is the fraction of uniparental adults (selfing rate).
- R is the decrease in effective population size caused by the mating system.
- T_k is the number of generations of selfing in the immediate ancestry of individual k.
- Θ_l is the scaled mutation rate $4Nu$ for locus l.
- Θ_l^* is the *effective* scaled mutation rate for locus l.

Here, R is given by

- $R = \lim_{N\to\infty} \frac{N^*}{N}$

where N is the population size, and N^* is the effective population size defined by the rate of parent-sharing. The effective scaled mutation rate is

- $\Theta_l^* = \Theta_l \cdot (1 - s^*/2)/R$.

In general, s^* and R are composite parameters: they are determined from the basic mating system parameters Ψ.

It is important to note that the genetic data contain information directly only about s^* and Θ_l^*. Therefore, the basic mating system parameters Ψ may be unidentifiable on the basis of genetic data alone. This can be solved by introducing additional data in the form on field observations on, for example, the fraction of males or the fraction of females in the population. When such information is unavailable, the generic model should be used.

7.1 Generic model

The generic model is agnostic about the particular mating system, and simply reports s^*, T_k and Θ_l^* as estimated from the genetic data. Since the mating system is not known, R cannot be calculated. This means that it is not possible to calculate Θ_l from Θ_l^*.

The generic model characterizes the mating system simply in terms of the selfing rate s^*. Thus, $\Psi = \{s^*\}$, and so s^* is a basic parameter, instead of being calculated from other parameters.

The following variables are estimated, with the field names given:

Variable	Name	Description
s^*	s*	Fraction of uniparental adults (selfing rate).
T_k	t[k]	Number of generations of selfing for individual k.
Θ_l^*	theta*[l]	*Effective* scaled mutation rate for locus l.

This variant is run by specifying `-M PopGen.Selfing.Generic` on the command line.

7.2 Pure Hermaphrodite

In the pure hermaphrodite model, each individual contributes both male and female gametes to the gene pool. There are two variants of this model.

7.2.1 Variant I

The first variant has $\Psi = \{s^*\}$. This variant treats s^* as a basic parameter, and does not model inbreeding depression. This variant is similar to the generic model, except that R can be calculated because the mating system is known to be a pure hermaphrodite mating system. This allows Θ_l to be calculated also. However, note that R is always 1 for the pure hermaphrodite model.

The following variables are estimated, with the field names given:

Variable	Name	Description
s^*	s*	Fraction of uniparental adults (selfing rate).
T_k	t[k]	Number of generations of selfing for individual k.
R	R	Decrease in parent-sharing effective population size
Θ_l^*	theta*[l]	*Effective* scaled mutation rate for locus l.
Θ_l	theta[l]	Scaled mutation rate $4Nu$ for locus l.

This variant is run by specifying `-M PopGen.Selfing.Herm` on the command line.

7.2.2 Variant II

The second variant has $\Psi = \{\tilde{s}, \tau\}$. This variant treats s^* as a composite parameter.

The following variables are estimated, with the field names given:

Variable	Name	Description
$\tilde{s}$	s~	Fraction of uniparental seeds.
τ	tau	Relative viability of selfed seeds.
s^*	s*	Fraction of uniparental adults (selfing rate).
T_k	t[k]	Number of generations of selfing for individual k.
R	R	Decrease in parent-sharing effective population size
Θ_l^*	theta*[l]	*Effective* scaled mutation rate for locus l.
Θ_l	theta[l]	Scaled mutation rate $4Nu$ for locus l.

Here the user must modify `HermID.hs` to add additional information about $\tilde{s}$ or τ.

This variant is run by specifying `-m HermID.hs` on the command line.

7.3 Androdioecy

In the androdioecious model, the population consists of some fraction p_m of males, with the rest of the individuals being hermaphrodites. Hermaphrodites produce males gametes, but only fertilize their own eggs. There are two variants of this model.

7.3.1 Variant I

The first variant has $\Psi = \{s^*, p_m\}$. This variant treats s^* as a basic parameter, and does not model inbreeding depression.

The following variables are estimated, with the field names given:

Variable	Name	Description
s^*	s*	Fraction of uniparental adults (selfing rate).
p_m	p_m	The fraction of males
T_k	t[k]	Number of generations of selfing for individual k.

Variable	Name	Description
R	R	Decrease in parent-sharing effective population size
Θ_l^*	theta*[l]	*Effective* scaled mutation rate for locus l.
Θ_l	theta[l]	Scaled mutation rate $4Nu$ for locus l.

Here the user must modify `Andro.hs` to add additional information about p_m.

This variant is run by specifying `-m Andro.hs` on the command line.

7.3.2 Variant II

The second variant has $\Psi = \{\tilde{s}, \tau, p_m\}$. This variant treats s^* as a composite parameter.

The following variables are estimated, with the field names given:

Variable	Name	Description
$\tilde{s}$	s~	Fraction of uniparental zygotes.
τ	tau	Relative viability of selfed zygotes.
p_m	p_m	The fraction of males
s^*	s*	Fraction of uniparental adults (selfing rate).
T_k	t[k]	Number of generations of selfing for individual k.
R	R	Decrease in parent-sharing effective population size
Θ_l^*	theta*[l]	*Effective* scaled mutation rate for locus l.
Θ_l	theta[l]	Scaled mutation rate $4Nu$ for locus l.

Here the user must modify `AndroID.hs` to add additional information about p_m and also about $\tilde{s}$ or τ.

This variant is run by specifying `-m AndroID.hs` on the command line.

7.4 Gynodioecy

In the gynodioecious model, the population consist of some fraction p_f of females, with the rest of the individuals being hermaphrodites. Hermaphrodites produce pollen, and so can contribute male gametes to females, other hermaphrodites, and themselves.

In this model $\Psi = \{\tilde{s}, \tau, p_f, \sigma\}$. Here s^* is a composite parameter.

The following variables are estimated, with the field names given:

Variable	Name	Description
$\tilde{s}$	s~	Fraction of hermaphrodite seeds set by self-pollen.
τ	tau	Relative viability of selfed seeds.
p_f	p_f	The fraction of females
σ	sigma	Seed production rate of females relative to hermaphrodites.
s^*	s*	Fraction of uniparental adults (selfing rate).
T_k	t[k]	Number of generations of selfing for individual k.
R	R	Decrease in parent-sharing effective population size
H	H	Fraction of non-selfed individuals with a hermaphrodite seed-parent.
Θ_l^*	theta*[l]	*Effective* scaled mutation rate for locus l.
Θ_l	theta[l]	Scaled mutation rate $4Nu$ for locus l.

Here the user must modify `Gyno.hs` to add additional information about 3 out of the 4 components of Ψ.

This variant is run by specifying `-m Gyno.hs` on the command line.

8 Specifying additional information

When the mating system parameters Ψ contain more than one degree of freedom, the mating system parameters are not identifiable from genetic data alone. Therefore, the user must obtain a model-description module (`HermID.hs`, `Andro.hs`, `AndroID.hs`, or `Gyno.hs`) and modify this module to add additional information to make the mating system parameters identifiable. In general, if Ψ contains n variables, then additional information about $n-1$ of them must be incorporated.

8.1 Modifying model-description modules

BES model-description modules (such as `Andro.hs`) use a Haskell-like syntax. In this syntax,

1. Function application `f(x,y)` is written `f x y`.
2. Comments are introduced by `--`.
3. In a `do` block, an `observe` command introduces data. Commenting out the `observe` statements removes the data from the model file.

In the model-description template files (e.g. `HermID.hs`), the comments illustrate possible ways to introduce variables.

8.2 Methods for adding additional information

Additional information about a variable can be added in 3 ways.

1. Add observations that depends on that variable.
2. Fix the variable to a known constant value.
3. Place a subjective prior on the variable.

8.2.1 Introduce a variable with a prior and place observations on it.

```
tau <- random $ sample $ uniform 0.0 1.0
observe (binomial 20 tau) 10
```

8.2.2 Fix a variable to a known constant value.

If you know the value of a variable, you can fix it to a constant:

```
let tau = 1.0
```

If you have observations about a variable (e.g. p_m) then do not fix that variable to a constant. If you fix a variable to a constant, then that variable cannot be estimated since its value is already known.

8.2.3 Place a subjective prior on a variable

This approach doesn't actually make the parameter *identifiable*, since this approach affects only the prior, and not the likelihood.

```
tau <- sample $ beta 2.0 8.0
```

As a result, it is not possible to compare the posterior (with data) and the prior (without data) to assess the impact of the data. This approach is therefore not recommended.

9 Bayesian priors and posteriors

In Bayesian parlance, *prior* means "before the data", and *posterior* means "after the data". The Bayesian approach places prior distributions on parameters. As a result, parameters become random variables, and can have distributions. This differs from the maximum likelihood setting, where parameters are not random. However, the researcher must interpret prior and posterior distributions carefully in order to avoid drawing erroneous conclusions.

9.1 Uninformative priors

We take the "objective Bayesian" approach, and seek priors that have minimal influence on the analysis. In general, such priors have a broad range. This range should not be characterized simply in terms of the mean or median, but in terms of the Bayesian Credible Interval (BCI), and in terms of the shapes of the tails of the distribution.

Note that, although the goal is to obtain priors with minimal influence, such priors are not actually "uninformative". Specifically, a uniform prior is not "uninformative". The choice of an appropriate prior should be undertaken with care in order to avoid ruling out any plausible outcome *a priori*.

9.2 Comparing the posterior and the prior

The posterior distribution combines information the prior distribution and the likelihood. In order to determine if the shape of the posterior is being largely driven by the prior, or largely driven by the likelihood, one should compare the posterior and the prior distributions.

9.3 Priors on composite parameters

For models, such as the gynodioecious model, where s^* is a composite parameter, no prior distribution is placed on s^* directly. However, s^* still has a prior distribution that is the combined result of the priors on all the basic parameters that it is computed from.

In order to determine what the shape of the prior on s^* is, one can run the model without data. This can be done by commenting out the `observe` statements.