DFE-alpha: inference of the distribution of fitness effects of new mutations and the rate of adaptive molecular evolution

Peter D. Keightley

Institute of Evolutionary Biology, University of Edinburgh, Charlotte Auerbach Rd, Edinburgh EH9 3FL, UK

Documentation for version 2.15 9th February 2016

License and disclaimer

The software is provided under the terms of the GNU General Public License. The program is distributed free of charge without any warranty whatsoever. See the GNU General Public License http://www.gnu.org/licenses/gpl.html for details.

1. Introduction

DFE-alpha was initially written to estimate the distribution of fitness effects (DFE) of new deleterious mutations using within-species nucleotide polymorphism data. It was subsequently extended to estimate the rate of adaptive molecular evolution, and then to infer the rate and fitness effects of advantageous mutations. The DFE is estimated by maximum likelihood (ML) based on site frequency spectra (SFSs) for two sets of nucleotide sites, one set assumed to be subject to mutation, selection and genetic drift and the other set of sites assumed to be evolving neutrally. The DFE is assumed to be a gamma distribution with shape parameter beta (β), or a model in which all mutational effects are equal can be run. The mean mutational effect is estimated on a scale $N_e s$, where N_e is a measure of the recent effective population size and s is the selection strength acting on a new mutation.

The program to estimate the DFE is called est_dfe, and can be run in two modes, depending on whether the folded or unfolded SFS is analysed. It is more straightforward and the results are likely to be more robust from analysis of the folded SFS. If the folded spectra are analyzed, running the program est_alpha_omega then allows the proportion of adaptive substitutions (α) and the relative rate of adaptive substitution (ω_a, expressed relative to the neutral substitution rate) to be estimated by combining the parameter estimates of the DFE with between-species nucleotide divergence data. Under either the folded or unfolded SFS modes of operation, a simple model of recent demographic change is assumed, since demographic change can affect the selected and neutral SFSs in ways that resemble selection. A one-epoch unchanging population size, a two-epoch model with a change in population size from N_1 to N_2 t_2 generations in the past, or a three-epoch model with two step changes from N_1 to N_2 and from N_2 to N_3 t_2 and t_3 generations in the past, respectively, can be run. Note that inferring the parameters of the 3-epoch model usually takes several hours of processing time.

2. Citing DFE-alpha

If you use the programs and publish a paper, then please make an appropriate citation. If you use est_dfe to estimate the DFE using the folded SFS, then cite:

Keightley, P. D. and Eyre-Walker, A. (2007). Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177: 2251-2261.

If you use est_alpha_omega to estimate α and ω_a then cite:

Eyre-Walker, A. and Keightley, P. D. (2009). Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular Biology and Evolution* 26:

2097-2108.

If you use est_dfe to estimate the DFE and adaptive mutation rate and fitness effects using the unfolded SFS, then cite:

Schneider, A., Charlesworth, B., Eyre-Walker, A. and Keightley, P. D. (2011). A method for inferring the rate of occurrence and fitness effects of advantageous mutations. *Genetics* 189: 1427-1437.

3. Installation

The programs require the GNU Scientific Library (gsl). See http://www.gnu.org/software/gsl/ to download gsl and to find out how to install it on your system.

DFE-alpha comprises three programs:

est_dfe – estimates the DFE for deleterious mutations. If analysing the unfolded SFS, the rate and fitness effects of advantageous mutation are also estimated.

est_alpha_omega – estimates the proportion of adaptive substitutions (α) and the relative rate of adaptive substitution (ω_a) from the folded SFS.

prop_muts_in_s_ranges – estimates the proportions of deleterious mutations with fitness effects in different ranges of fitness effects on a scale $N_e s$.

Having downloaded the programs along with their Makefile, they are compiled by invoking the command:

```
$ make
```

The software is written in C to be compiled by gcc on a Linux system, and should also run on other Unix-based systems such as Mac OS X. It is unknown whether the software runs on Windows systems: you could try a Unix emulator.

4. Running the programs

All three programs will run if the object file name is typed (assuming that the current directory is in your path) without command line options, in which case the required program parameters are asked for one at a time. This may be useful for learning how to use the programs, but the programs are more effectively run with command line options:

```
$ est_dfe -c est_dfe_config_file.txt
```

```
$ est_alpha_omega -c est_alpha_omega_config_file.txt
```

```
$ prop_muts_in_s_ranges -c est_dfe_output_file -o output_file
```

When run with command line options, est_dfe and est_alpha_omega expect configuration (config) files, the name of which is provided after the -c flag. The program prop_muts_in_s_ranges expects as its input the main output file produced by est_dfe.

5. Data files, input and output files

The programs require data files for numerical integration that must be downloaded from the DFE-alpha website. If the programs are run without a config file (-c absent), then the path to the directory containing these data files must be supplied in a text file named directory_config.dat. Data files for the one- and two-epoch model are downloaded. If the three-epoch model is chosen, files will be dynamically created in the directory specified. This directory must be different from the two epoch data directory.

The files input and output by est_dfe are described in Table 1.

Table 1. Files input or output by est_dfe and est_alpha_omega. The default file names are assumed as defaults by the program if they are not specified in the corresponding program's configuration file.

Default file name	***Comment***	***Corresponding parameter in config file***
directory_config.dat	Contains a single line of text specifying the directory containing the data files for 1 and 2 epoch models.	data_path_1
directory_config_three_epoch.dat	Contains a single line of text specifying the directory containing the data files for 3 epoch model. (Must be different from data_path_1).	data_path_2
sfs.txt	Input file containing the site frequency spectra for analysis by est_dfe.	sfs_input_file
est_dfe.out	Main results file produced by est_dfe.	-
neut_egf.out	Neutral gene frequency vector file produced by est_dfe.	-
sel_egf.out	Selected gene frequency vector file produced by est_dfe.	-
est_alpha_omega.out	Results file produced by est_alpha_omega.	est_alpha_omega_results_file
divergence.txt	Input file for est_alpha_omega containing counts of nucleotide sites and differences.	divergence_file

5.1 Input file for est_dfe containing the SFS data

The single input file for est_dfe contains one or more pairs of SFSs for selected and neutral sites, and has the following format (comments in red):

```
No. SFSs with different numbers of alleles sampled ($m$) [integer on one line]
for ($i$ = 1 to $m$) {
      No. alleles sampled in SFS $i$ ($x_i$) [= no. elements in unfolded vector]
      Selected SFS vector [a list of 0..$x_i$ counts]
      Neutral SFS vector [a list of 0..$x_i$ counts]
}
```

With the exception of the SFS vectors, all data elements are single integers on separate lines.

- The SFS includes *all* sites and not just polymorphic sites. An SFS vector looks like this:
  # of sites with frequency $0/x_i$, # of sites with frequency $1/x_i$, ... # of sites with frequency x_i/ x_i
- The SFS is a list of space-separated numbers (integer or real number) on a single line.
- If folded SFSs are provided by the user, then their length must be the number of alleles sampled +1, and the upper half of the SFS vectors should be zeros.
- Even if the neutral or selected SFS only are analysed in a given run, both SFSs need to be provided in the file.

Example of input file with a single pair of SFSs with 11 alleles:

```
1
11
317781 1410 167 30 11 12 14 17 29 25 33 41
132554 3031 1411 1001 731 618 504 503 419 402 421 563
```

6. Running est_dfe with a config file

The config file contains a list of parameter names (strings), each followed by a single value (integer, real number or string) (Table 2). A line can be commented out by starting it with the "#" character.

Table 2. Parameters specified in est_dfe_alpha configuration file and their meanings.

Parameter name	***Possible value***	***Comments***	***Mandatory?***
data_path_1	string	Name of directory containing the data files for the 1 and 2 epoch models. If absent, default directory specified in directory_config.dat is used.	No
data_path_2	string	Name of directory containing the data files for the 3 epoch model. Must be different from data_path_1. If absent, default directory specified in directory_config_three_epoch.dat is used.	No
sfs_input_file	string	File containing the site frequency spectra read by the program for analysis. If absent, default file is read.	No
est_dfe_results_dir	string	Name of directory to which est_dfe results files are written. If absent, current directory is written to.	No
est_dfe_demography_results_file	string	File containing results of previous run of est_dfe. The demographic parameters are read in from this file by the program, and assumed as fixed. Applies only to the case of site_class = 1.	If site_class = 1.
fold	0, 1	Fold the SFS [1] or not [0].	Yes
epochs	1, 2, 3	Number of population size changes +1.	Yes
site_class	0, 1, 2	0 = analyse neutral SFS only, 1 = analyse selected SFS only, 2 = analyse neutral and selected SFSs simultaneously. Option 2 is not allowed if fold = 0 or epochs = 3.	Yes
search_n2	0, 1	Search for the best-fitting population size n2 in 2-epoch model.	No
n2	1-1000	Population size after first change in population size.	If epochs = 2 and search_n2 = 0
t2_variable	0, 1	t2 is variable [1] or not [0] in likelihood maximization.	If epochs = 2 or 3
t2	>=10	Duration of epoch after first population size change (an initial or fixed value)	If epochs = 2 or 3
mean_s_variable	0, 1	Mean effect of a deleterious mutation is variable [1] or not [0] in likelihood maximization.	If site_class = 1 or 2
mean_s	<0	Mean effect of a deleterious mutation (an initial or fixed value).	If site_class = 1 or 2
beta_variable	0, 1	Shape parameter of gamma distribution is variable [1] or not [0] in likelihood maximization.	If site_class = 1 or 2
beta	>0.05 or -99	Shape parameter of gamma distribution (an initial or fixed value); The value -99 specified the equal effects model (a fixed value).	If site_class = 1 or 2
p_additional	>=0, <1	Proportion of mutations in fixed class 1. Only applies if fold = 0.	If site_class = 1 and fold = 0
s_additional	>-1, <1	Effect of mutations in fixed class 1. Only applies if fold = 0.	If site_class = 1 and fold = 0

Paths to the data directories (see above) need to be provided following the parameters data_path_1 and data_path_2. The output files, including the main results file (est_dfe.out) are written to a directory specified following est_dfe_results_dir. This is created by the program if it does not exist, **or existing results files in it are overwritten** if it does exist.

The program runs in two distinct modes, depending on whether fold is selected. If fold = 1, then the SFS is folded, and there are no additional classes of adaptive (or deleterious) mutations modelled by est_dfe (the parameters p_additional and s_additional are ignored if present). If fold = 0, the unfolded SFS is analysed and a fraction p_additional of mutations with effects s_additional are fitted. Within a run, these are fixed values, and their ML estimates need to be found by a grid-search implemented by the user. It is possible, however, to include these as variable parameters and to fit more than one additional class of mutations by inputting parameters as prompted by the program (rather than via a config file).

A second key parameter that controls the way that the program runs is site_class, which can take three values.

- If site_class = 0 the neutral SFS only is analyzed to estimate demographic and mutation parameters.
- If site_class = 1 the selected SFS is analyzed assuming fixed values for the demographic and mutation parameters, generated by a previous run of est_dfe with site_class = 0. These fixed values are read from a file specified by est_dfe_demography_results_file.
- If site_class =2, then the selected and neutral SFSs are analysed together.

It is recommended that the program is run by analyzing the neutral SFS data with site_class = 0, then the selected SFS with site_class = 1. Examples of config files for a typical workflow, running with site_class = 0 first, followed by site_class = 1, are as follows:

Config file for analysis of folded neutral SFS, two epochs by est_dfe

```
data_path_1     /home/myhomedirectory/data/
data_path_2     /home/myhomedirectory/data-three-epoch/
sfs_input_file  sfs.txt
est_dfe_results_dir     results_dir_neut/
site_class 0
fold 1
epochs 2
search_n2 1
t2_variable 1
t2 50
```

Config file for analysis of folded selected SFS by est_dfe

```
data_path_1     /home/myhomedirectory/data/
data_path_2     /home/myhomedirectory/data-three-epoch/
sfs_input_file  sfs.txt
est_dfe_results_dir     results_dir_sel/
est_dfe_demography_results_file results_dir_neut/est_dfe.out
site_class 1
fold  1
epochs  2
mean_s_variable 1
mean_s -0.1
beta_variable 1
beta 0.5
```

7. Running est_alpha_omega with a config file

This program calculates the proportion of adaptive substitutions (α) and the rate of adaptive substitution expressed relative to the neutral substitution rate (ω_a). It uses the DFE parameters calculated by est_dfe from the folded SFS to estimate the fixation probability of a deleterious mutation, and from this the expected number of fixed selected sites in the absence of adaptation. The difference between observed number and this expected number estimates the number of adaptive substitutions. As for est_dfe, a data file directory data_path_1 needs to be defined. The input file containing the total numbers of sites and fixed sites is specified by divergence_file, which has the following format:

```
1 number_of_selected_sites number_of_selected_differences
0 number_of_neutral_sites number_of_neutral_differences
```

The config file parameters are listed in Table 3. The program requires the DFE parameter estimates from a previous run of est_dfe (under site_class = 1 or 2) to be present as est_dfe_results_file. In addition, if polymorphism is removed from divergence (advised), it requires files to be specified containing neutral and selected gene frequency vectors (neut_egf_file and sel_egf_file, respectively), generated by a run of est_dfe with site_class = 2 or by separate runs under site_class = 0 and 1, respectively. The method to remove polymorphism is described in Keightley, P. D. and Eyre-Walker, A. (2012). Estimating the rate of adaptive molecular evolution when the evolutionary divergence between species is small. *Journal of Molecular Evolution* 74: 61-68.

Table 3. Parameters specified in est_alpha_omega configuration file and their meanings

Parameter name	*Possible value*	*Comments*	*Mandatory?*
data_path_1	string	Path to the directory containing the data files used by the program. If absent, default directory specified in the file directory_config.dat is used.	No
divergence_file	string	Input file containing numbers of nucleotide differences and sites. If absent default file is used.	No
est_alpha_omega_results_file	string	Output file from the program. If absent default file is used.	No
est_dfe_results_file	string	Results file produced by est_dfe, which is read by est_alpha_omega. If absent, default file is used.	No
neut_egf_file	string	File containing the neutral gene frequency vector produced by est_dfe, which is read by est_alpha_omega. If absent, default file is used.	No
sel_egf_file	string	File containing the selected gene frequency vector produced by est_dfe, which is read by est_alpha_omega. If absent, default file is used.	No
do_jukes_cantor	0, 1	Carry out Jukes-Cantor correction [1] or not [0] when calculating nucleotide divergence.	Yes
remove_poly	0, 1	Remove polymorphism contributing to divergence [1] or not [0] when estimating alpha and omega_a.	Yes

Example config file for est_alpha_omega

```
data_path_1                /home/myhomedirectory/data/
divergence_file            divergence.txt
est_alpha_omega_results_file    est_alpha_omega.out
est_dfe_results_file       results_dir_sel/est_dfe.out
neut_egf_file              results_dir_neut/neut_egf.out
sel_egf_file               results_dir_sel/sel_egf.out
do_jukes_cantor            1
remove_poly                1
```

8. Running prop_muts_in_s_ranges

This program estimates the proportion of deleterious mutations with effects in four different ranges of fitness effects on a scale $N_e s$. It uses an estimate of the DFE parameters from a previous run of est_dfe (run with site_class = 1 or 2) to calculate these proportions. It is invoked by issuing the command of the following form:

```
$ prop_muts_in_s_ranges -c  results_dir_sel/est_dfe.out -o output_file
```

The DFE parameter estimates are read from the file following -c and the results are output to the file following -o. Default file names are built-in. The output file has the following format (comments in red):

```
for ($i$ = 1 to 4)
{
    lower_Nes_value$_i$   upper_Nes_value$_i$ proportion_of_mutations$_i$
}
```

The value `-99.000000` for `upper_Nes_value`$_4$ means infinity.