DoFE 4.0

Distribution of Fitness Effects Software

Adam Eyre-Walker
University of Sussex
November 8th 2016

Introduction

This software is designed to run a number of analyses aimed at estimating various components of the distribution of fitness effects. Currently the software calculates the proportion of substitutions fixed by adaptive mutation according to the methods of Fay, Wycoff and Wu (2001), Smith and Eyre-Walker (2002), Bierne and Eyre-Walker (2004) and Eyre-Walker and Keightley (2009). It also implements the method of Eyre-Walker, Woolfit and Phelps to estimate the distribution of fitness effects (DFE) of deleterious mutations and the methods to estimate the neutrality described in Stoletzki and Eyre-Walker (2011).

Methods

Fay, Wycoff and Wu (2001), Smith and Eyre-Walker (2002), Bierne and Eyre-Walker (2004), Stoletzki and Eyre-Walker (2011)

All these methods share a common data format that can be in one of two forms. The first line in both is reserved for a description, which appears as a header when the data is displayed. The data should then be arranged with each gene on a separate line with the data separated by tabs in the following order

Gene name Dn Ln(divergence) Pn Ln(polymorphism) Ds Ls(divergence) Ps Ls(polymorphism)

Or

Gene name Dn Pn Ds Ps

Where Dn and Pn are numbers of non-synonymous substitutions and polymorphisms respectively and Ln(divergence) and Ln(polymorphism) are the numbers of sites in each case. I wrote the software to allow different numbers of sites because the divergence and polymorphism alignments can sometimes have different numbers of sites. Ds and Ps are as above but for silent sites. See

example data file FWW01.txt. All methods are found under the *Adaptive* menu except the neutrality index methods of Stoletzki and Eyre-Walker (2011), which are found under the *Other* menu. Please note, there are no routines to calculate the Direction of Selection (DoS) statistic; this can be readily calculated using Excel.

Eyre-Walker, Woolfit and Phelps (2006)
The format is as follows. The first line is reserved for description. It is then followed by a gene on each line with each field separated by a tab:

Gene name number of chromosomes Ln(poly) Pn1 Pn2 ….Pnk
  Ls(poly) Ps1 Ps2 …Psk

where k = Floor(n/2) and n is the number of chromosomes sampled. Note that the number of sampled chromosomes must be the same for all genes. See example data file EWWP06.txt.

The method does not do the integrations required by this method but instead uses a look-up table; this needs to be loaded prior to the analysis (you are prompted for this). The lookup table needs to be generated by the program LookUpTableGenerator (see below).

The method gives 95% credibility intervals for the shape parameter, the mean strength of selection and the standard error associated with the proportion of mutations in each part of the DFE.

Eyre-Walker and Keightley (2009)
This method incorporates the method of Eyre-Walker, Woolfit and Phelps into a McDonald-Kreitman type estimation of the proportion of advantageous substitutions (Eyre-Walker and Keightley 2009)(note this is the second method presented in the appendix of this paper; if you want to run the main method go to http://homepages.ed.ac.uk/eang33/). The data format is as follows; the first line is reserved for description, then on successive lines we have for each gene (or if you have summed the data across genes, a single line):

Gene name number of chromosomes Ln(poly) Pn1 Pn2 ….Pnk
  Ls(poly) Ps1 Ps2 …Psk Ln(div) Dn Ls(div) Ds

See example data file EWK09.txt. As with the method above, this method does not do the integrations required by this method but instead uses a look-up table; this needs to be loaded prior to the analysis (you are prompted for this). The lookup table needs to be generated by the program LookUpTableGenerator (see

below). Please note the lookup tables have changed since release 2.1, so you need to compile these – my apologies. Both this method and EWWP use the same look-up tables.

The implementations of the EWWP06 and the EWK09 methods allow you to exclude singletons. The EWK09 method also allows you to estimate omegaA, the rate of adaptive evolution relative to the rate of synonymous substitution (Gossmann et al. 2010) and to take into account that some of the divergence that you observe between two randomly chosen sequences is actually polymorphism.

The method gives 95% credibility intervals for alpha, omegaA, the shape parameter and the mean strength of selection and the standard error associated with the proportion of mutations in each part of the DFE.

Both methods can be run over multiple datafiles automatically. If you separate each datafile by a line with a hash, the program will run an analysis of the first dataset, then the second…etc outputting the results to the screen. Note that it only prompts you for a lookup table for the first dataset, so all datasets must have the same number of chromosomes sampled. Each datafile should be proceeded by a descriptor. See EWK09 multifile.txt for example.

Gossmann, Woolfit and Eyre-Walker (2012)
I don't recommend the use of this routine; estimating the parameters by ML is a better approach, since the method of Gossmann et al is biased. Contact me directly about this.

Checking the chain has converged

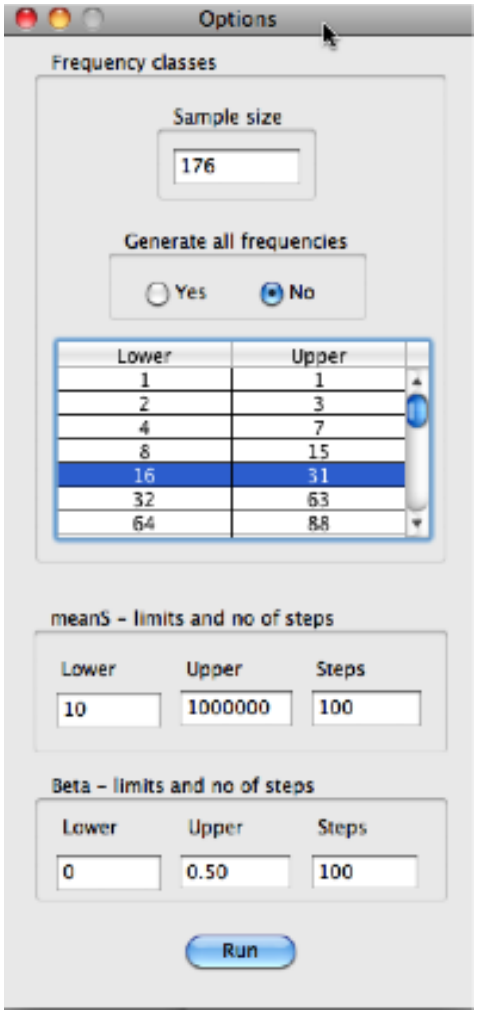
You should check that the chain converged for the EWWP, EWK and GMEW methods. This can be done visually, by saving the output from the MCMC (you are given this option at the end of the analysis) and inspecting the results in Excel or some similar program. Plot the results for each parameter against the step number and check

1. That there are no overall trends. None of the parameters should systematically increase or decrease. The parameter which has the strongest tendency to do this is the mean strength of selection; this behaviour can usually be corrected by choosing a higher starting value, or running the chain for longer.
2. That none of the parameters are close to their maximum or minimum values as set in the lookuptable. If they are then you need to recompile the lookup table and repeat the analysis. Again the mean strength of selection

can get very large and sometimes the shape parameter can get very small.

Using LookUpTableGenerator

Both The EWWP and EWK methods use lookup tables during their estimation procedure. These need to be generated before you can run your analysis using the program LookUpTableGenerator; I hope to integrate this into a single program in the near future. LookUpTableGenerator is fairly self-explanatory. If you have sampled a relatively small number of chromosomes then choose the "Generate all frequencies" option. However, if you have sampled hundreds of chromosomes then it might worth grouping frequency categories to increase the speed of the analysis. I recommend keeping singletons as a group then combining 2-3, 4-7, 8-15…etc as shown here:



NOTE, you will have to combine the folded site frequency spectrum in the same manner.

CAUTION: If you have huge amounts of data, particularly for the selected sites, you may have to recompile the lookuptable to restrict the analysis to a narrow range of values; this is because with large amounts of data the accuracy of the estimates becomes better than the accuracy of the lookuptable, which by its nature is discrete. As a precautionary measure analyses should be re-run if the confidence intervals on meanS are less than an order of magnitude apart.

Running an analysis

To run the analysis you need to load the data using the file menu (the welcome screen can be hidden by clicking on it or hitting any key). Then go to the adaptive, deleterious or other menus and select your chosen analysis. Most of this should be self explanatory but in the Bierne and Eyre-walker method, and in the Eyre-Walker, Woolfit and Phelps, and Eyre-Walker and Keightley methods you need to provide starting values. The method will suggest sensible starting values but these can be changed if the method does not converge (Bierne and EW) or takes too long for the chain to settle (EW, Woolfit and Phelps, and EW and Keightley)

Contact

I am happy to help with any of the analyses that can be performed by DoFE. Also please report any bugs to me and I welcome any comments about how to improve the software.

Adam Eyre-Walker (a.c.eyre-walker@sussex.ac.uk)

References

Bierne, N., and A. Eyre-Walker. 2004. Genomic rate of adaptive amino acid substitution in *Drosophila*. Mol. Biol. Evol. **21**:1350-1360.

Eyre-Walker, A., and P. D. Keightley. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol.

Eyre-Walker, A., M. Woolfit, and T. Phelps. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics **173**:891-900.

Fay, J., G. J. Wycoff, and C.-I. Wu. 2001. Positive and negative selection on the human genome. Genetics **158**:1227-1234.

Gossmann, T., B.-H. Song, A. J. Windsor, T. Mitchell-Olds, C. J. Dixon, M. V. Kapralov, D. A. Fialtov, and A. Eyre-Walker. 2010. Genome wide analyses reveal little evidence of adaptive evolution in many plant species. Mol. Biol. Evol. **27**, 1822-1832.

Gossmann, T., Woolfit, M. and Eyre-Walker, A. 2011 Quantifying the variation in the effective population size within a genome. *Genetics* **189**, 1389-1402.
Smith, N. G. C., and A. Eyre-Walker. 2002. Adaptive protein evolution in *Drosophila*. Nature **415**:1022-1024.