grapes : (Greetings_here_is_yet_another) **Rate of Adaptive Protein Evolution Software**

Introduction

grapes is a program estimating the adaptive and non-adaptive amino-acid substitution rates from polymorphism and divergence coding sequence data. It is written in C++ and based on the Bio++ libraries. This file is relevant to version 1.1.

grapes essentially re-implements Adam Eyre-Walker's **DoFE** program (Eyre-Walker and Keightley 2009) using a different optimization procedure and with a couple of additional features. Details of the methods are available from Galtier (2016). A similar approach and program have been developed by Tataru et al. (2017). See Rousselle et al. (2018) for a comparison of methods and clarification of their relationships.

Installation

The distributed binaries for GNU/Linux are statically linked and ready to run.

Compilation

If, for any reason, a source compilation is needed, here are the dependencies :

- gcc >= 4.4
- Bio++ >= 2.0 (bpp-core, bpp-phyl, bpp-seq)
- libgsl >= 1.0

The makefile should be adapted to fit with Bio++ path.

Execution

grapes -in input_file.dofe -out output_file.csv -model model_name [options]

Input

grapes needs information on the number of synonymous and non-synonymous substitutions between focal and outgroup species, number and frequency of synonymous and non-synonymous SNPs (SFS),

number of synonymous and non-synonymous sites in the divergence and polymorphism data sets. These are passed via a DoFE file as described in the documentation of the DoFE program: http://www.lifesci.susx.ac.uk/home/Adam_Eyre-Walker/Website/Software.html. Note that divergence data (the last four numbers) are optional.

grapes can handle both folded (as in **DoFE**) and unfolded SFS data. If SFS's are unfolded, this must be indicated via an extra line in the input file containing: `#unfolded`. See examples at the bottom of this file.

Analysis

grapes will estimate the distribution of fitness effect of mutations (DFE), rate of adaptive evolution (ω_a), rate of non-adaptive evolution (ω_{na}), and proportion of adaptive substitutions (α) by fitting a population genetic model to SFS + divergence data in the maximum likelihood framework, using the nuisance parameters r_i's introduced by Eyre-Walker et al. (2006). **grapes** will also perform more basic analyses, namely estimating α as $1 - [(\pi_N/\pi_S)/(d_N/d_S)]$, referred to as Neutral model, and using the corrected version of Fay, Wickoff and Wu (2002 Nature 415:1024), referred to as FWW.

Output

Basic output is written in the terminal. This corresponds to estimates of the adaptive and non-adaptive rates, main model parameters and likelihoods. Detailed output is written in a `csv` file.

Options

-model `GammaZero|GammaExpo|ScaledBeta|DisplGamma|FGMBesselK|all`

The most important option is the name of the assumed DFE model. Five distinct models are implemented in addition to the Neutral model, namely `GammaZero` (=Gamma), `GammaExpo`, `DisplGamma`, `ScaledBeta`, and `FGMBesselK`. See Galtier (2016) for details on what these models mean. One can either use one of the six models (e.g., `-model GammaExpo`), or do the six in a single run by passing `-model all`.

`-nearly_neutral <float>` : defines the threshold of $N_e.s$ above which a mutation is considered adaptive (S_{adv} in Galtier 2016, default=5)

`-FWW_threshold <float>` : minimal allele frequency in FWW alpha estimation (default=0.15)

`-no_div_data` : estimates DFE and calculates non-adaptive/adaptive rates only based on polymorphism data; divergence data, if any, are ignored (default=false)

`-no_div_param` : calculates non-adaptive/adaptive rates only based on the estimated DFE; will be set to false if `-model GammaZero` or `-no_div_data` is passed (default=false)

`-no_syn_orient_error` : force equal synonymous and non-synonymous mis-orientation rate (default=false)

`-anc_to_rec_Ne_ratio <float>` : ratio of ancient (divergence) to recent (polymorphism) effective population size; this will not alter DFE estimation, but modify calculation of the non-daptive and adaptive rates (default = 1.)

`-nb_rand_start <int>` : number of random starting values in model optimization (default=0); setting positive values will slow down the program but decrease the probability of being trapped in local optima.

`-fixed_param <control_file_name>` : this option should be used if one does not want to optimize every parameter, but rather have some parameters fixed to predefined values; parameter names and predefined values are passed via a control file; see example at the bottom of this file (default=none).

Example command lines

```
grapes -in infile.dofe -out outfile.csv -model GammaZero
```
→ "second approach" (with r_i's) in Eyre-Walker et al. 2009
→ "Gamma" in Galtier 2016 and Rousselle et al. 2018
→ "α_{div}, deleterious DFE" in Tataru et al. 2017

```
grapes -in infile.dofe -out outfile.csv -model GammaExpo -no_div_param
```
→ "α_{DFE}, full DFE" in Tataru et al. 2017
→ "GammaExpo, [-A]" in Galtier 2016
→ "GammaExpo*" in Rousselle et al. 2018

```
grapes -in infile.dofe -out outfile.csv -model GammaExpo -no_div_data
```
→ "α_{DFE}, full DFE, polymorphism data alone" in Tataru et al. 2017

Developed by

Nicolas Galtier
Institute of Evolutionary Sciences
CNRS – University Montpellier

Supported by

Montpellier Bioinformatics & Biodiversity platform
http://mbb.univ-montp2.fr

References

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

Eyre-Walker A, Keightley PD. 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.

Galtier N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. ***PLoS Genetics*** 12:e1005774.

Rousselle M., Mollion M., Nabholz B., Bataillon T., Galtier N. 2018. Overestimation of the adaptive substitution rate in fluctuating populations. ***Biology Letters*** 14:20180055.

Tataru P, Mollion M, Glémin S, Bataillon T. 2017 Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. ***Genetics*** 207:1103–1119.

Example input files:

- folded DoFE file:

```
Allolobophora_chlorotica+Aporerctodea_icterica (4852 genes)
all_genes   12    254286      4360.47     1987.61     1339.62     1640.54
      1087.89     581.969     92842.3     18049.3     9568.58     6722.38
      6820.24     5353.69     2607.35     340000      10578 111000      24890
```

First line is a header/comment line.
Second line contains the data:
entry 1 (`all_genes`): any string (dataset description)
entry 2 (`12`): sample size (here, 6 diploid individuals)
entry 3 (`254286`): number of non-synonymous sites, polymorphism data
entry 4 → 3+*n*/2, where *n* equals sample size: non-synonymous SFS (number of non-synonymous singletons, doubletons, etc...)
entry 3+*n*/2+1 (`92842.3`): number of synonymous sites, polymorphism data
entry 3+*n*/2+2 → 3+*n*/2+1+*n*/2, where *n* equals sample size: synonymous SFS (number of synonymous singletons, doubletons, etc...)
entry 3+*n*/2+1+*n*/2+1 (`340000`): number of non-synonymous sites, divergence data
entry 3+*n*/2+1+*n*/2+2 (`10578`): number of non-synonymous substitutions, divergence data
entry 3+*n*/2+1+*n*/2+3 (`111000`): number of synonymous sites, divergence data
entry 3+*n*/2+1+*n*/2+4 (`24890`): number of synonymous substitutions, divergence data

- unfolded DoFE file:

```
Microtus_arvalis+Microtus_glareolus.fas (2943 genes)
#unfolded
all_genes   12    1.61188e+06 2008.19     590.033     165.33      109.824     85.956
      73.6923     61.4066     63.5934     65.3077     104.637     119.604     361114
      5046.4      2095.86     836.505     533.066     403.44      357.615
      327.066     318.044     322.549     518.198     757.527     1.48228e+06 13089
      432574      38940
```

Same as above with additional `#unfolded` line, and SFS's containing *n*-1 entries instead of *n*/2.

Example control file:

```
GammaExpo,negGshape,0.4
GammaExpo,negGmean,10000
```

Each line contains model name, parameter name and fixed parameter value.

Parameter names are listed in the table below:

Model	Parameter name	Parameter meaning
GammaZero	negGmean	Gamma distribution mean, negative effects
GammaZero	negGshape	Gamma distribution shape, negative effects
GammaExpo	negGmean	Gamma distribution mean, negative effects
GammaExpo	negGshape	Gamma distribution shape, negative effects
GammaExpo	posGmean	Exponential distribution mean, positive effects
GammaExpo	pos_prop	Proportion of beneficial mutations (positive s)
DisplGamma	negGmean	Un-displaced Gamma distribution mean
DisplGamma	negGshape	Un-displaced Gamma distribution shape
DisplGamma	s0	Displacement
ScaledBeta	Ba	Beta distribution first shape parameter
ScaledBeta	Bb	Beta distribution second shape parameter
ScaledBeta	med_prop	Proportion of mutations with Ne.s above -25
FGMBesselK	BKm	*m* in equation 8, Lourenco et al 2011*
FGMBesselK	BKnsz2	n/z^2 in equation 8, Lourenco et al 2011*
FGMBesselK	BKsigma	σ in equation 8, Lourenco et al 2011*
FGMBesselK	BKscale	scaling parameter, FGM Bessel distribution

*Lourenco et al. 2011 Genetics 65:1559