multi_grapes : **a multi-SFS version of the grapes program**

Introduction

multi_grapes is a program fitting a population genetic model to synonymous + non-synonymous Site Frequency Spectrum (SFS) data, and is essentially a multi-SFS version of the program **grapes** (https://github.com/BioPP/grapes) . This file is relevant to version 1.1.

The main new feature of multi_grapes, compared to grapes, is the option of fitting model parameters assumed to be shared by distinct data sets – specifically, the shape of the distribution of fitness effects (DFE). This is the main focus of Galtier & Rousselle 2020, and the reason why this program was developed. In addition, new models of the DFE are implemented in multi_grapes that are absent in grapes.

These new option come with costs: (1) several of the models implemented in grapes are not available in multi_grapes; (2) no estimate of the adaptive substitution rate is performed by multi_grapes.

Installation

See the Grapes manual.

Compilation

See the Grapes manual.

Execution

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

Input

multi_grapes needs information on the number and frequency of synonymous and non-synonymous SNPs (SFS) and number of synonymous and non-synonymous sites, for one or several 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.

multi_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. Multiple data sets should

appear as different lines of the DoFE file (see examples at the bottom of this file)

Analysis

multi_grapes will estimate the distribution of fitness effect of mutations (DFE) by fitting a population genetic model to SFS + divergence data in the maximum likelihood framework. Data sets will first be analyzed separately (=with all parameters specific to each data set, as in grapes) then jointly (=with one parameter, the shape of the DFE, shared among data sets).

Output

Basic output is written in the terminal. This corresponds parameter estimates and likelihoods. Detailed output is written in a csv file.

Options

The most important option is the name of the assumed DFE model. Two distinct models are implemented in addition to the Neutral model, namely GammaZero (=Gamma) and ReflectedGamma. See Galtier (2016) and Galtier & Rousselle (2020) for details on what these models mean. In addition, one can add a class of lethal mutations using the -p_lethal option (see below).

Additional options:

-fold : fold the SFS (if unfolded)

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

-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.

-no_separate : do not perform separate analysis.

-no_shared : do not perform shared analysis.

-shared_shape <int>,<int>,...: fixed shared shape parameters (all the listed values will be fitted successively); in the absence of this option, the shared shape parameter will be optimized.

-p_lethal <float> : proportion of lethal mutations (this parameter cannot be optimized).

Developed by

Nicolas Galtier
Institute of Evolutionary Sciences
CNRS – University Montpellier

Supported by

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

References

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

Galtier N., Rousselle M. 2020. How much does Ne vary among species? **biorxiv** https://doi.org/10.1101/861849

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.

- multi-species DoFE file:

```
fourmis/    Wed,03      Apr2019     16:32:30+0200
#unfolded
Formica_cunicularia     8     400701      1202.23     300.657     109.941
      77.6465     65.5697     82.2242     80.5131     115533      1329.47
      570.259     258.343     164.596     130.121     147.059     130.028
      493898.9871 2212.76516522424  192183.30712      3136.55346427778
Formica_fusca     10    516075      1388.02     219.911     126.112     105.307
      88.7185     71.5122     65.3511     64.4724     84.6274     194430
      1655.77     458.135     300.317     231.845     192.99      150.888
      131.223     134.909     170.529     964535.9    244.7114    391115.4
      425.7194
Formica_pratensis 12    586753      979.616     283.847     140.758     98.7363
      69.4736     52.8  34.0835     29.3791     31.8214     34.1918     35.6714
      222142      865.268     373.901     213.662     163.245     128.126
      106.923     94.1758     76.3819     71.6703     76.6253     80.9967
      931598.3    513.256     377392.1    752.3783
Formica_sanguinea 16    613806      1860.31     435.443     221.748     148.063
      105.391     81.9975     75.9594     55.6502     47.0437     40.1243
      35.0322     25.6684     31.3399     36.6223     48.8784     232166
      2124.23     860.257     512.793     343.709     268.326     207.418
      179.786     147.136     126.418     122.138     107.387     94.7327
      81.0537     73.5026     107.048     942195.7    450.9158    382511.8
      660.5132
```