dfoil Documentation

Release 2017-06-14

James B. Pease

Jun 21, 2017

CONTENTS:

1 Getting Started **1**

1.1 What is dfoil? . . . 1
1.2 How do I cite this sofware? . . . 1
1.3 Requirements . . . 1
1.4 Installation . . . 2
1.5 Preparing your data . . . 2
1.6 Running dfoil.py . . . 3
1.7 Output Format . . . 4
1.8 Releases . . . 4
1.9 License . . . 5

2 Program Parameter Descriptions **7**

2.1 dfoil . . . 7
2.2 dfoil_analyze . . . 11
2.3 dfoil_sim . . . 12
2.4 fasta2dfoil . . . 15

3 Indices and tables **17**

CHAPTER

ONE

GETTING STARTED

1.1 What is dfoil?

$D_{\text{FOIL}}$is a method for testing introgression in a five-taxon symmetric phylogeny.

1.2 How do I cite this sofware?

If you use this program, please cite:

James B Pease, Matthew W. Hahn. 2015. "Detection and Polarization of Introgression in a Five-taxon Phylogeny" Systematic Biology. 64 (4): 651-662. http://www.dx.doi.org/10.1093/sysbio/syv023 doi: 10.1093/sysbio/syv023

Please also include the link <https://www.github.com/jbpease/dfoil> in your publication.

1.3 Requirements

- Python 2.7.x or 3.x
- Scipy: http://www.scipy.org/
- Numpy: http://www.numpy.org/
- matplotlib 1.5.3+: http://www.matplotlib.org/

1.3.1 Optional

To run simulations within dfoil_sim.py you will also need: * ms: http://home.uchicago.edu/rhudson1/source/mksamples.htm

1.4 Installation

No installation is necessary, just download dfoil and run scripts through Python. `git clone https://www.github.com/jbpease/dfoil` Other required software should be installed according to their individual instructions.

1.5 Preparing your data

1.5.1 Generating an AB-pattern count file from a FASTA file.

Use the script fasta2dfoil (see options below) to generate the count file from a FASTA.

1.5.2 Using a prepared count file:

One or more input files can be specified. These files can have any number of header lines at the beginning (including none), but **all header lines must start with '#'**

Data fields should be tab/space separated files with two starting fields:

- `CHROMOSOME` (this can be dummy values)
- `POSITION` (also can be dummy values)

followed by the pattern counts in this order:

`AAAAA AAABA AABAA AABBA ABAAA ABABA ABBAA ABBBA, BAAAA BAABA BABAA BABBA BBAAA BBABA BBBAA BBBBA`

for the four-taxon test the patterns are `AAAA AABA ABAA ABBA BAAA BABA BBAA BBBA`

Hint: these patterns are in 'binary' order 0000, 0010, 0100...

Important: The order of taxa must be P1 P2 P3 P4 O, such that: * "O" is the outgroup * P1 and P2 are a monophyletic pair of taxa * P3 and P4 are a monophyletic pair of taxa * P3 and P4 divergence occurs before (in forward time) the divergence of P1 and P2 (The choice of P1/P2 and P3/P4 within the pairings is arbitrary)

1.6 Running dfoil.py

1.6.1 Basic usage

`python dfoil.py --infile INPUTFILE1 [INPUTFILE2 ...] --out OUTPUTFILE1 [OUTPUTFILE2 ...]`

1.6.2 Pre-check

The data will undergo a precheck (use `--skip-pre-check` to turn off, or `--pre-check-only` to only run the pre-check. This will check for common issues in count data that might affect the result or violate the assumptions.

Common issues include: * An accelerated rate of substitutions (i.e. an excess of B's) on a specific branch relative to its sister taxon * Mis-labeling of P1/P2 and P3/P4 (remember P3/P4 divergence should come first, in forward time)

1.6.3 Modes

- dfoil = standard DFOIL for five-taxa (this is default)
- dfoilalt = dfoil without single-B patterns (ABAAA, etc.)
- partitioned = Partitioned D-statistics (Eaton & Ree 2013)
- dstat = four-taxon D-statistic (Green et al. 2010)
- dstatalt = four-taxon D-statistic with inverse patterns added (use with caution)

1.6.4 Advanced Weight Parameters

The parameters *–beta1*, *–beta2*, and *–beta3* are weighting factors (0 <= b <= 1), for single-B, double-B, and triple-B patterns, respectively. Ordinarily you will not need to set these.

By default these are set as:

- dfoil = 1,1,1
- dfoilalt = 0,1,1
- dstat: 0,1,N/A (no triple-B)
- dstatalt: 1,1,N/A (no triple-B)
- partitioned: N/A (does not use weighting parameters)

1.7 Output Format

One or more output files are specified (equal to number of inputs). The outputs will have fields:

- `CHROMOSOME`
- `POSITION`

then for each D-statistic:

- `Dxx_left` (left term value)
- `Dxx_right` (right term value)
- `Dxx_stat` (D-statistic value)
- `Dxx_chisq` (Chi_Squared value)
- `Dxx_pval` (Chi_Squared P-value)

(where "xx" will be replaced with FO, IL, FI, OL)

1.8 Releases

1.8.1 2017-06-14

Major upgrade to Sphinx documentation. Integrated the pre-check (formerly pre-dfoil.py) into the main script. Minor fixes to syntax.

1.8.2 2017-01-29

Fixes for visual graph outputs, *–plot_labels* has been fixed, replaced *colornoanc* and *colornoanc-dark* with *–plot_noanc* option, changes to color pallette for code comprehension. Replaced *–plot-path* with just *–plot*, and *–plot show* is now deprecated.

1.8.3 2015-11-23

More fixes for Python3 compability, added 'pre-dfoil.py' that checks for issues in count files before running dfoil.py

1.8.4 2015-04-17

Minor updates and citation information, Publication Release Version

1.8.5 2014-04-28

Fixes for Python3 compatibility

1.8.6 2014-02-07

Re-release version on GitHub

1.9 License

CHAPTER

TWO

PROGRAM PARAMETER DESCRIPTIONS

2.1 dfoil

2.1.1 Description

DFOIL: Directional introgression testing a five-taxon phylogeny dfoil - Calculate DFOIL and D-statistics stats from one or more count files. James B. Pease http://www.github.com/jbpease/dfoil

USAGE: dfoil.py INPUTFILE1 ... –out OUTPUTFILE1 ...

2.1.2 Parameters

`-h/--help`

Description: show this help message and exit

Type: boolean flag

`--infile` (required)

Description: input tab-separated counts file

Type: None; **Default:** None

`--out` (required)

Description: outputs tab-separated DFOIL stats

Type: None; **Default:** None

--beta1

Description: beta1 coefficient for single-B patterns, defaults: DFOIL/Dstatalt=1.0, DFOILalt,Dstat=0,Dpart=N.A.

Type: float; **Default:** None

--beta2

Description: beta2 coefficient for double-B patterns, defaults: Dpart=N.A., others=1.0

Type: float; **Default:** 1.0

--beta3

Description: beta3 coefficient for triple-B patterns defaults: DFOIL/DFOILalt=1.0, Dstat/Dpart=N.A.

Type: float; **Default:** 1.0

--mincount

Description: minium number of D denominator sites per window

Type: integer; **Default:** 10

--mintotal

Description: minimum total number of sites in a region

Type: integer; **Default:** 50

--mode

Description: dfoil = DFOIL, dfoilalt = DFOIL without single-B patterns, partitioned = Partitioned D-statistics, dstat = Four-Taxon D-statistic, dstatalt = Four-Taxon D-statistic with single-B patterns

Type: None; **Default:** dfoil

Choices: ['dfoil', 'dfoilalt', 'partitioned', 'dstat', 'dstatalt']

--plot

Description: write plot to file path(s) given

Type: None; **Default:** None

--plot_background

Description: 0-1.0 background intensity (0=none, default=0.3)

Type: float; **Default:** 0.3

--plot_color

Description: choose color mode

Type: None; **Default:** color

Choices: ['color', 'colordark', 'bw', 'bwdark']

--plot_height

Description: height of plot (in cm)

Type: float; **Default:** 8.0

--plot_hideaxes

Description: hide axes labels

Type: boolean flag

--plot_hidekey

Description: hide plot key

Type: boolean flag

--plot_labels

Description: taxon labels

Type: None; **Default:** None

--plot_lineweight

Description: line weight for dplots (default=1pt)

Type: float; **Default:** 1.0

--plot_noanc

Description: **do not plot background for** ancestral introgression

Type: boolean flag

--plot_smooth

Description: average D-stats over this number of points

Type: integer; **Default:** None

--plot_totals

Description: add a background plot of total site counts

Type: boolean flag

--plot_width

Description: width of plot (in cm)

Type: float; **Default:** 24.0

--plot_yscale

Description: Y-axis min-max value, default is 1

Type: float; **Default:** 1.0

--pre-check-only

Description: Only run the data pre-check (formely pre-dfoil.py)

Type: boolean flag

--pvalue

Description: minimum P-value cutoff for regions, can specify one P-value for all four tests or two separate ones for DFO/DIL and DFI/DOL (or D1/D2 and D12 for 'partitioned')

Type: float; **Default:** [0.01, 0.01]

--runlength

Description: if two introgressing windows are separated by this many windows of non-introgression color in the intervening windows to create a more continuous visual appearance

Type: integer; **Default:** 0

--skip-pre-check

Description: Skip running the data pre-check (formely pre-dfoil)

Type: boolean flag

--zerochar

Description: list of strings used in place of zeros in the input file default is [".", "NA"]

Type: None; **Default:** ['.', 'NA']

2.2 dfoil_analyze

2.2.1 Description

DFOIL: Directional introgression testing a five-taxon phylogeny dfoil_analyze: Given a dfoil output file, gives summary statistics to stdout James B. Pease http://www.github.com/jbpease/dfoil

2.2.2 Parameters

infile

Description: dfoil output file

Type: None; **Default:** None

-h/--help

Description: show this help message and exit

Type: boolean flag

--ndigits

Description: number of decimal places

Type: integer; **Default:** 3

2.3 dfoil_sim

2.3.1 Description

DFOIL: Directional introgression testing a five-taxon phylogeny dfoil_sim - simulation of sequences for testing dfoil James B. Pease http://www.github.com/jbpease/dfoil

2.3.2 Parameters

outputfile

Description: output site count filename

Type: file path; **Default:** None

-h/--help

Description: show this help message and exit

Type: boolean flag

--coaltimes

Description: coalescent times in 4Ne units

Type: float; **Default:** (3, 2, 1, 1)

--mdest

Description: 1-based index of migration recipient population

Type: integer; **Default:** None

--mrate

Description: per individual per generation migration rate (default=5e-4)

Type: float; **Default:** 0.0005

--msfile

Description: use pre-computed ms output file instead of running ms.

Type: None; **Default:** None

--msource

Description: 1-based index of migration source population

Type: integer; **Default:** None

--mspath

Description: path to ms executable

Type: None; **Default:** ms

--mtimes

Description: time bounds for the migration period

Type: float; **Default:** None

--mu

Description: per site per generation mutation rate (default=7e-9)

Type: float; **Default:** 7e-09

--nconverge

Description: number of convergent sites per window

Type: integer; **Default:** 0

--nloci

Description: number of windows to simulate

Type: integer; **Default:** 100

--popsize

Description: Ne, effective population size (default=1e6)

Type: integer; **Default:** 1000000.0

--quiet

Description: suppress screen output

Type: boolean flag

--recomb

Description: per site per generation recombination rate (default=0)

Type: float; **Default:** 0.0

--rho

Description: specific rho = 4*Ne*mu instead of using –recomb

Type: float; **Default:** None

--window

Description: length (bp) of windows

Type: integer; **Default:** 100000

2.4 fasta2dfoil

2.4.1 Description

DFOIL: Directional introgression testing a five-taxon phylogeny James B. Pease http://www.github.com/jbpease/dfoil

fasta2dfoil - This script takes one or more FASTA files containing 5 or 4 taxa and counts site patterns for use in DFOIL/Dstat analysis. To combine multiple FASTA files, each file should be sequences from one locus (i.e., one entry in the final table) and the names of sequences must be identical in all files.

2.4.2 Parameters

fastafile

Description: one or more input fasta files for each locus

Type: None; **Default:** None

`-h/--help`

Description: show this help message and exit

Type: boolean flag

`--names/-n` (required)

Description: Order of the 5 (or 4) taxa, names must be consistent in all input files, outgroup should be last

Type: None; **Default:** None

`--out/-o` (required)

Description: output count file, one entry per fasta

Type: None; **Default:** None

CHAPTER

THREE

INDICES AND TABLES

- genindex
- modindex
- search