sfs: Compute Site Frequency Spectra

Bernhard Haubold
Max-Planck-Institute for Evolutionary Biology, Plön, Germany

November 28, 2018

1 Introduction

For a sample of n haplotypes let $f_i(n)$ be the number of sites where i haplotypes carry a mutation. The vector

$$f_1(n), f_2(n), ..., f_{n-1}(n)$$

is called the *site frequency spectrum*. The package `sfs` contains two programs, `ms2sfs` and `bootSfs`, for dealing with site frequency spectra. `Ms2sfs` takes multiple haplotype samples simulated with `ms` (Hudson, 2002) as input and prints the corresponding site frequency spectra. `BootSfs` takes one or more site frequency spectra as input and bootstraps them.

2 Getting Started

`Ms2sfs` and `bootSfs` were written in C on a computer running Linux and should work on any standard UNIX system. However, please contact me at `haubold@evolbio.mpg.de` if you have any problems with the programs.

- Change into the package directory

  ```
  cd sfs
  ```

- Generate `ms2sfs` & `bootSfs`

  ```
  make
  ```

- Run the test scripts

  ```
  make test
  ```

  The executables `ms2sfs` and `bootSfs` are now located in the directory `build`. Place them into your PATH.

2.1 ms2sfs

- List the options

  ```
  ms2sfs -h
  ```

- Test it on a data set consisting of 2 simulated samples of five (odd) haplotypes.

  ```
  ms2sfs data/msOdd.dat
  ```

- Compute the folded site frequency spectrum

```
ms2sfs -f data/msOdd.dat
```

- Repeat for one sample of six (even) haplotypes

```
ms2sfs data/msEven.dat
ms2sfs -f data/msEven.dat
```

- Apply `ms2sfs` to 10^4 simulated haplotypes

```
ms 10 10000 -t 10 | ms2sfs | tail
```

where `ms` is the coalescent simulator by Hudson (2002).

2.2 `bootSfs`

- List options

```
bootSfs -h
```

- Read in one site frequency spectrum and bootstrap it twice:

```
bootSfs -i 2 data/test.sfs
```

- In case a particular run needs to be repeated exactly, set the seed for the random number generator:

```
bootSfs -i 2 -s 13 data/test.sfs
```

3 Change Log

- Version 0.1 (September 25, 2017)
  - First working version.
- Version 0.2 (October 23, 2017)
  - Polished interface.
- Version 0.3 (November 17, 2017)
  - Implemented folding of SFS (`-F`).
- Version 0.4 (November 29, 2017)
  - Enable analytic computation of SFS (`-T` to specify θ and `-n` to specify sample size.
- Version 0.5 (December 1, 2017)
  - Fixed error in folding.
- Version 0.6 (December 18, 2017)
  - Cleaned up interface.
- Version 0.7
  - Implemented `-r` for printing raw counts.
- June 13, 2018
  - Posted on `github`.

References

R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18: 337–338, 2002.