PedMix tutorial

Jingwen Pei
April, 2017

1 Introduction

PedMix is HMM (Hidden Markov Model) based framework designed to infer the admixture proportion of parents, grandparents and even great grand parents. Standard method inferring the ancestry of focal individual implicitly assume the same admixture proportion in both parents. However, this is unrealistic for many human populations, especially for the the recently admixed populations. *PedMix* consider a single diploid individual from an admixed population from two ancestral populations A and B. We assume phased genotype $G = (H_1, H_2)$ and allele frequency in both ancestral populations for each SNP are given. And also the admixed population is assumed to be admixture from A and B g generations ago, where g is known.

2 Prerequisite

g++ version later than 4.4.7 has been tested to compile *PedMix* successfully. We suggest to enable openmp in compilation to enhance the performance of *PedMix*.

To obtain Maximum Likelihood estimates of admixture proportions, we apply the Boyden-Fletcher-Goldfarb-Shanno (BFGS) method of optimization, using an implementation of the limited-memory version of the algorithm (L-BFGS) written in C. L-BFGS library needs to be installed to compile *Pedmix*. Please download the source code of L-BFGS package from:`https://github.com/chokkan/liblbfgs`, then compile with following commands:

```
$ ./configure
$ make
$ make install
```

PedMix requires the path where library and header are installed in Makefile. One can also install library and header under an alternative directory with specification of configure command. For more details of installation and method, please refer to website: `http://www.chokkan.org/software/liblbfgs`.

3 Download and Installation

The compiled version of *PedMix* is now available in: `https://github.com/pjweggy/PedMix`.

Source code will be later available under the same path. After installation of L-BFGS library, please change the path of $LBFGS_INC and $LBFGS_LIB to your installation path of header and library in Makefile (first two lines). Then run the following commands:

```
$ make clean
$ make
```

openmp is enabled with -fopenmp specified in Makefile.

4 Input Data format

The input of phased genotype data has the format of the following:

```
//
1 0.0000 0.5440 0.0002475900 1 0
1 0.7200 0.0000 0.0001201200 0 0
1 0.9500 0.0320 0.0003691200 0 0
1 0.9140 0.2980 0.0001483800 0 1
1 0.6620 0.0000 0.0021167700 0 0
1 0.5000 0.0000 0.0000291600 0 0
1 0.5000 0.0000 0.0000531000 0 0
1 0.6060 0.0000 0.0013107000 0 0
1 0.4280 1.0000 0.0000137400 1 1
1 0.4280 1.0000 0.0000927000 1 1
1 0.3660 0.9920 0.0001067100 1 1
```

The first column specifies the allele type that is used to count allele frequency. The second column specifies the allele frequency of such allele type in first ancestral population. The third column specifies the allele frequency in second ancestral population. The fourth column specifies the genetic distance (in centimorgan, cM) from current site to the next site. Thus, the last site should have distance 0.0, but it won't affect if greater than 0.0 since it will not be used. The last two columns specify two haplotypes (phased genotypes).

One can contain multiple locus in one input data. *PedMix* would consider multiple locus together to achieve maximum likelihood. Each loci should start with delimiter "//".

5 Usage

6 Tips

6.1 How to choose parameters

6.2 loci length

6.3 number of loci

6.4 phasing error

6.5 filtering strategy: LD pruning vs f-based pruning

6.6 Running Time

To make most use of computational resources, we suggest to use multiple threads for multi-locus data. It's efficient when the number of locus is the multiple of number of threads.

7 How to cite