Illustrations for evolBoosting

Kao Lin[1], Zongfeng Yang[1] and Haipeng Li[1] *

July 24, 2012

This document explains the principle of the R package "evolBoosting" and shows how to use it. This is a proper citation for this package:

K. Lin, H. Li, C. Schloetterer and A. Futschik(2011) Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. *Genetics* 187: 229-244.

1 Principle

The basic idea of this this package was described in Lin et al. 2011. However, some details might be different comparing to the method or data in Lin et al. 2011. The principle of this package is to use a machine learning method, boosting, to study the difference between selected population DNA samples and neutral samples. Then, boosting will produce a predictor via training(study) process. This predictor can then be used to predict new samples are whether selected or neutral.

The boosting method was described in Buehmann and Hothorn 2007. This package also depends on the package *mboost*, which is developed by Buehmann and Hothorn.

Before training or predicting, a set of summary statistics will be computed to represent the population DNA samples. These summary statistics include θ_w(Watterson 1975), θ_π(Tajima 1983), θ_h(Fay and Wu 2000), θ_l(Zeng et al. 2006), Tajima's D(Tajima 1989), Fay and Wu's H(Fay and Wu 2000, Zeng et al. 2006), MFDM(Li 2011), and iHH(Sabeti et al. 2002, Voight et al. 2006). Except iHH, all the statistics were computed in a sliding window, which slides from the left to the right of the sequence region. iHH will be computed in each window when 2 windows extends from the middle point to the two ends of the sequence region.

For example, if window size is 0.2 and sliding-step size is 0.1, all the summary statistics except iHH will be computed in each of the following regions: 0~0.2, 0.1~0.3, 0.2~0.4, 0.3~0.5, 0.4~0.6, 0.5~0.7, 0.6~0.8, 0.7~0.9, 0.8~1.0. iHH is integrated EHH, which will be integrated from the middle point to both left and right. In this example, iHH will be integrated from 0.5 to 0.4, 0.3, 0.2, 0.1 and 0 on the left and from 0.5 to 0.6, 0.7, 0.8, 0.9 and 1.0 on the left. Thus, we will have 9(sliding windows)*7(summary statistics except iHH) + 10(iHHs)=73 values of different summary statistics.

*Chinese-Academy-of-Sciences-Max-Plank-Gesellschaft Partner Institute for Computational Biology, 320 Yueyang Road, Shanghai, China. Contact: linkao@picb.ac.cn

In most of the cases, you need to simulate some population DNA samples before using this package. You can use *ms*(Hudson 2002) to simulate neutral samples, and use *msms*(Ewing and Hermisson 2010) to simulate neutral and selected samples.

The data processed by this package must be in *ms*/*msms* format or fasta format(Pearson and Lipman 1988).

2 Usage

Generally speaking, if you want to predict whether your sample(s) is(are) selected or neutral, four steps are needed.

1. Estimate some basic parameters(mutation rate, recombination rate, etc.) from your sample(s). This package doesn't provide this function, you can get the values from literatures or use another software, e.g., DnaSP(Librado and Rozas 2009) or MEGA(Tamura et al. 2011). Then use these parameters to simulate a set(e.g. 100) of neutral samples(using *ms* or *msms*) and selected samples(using *msms*).

2. Compute summary statistics from the simulated samples and your sample(s) using the function *statComputing*.

3. Combine the summary statistics from the simulated neutral and selected samples, and use the function *training* to get a predictor.

4. Use the predictor and the summary statistics from your sample(s) by the function *predicting* to make a prediction for your sample(s).

3 Example

A full example is shown below. In this example, all the samples contains 10 haplotype chromosomes with a population mutation rate = 0.005(per site) and a population recombination rate = 0.02(per site). The file "ms.out" contains 100 neutral samples. The file "msms.out" contains 100 recent strong selected samples(selective coefficient $2N_es = 500$, fixating time $= 0.001 * 4N_e$ generations ago). The samples in these two files are for training. The file "msms_for_test.out" contains 10 samples which have undergone not-very-recent weak seletion(selective coefficient $2N_es = 200$, fixating time $= 0.01 * 4N_e$ generations ago). The file "test.fasta" contains a fasta format sample with 1 outgroup sequence and 10 population sequences. "msms_for_test.out" and "test.fasta" are for testing.

The file of the fasta format must contain an outgroup sequence with the name "outgroup". Thus, if the sample size is n, the fasta file should contain $n + 1$ sequences in which the additional sequence is the outgroup sequence.

You can find these files in the sub-directory "extdata" in the package directory. The full path of the files can be achieved by using the R function $system.file$(See below).

```
> library("evolBoosting")
```

data(in principle you need use ms and/or msms to simulate training samples):

```
> ms <- system.file("extdata/ms.out", package = "evolBoosting")
> msms <- system.file("extdata/msms.out", package = "evolBoosting")
> msms_for_test <- system.file("extdata/msms_for_test.out", package = "evolBoosting")
> fasta <- system.file("extdata/test.fasta", package = "evolBoosting")
```

training(here ms contains neutral samples, msms contains selected samples):

```
> train1 <- statComputing(ms, 0.2, 0.2, y = 0)
> train2 <- statComputing(msms, 0.2, 0.2, y = 1)
> train <- rbind(train1, train2)
> predictor <- training(train)
```

predicting:

```
> test1 <- statComputing(msms_for_test, 0.2, 0.2)
> test2 <- statComputing(fasta, 0.2, 0.2)
> result1 <- predicting(predictor, test1)

[1] "The proportion of samples estimated as selection:"
[1] 0.6

> result2 <- predicting(predictor, test2)

[1] "The proportion of samples estimated as selection:"
[1] 1
```

show results(0 = neutral, 1 = selected):

```
> result1

      [,1]
 [1,]    0
 [2,]    1
 [3,]    1
 [4,]    1
 [5,]    1
 [6,]    1
 [7,]    0
 [8,]    0
 [9,]    0
[10,]    1

> result2

     [,1]
[1,]    1
```

Since "msms_for_test.out" contains samples with weak selective signal, thus four of them were predicted as neutral samples(0). The other 6 samples were predicted as selected samples(1). The sample in the fasta file was predicted as selected(1).

4 Reference

P. Buehmann and T. Hothorn(2007) Boosting algorithms: regularization, prediction, and model fitting. *Stat. Sci.* 22: 477-505.

G. Ewing and J. Hermisson(2010) MSMS: A coalescent simulation program including recombination, demographic structure, and selection at a single locus. *Bioinformatics* 26:2064-2065.

J.C. Fay and C.I. Wu(2000) Hitchhiking under positive Darwinian selection. *Genetics* 155:1405-1413.

R.R. Hudson(2002) Generating samples under a Wright-Fisher neutral model. *Bioinformatics* 18:337-338.

H. Li(2011) A new test for detecting recent positive selection that is free from the coufounding impacts of demography. *Mol Biol Evol* 28:365-375.

P. Librado and J. Rozas (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451-1452.

K. Lin, H. Li, C. Schloetterer and A. Futschik(2011) Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. *Genetics* 187: 229-244.

W.R. Pearson and D.J. Lipman(1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* 85:2444-2448.

P.C. Sabeti, D.E. Reich, J.M. Higgins, H.Z.P. Levine, D.J. Richter et al.(2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832-837.

F. Tajima(1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437-460.

F. Tajima(1989) Statistical method for testing the nuetral mutation hypothesis by DNA polymorphism. *Genetics* 123:585-595.

K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei and S. Kumar (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28:2731-2739.

B.F. Voight, S. Kudaravalli, X. Wen and J.K. Pritchard(2006) A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72.

G.A. Watterson(1975) On the number of segregating sites in genetical models without recombination. *Thero. Popul. Biol.* 7:256-276.

K. Zeng, S. Shi and C.I. Wu(2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174:1431-1439.