PhiPack: PHI test and other tests of recombination

Trevor Bruen

trevor@mcb.mcgill.ca

Contents

1 Introduction

The PhiPack software package implements (in C) a few tests for recombination and can produce refined incompatibility matrices as well. Specifically, PHIPack implements the 'Pairwise Homoplasy Index' ([1]), Maximum χ^2 ([4]) and the 'Neighbour Similarity Score' ([2]). The program Phi can be run to produce a p-value of recombination within a data set and the program profile can be run to determine regions exhibiting strongest evidence mosaicism.

2 Installation

Download the PhiPack.tar file from the webpage. Copy it into a directory of your choice (Mac usually Applications folder). Now open a terminal window (on a Mac OS X you can find this in Applications/Utilities). In the terminal window type

```
> tar -xvf PhiPack.tar
```

Next type (assuming the program is installed in the Applications folder)

```
> cd /Applications/PhiPack/src
> make
> cd ..
```

3 Running

After completing the Installation steps the program can be run by typing (assuming the program is installed in /Applications)

```
> cd /Applications/PhiPack/
> ./Phi -f noro.fasta
```

This gives a p-value of observing the noro.fasta sequences under the null hypothesis of no recombination calculated using the PHI statistic.

By typing:

```
> ./Phi
```

a list of options for running the program are presented.

4 Input format

The PHIPack program accepts two main types of files: PHYLIP and FASTA. The basic syntax is:

```
./Phi -s|-r|-f filename
```

One of the three options **-s, -r** and **-f** must be chosen.

4.0.1 PHYLIP (-s or -r)

The basic PHYLIP format accepted by the program is:

```
4 20
Seq_A       ACGGGGG....
...
Seq_D       ACGGGGG....
```

In this example, there are 4 sequences of of length 20. By using the **-s** option the sequence names can only be exactly 10 characters long, but using the **-r** option the sequence names can be very long but must be terminated by at least two spaces. For example the following is invalid for **-s** but valid for **-r**.

```
Seq_ABCDEFGHIJ  ACGTGG...
```

On the other hand the following is valid for **-s** but not valid for **-r**

```
Seq_567890ACGTGG...
```

Phylip files produced by PAUP* should be run with the **-s** option. Interleaved files can usually be read as well.

4.0.2 FASTA (-f)

The FASTA format accepted by the program is:

```
>Seq_A
ACGGGGG....
...
>Seq_D
ACGGGGG....
```

The name of the sequence should be followed by a newline and then the entire sequence.

5 Tests

By default the PHI Test is run on each alignment. Both Max χ^2 and NSS can be run using the **-o** option. More detailed output is available by typing **-v**. To change the number of permutations use **-p #**.

5.0.3 Phi (Φ_w)

The **-p** will cause the program to report the value of the Phi statistic under a direct permutation test as well as the normal alternative to the permutation test. Using the option **-w #** can change the window-size from the default ($w = 100$).

5.0.4 Max χ^2 and NSS

Both Max χ^2 and NSS use permutations tests to assess significance. A p-value estimated by a permutation test is reported for both of this statistics. For Max χ^2, a fixed window-size of 2/3 the number of polymorphic sites is used, which has been proposed previously [5, 6].

6 Profile

The program profile can be run to determine which regions exhibit the strongest evidence of mosaicism. This is particularly useful if only a few recombination events have occurred, but the alignment is somewhat long and has strong recurrent mutation (e.g. viral alignment of 10000 base pairs).

The idea is to calculate the Phi statistic along a sliding-window. The input format is the same as for the program Phi, but the option **-n** can be used for the scanning window size and **-m** can be used for the step size.

One output file is produced - Profile.csv. The first column contains the center of the window whereas the second column contains the p-value of recombination within that window. No multiple test correction is made thus significance of the observed statistic is likely an over-estimate.

7 Plots

Additionally, a refined incompatibility graphic matrix can be produced. This is similar to a compatibility matrix produced by Reticulate [3, 7, 2]. Sites that are compatible are yellow whereas the darkest sites will tend to be the most incompatible [1]. A graph "matrix.ppm" can be produced using the **-g** option. To convert this graph into a bmp type

```
>./ppma_2_bmp matrix.ppm matrix.bmp
```

A smaller graph can be produced with the **-g i** option.

8 SequenceTypes

The default option is nucleotide (DNA) sequences, but other sequence types can be accepted as well. The basic syntax is

```
./Phi ... -t [D|A|O]
```

The difference in all three modes is how letter are interpreted. The symbols '?' and '-' are always interpreted as missing data and gaps respectively but if amino acids are selected (with **-t A**), letters such as **N** will be interpreted as states instead of missing data (for DNA sequences).

8.0.5 Ambiguous/Missing Data

For both the Phi and NSS ambiguous/missing states are ignored, although the sites are not. For example if one site had **A, C, ?, G, C,A** the taxa with the '?' would be ignored for the purposes of calculating compatibility. For Max χ^2 all sites with ambiguous/missing taxa are removed from the alignment prior to calculation of the statistic.

References

[1] Trevor Bruen, H. Philippe, and D. Bryant. A simple and robust statistical test for detecting the presence of recombination. *Genetics*, 172:2665–2681, 2006.

[2] I B Jakobsen and S Easteal. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Comput Appl Biosci*, 12(4):291–5, 1996.

[3] Walter J. Le Quesne. A method of selection of characters in numerical taxonomy. *Systematic Zoology*, 18(2):201–205, 1969.

[4] J. Maynard Smith. Analyzing the mosaic structure of genes. *J Mol Evol*, 34(2):126–9, 1992.

[5] D Posada and K A Crandall. Evaluation of methods for detecting recombination from dna sequences: computer simulations. *Proc Natl Acad Sci U S A*, 98(24):13757–62, 2001.

[6] David Posada. Evaluation of methods for detecting recombination from dna sequences: empirical data. *Mol Biol Evol*, 19(5):708–17, 2002.

[7] P.H.A. Sneath, M.J. Sackin, and R.P. Ambler. Detecting evolutionary incompatibilities from protein sequences. *Systematic Zoology*, 24(3):311–332, 1975.