minmax-chisq: Obtaining a Reduced Amino Acid Alphabet with More Homogeneous Composition

Version 1.1

Edward Susko

Department of Mathematics and Statistics, Dalhousie University

Introduction

The program `minmax-chisq` obtains the minmax chi-squared bins for a sequence alignment. These are part of the methods described in Susko and Roger (2007); please cite this reference when using the software.

For a given number of bins, the routine chooses the bins that minimize the maximum chi-squared statistic:

$$t_s = n \sum_i (\pi_{is} - \pi_i)^2 / \pi_i \tag{1}$$

where n is the sequence length, π_{is} is the frequency of the ith bin for the sth species and π_i is the overall frequency of the ith bin. The p-value for a test of the hypothesis that the frequency for the sth species is the same as the overall frequency, can be calculated as $P(T > t_s)$ where T has a $\chi^2_{n_b-1}$ distribution. The same sequence that will give the maximum t_s will give the minimum p-value and so minimizing the maximum chi-squared statistic is equivalent to maximizing the corresponding minimum p-value.

Installation

The program `minmax-chisq` is compiled from C source code. To install `minmax-chisq`

1. Download and unpack the software:

   ```
   $ tar zxf minmax-chisq.tar.gz
   ```

   his will create a directory `minmax-chisq` and a test input file.

2. Change directories to `minmax-chisq` and create the main program file with the `make` command.

   ```
   $ cd minmax-chisq
   $ make
   ```

This should create the program file `minmax-chisq` which can be copied to a location in your `PATH`. To test the program, still in the `minmax-chisq` directory, issue the commands

```
$ ./minmax-chisq -s33 -n1000 -obinfile.check < nematode.phy
```

The output should be comparable to the output in `nematode-minmax.out`. The output file `binfile.check` should be comparable to `nematode-binfile.dat`.

The source code has been compiled and tested using gcc version 4.2.3 on an Ubuntu Linux distribution (Release 8.04). While the program has not been tested on another platform, it should compile under any Linux distribution as well as Mac OS X.

Usage

The program can be run at the command line with the command

```
$ minmax-chisq [options] < seqfile
```

Here `seqfile` is the name of the file containing the sequence data. The file should conform to PHYLIP standards for input with 10 character long names padded by blanks. The names should match the names used in the input treefile. Input can be either interleaved or sequential with one caveat: The lines 2 through $m + 2$, where m is the number of taxa, must contain the name of taxa followed by sequence data. For instance the start of a sequence file might be

```
6 3414
Homsa      ANLLLLIVPI LI...
Phovi      INIISLIIPI LL...
...
```

but not

```
6 3414
Homsa
ANLLLLIVPI LI...
Phovi
INIISLIIPI LL...
...
```

which would be allowed under the sequential format by PHYLIP.

The options are specified on the command line starting with '-'. For instance, the test example,

```
$ ./minmax-chisq -s33 -n1000 -obinfile.check < nematode.phy
```

indicates that the starting seed for random starting bins is 33, that 1000 choices of random starting bins should be tried and that the optimal bins should be output to `binfile.check`.

The options are

`-l number`. The first number of bins to consider. The default value is 2.

`-u number`. The last number of bins to consider. The default value is 20. Jointly, `-l` and `-u` indicate what numbers of bins to consider. For instance,

```
$ minmax-chisq -l3 -u3 < nematode.phy
```

obtains the minmax chi-squared bins when the number of bins is 3. the default is to obtain the minmax chi-squared bins for every possible choice of numbers of bins from 2 through 20.

`-o filename`. The file to output bins to. the default is to output to the screen, which should be sufficient for a single choice of bins. The output format has one line for each choice of the number of bins. The *i*th lines gives the optimal bins for the *i*th choice of the number of bins. For instance, the call

```
$ minmax-chisq -s33 -n1000 -l3 -u5 -obinfile.check < nematode.phy
```

indicates that the choices for the number of bins are 3,4 and 5. The second line of `binfile.check` contains the optimal bins when there are 4 bins.

Each line of the output file gives the optimal bins. These are indicated by a row of 20 integer codes, one for each amino acid, indicating the bin that amino acid is in. Amino acids are ordered alphabetically:

```
  alanine, arginine, asparagine, aspartic, cysteine,
  glutamine, glutamic, glycine, histidine, isoleucine,
  leucine, lysine, methionine, phenylalanine, proline,
  serine, threonine, tryptophan, tyrosine, valine
```

which is the same ordering used by most phylogenetic packages including PAML, PHYLIP and TREE-PUZZLE.

As an example, the call

```
$ minmax-chisq -s33 -n1000 -l3 -u5 -obinfile.check < nematode.phy
```

had, with the exception of the first line, the output file `binfile.check`,

```
A R N D C Q E G H I L K M F P S T W Y V
0 2 1 1 2 2 0 0 0 1 1 2 0 0 0 1 1 0 0 1
1 3 0 2 3 3 2 0 1 1 1 3 1 1 1 0 0 1 0 1
1 0 3 1 0 0 4 3 3 1 1 0 1 1 1 2 2 2 2 1
```

We see that the optimal solution for 4 bins, given in the second row, had the amino acids arginine (R), cysteine (C), glutamine (Q) and lysine (K) in the same bin, labeled 3.

-i filename. The file with starting bins. This should have one line for each choice of the number of bins. For instance, if the program was called with

```
$ minmax-chisq -s33 -n1000 -l3 -u5 -ibinfile.in < nematode.phy
```

The second line of the file binfile.in should have the starting bins when there are four bins. The format for these is the same as for the output bins. There should be twenty entries, the ith entry indicating the bin that the ith amino acid corresponds to. Bins should be labeled from 0 to $m-1$, where m is the number of bins for the line.

If an input file is not given the routine chooses starting bins randomly.

-n number. The number of random choices of bins to consider for each bin size. The default is 1000. The routine uses bin rearrangements, exchanging items from one bin to another, to improve upon an initial guess at the optimal number of bins. This strategy is not guaranteed to give the optimal choice and so it is valuable to try a number of different starting choices.

-s seed. An integer giving the starting seed for any random generation done by the program. By default this is 42.

Output to the Screen

In the case that an output file for the bins is specified, one line of output is given for each choice of the number of bins. The first number indicated is the number of bins and the second number is the smallest p-value, across species, for chi-squared tests of homogeneity using this choice of bins. For instance, the output from

```
$ ./minmax-chisq -s33 -n1000 -l2 -u4 -obinfile.check < nematode.phy
2 0.55297
3 0.44757
4 0.39179
```

indicates that the minimum p-values for 2,3 and 4 bins were all bigger than 0.35, implying that there is no indication of compositional heterogeneity for these choices of bins.

If no output file for bins is indicated, these will also be output to the screen. For instance, the previous example had the output

```
$ ./minmax-chisq -s33 -n1000 -l2 -u4 < nematode.phy
2 0.55297
0 0 1 0 0 1 1 0 1 0 0 1 0 0 1 0 0 1 1 1
3 0.44757
0 2 1 1 2 2 2 0 0 1 1 2 0 0 0 1 1 1 0 1
4 0.39179
1 2 0 3 2 2 3 0 1 1 1 0 1 1 1 0 0 2 2 1
```

Limitations

Very few reasons for error are output.

References

Susko, E. and Roger, A.J. (2007). On reduced amino acid alphabets for phylogenetic inference. Mol. Biol. Evol. 24:2139–2150.