Documentation for **Homo Version 1.4**

A program for analysing compositional heterogeneity across aligned sequence data

April 2019

Open Source Software License Agreement (variation of the BSD / MIT License)

Third Party Components

The following third-party components are distributed with the Software. You agree to comply with the license terms for these components as part of accessing the Software. Other third-party software may also be identified in separate files distributed with the Software.

Homo uses four functions (i.e., xGaussian_Distribution_Tail, xChi_Square_Distribution_Tail, Sum_Poisson_Terms, and Sum_Over_Odd_Terms) from two source code files (i.e., gaussian_distribution_tail.c, and chi-square_distribution_tail.c) written by Dr Richard L. Horn (athretprogr@gmail.com).

Dr Horn granted CSIRO and USYD a license to use these source code files without restrictions.

The functions provided by Dr Horn replace some of the functions made available to the public from his Mathematics Source Library C & ASM (http://www.mymathlib.com)

Credits

This software was developed by Lars S Jermiin while employed at CSIRO and USYD. Currently, LSJ is at Research School of Biology, The Australian National University, Acton, ACT 2601, Australia (lars.jermiin@anu.edu.au).

Introduction

Most model-based molecular phylogenetic methods assume evolution under stationary, reversible, and globally homogeneous conditions, implying that it would be unwise to use these methods if the data evolved under more complex conditions. **Homo** allows users to test whether pairs of sequences in alignments of nucleotides or amino acids are consistent with evolution under stationary, reversible, and globally homogeneous conditions (for details, see Ababneh et al. 2006b).

Homo conducts a matched-pairs test of symmetry (Bowker 1948, Ababneh et al. 2006a) on all pairs of sequences in the alignment. For each test, the probability (p) of obtaining the test statistic (X_S) by chance is returned. Assuming evolution under stationary, reversible, and globally homogeneous conditions, the distribution of p-values will be uniform, implying that 5% of the tests will produce p-values < 0.05 (Schweder and Spjotvoll 1982, Vera-Ruiz et al. 2014).

Homo also returns four compositional distances for each sequence pair. These distances are inferred from vectors with the relative frequencies of homologous characters in the alignment using the well-known Euclidean distance metric as well as a less well-known Aitchison distance metric (Egozcue and Pawlowsky-Glahn 2011). The compositional distances may be used to infer networks or trees, which visualize the compositional dissimilarities among the sequences.

The output is a:

- Brief summary (printed to the terminal),
- Full summary (printed to a file), and
- Five tables (printed to files), one with the p-values and four with the compositional distances.

The summaries allow users to determine whether some of the sequences violate the assumption of evolution under stationary, reversible, and globally homogeneous conditions.

The table with p-values allows users to produce a heat map, which may be used to identify subsets of sequences that are inconsistent with the assumption of evolution under stationary, reversible, and globally homogeneous conditions.

The tables with compositional distances allow users to visualize (e.g., using binary trees or networks) how the sequences cluster compositionally and how compositionally different the sequences are.

The matched-pairs test of symmetry

Consider an alignment of nucleotide or amino-acid sequences of length n. If the sequences contain nucleotides, then the alphabet contains $m = 4$ letters; on the other hand, if they contain amino acids, then $m = 20$. The alignment of any pair of these sequences can be displayed in a divergence matrix:

$$\mathbf{N}_{ij} = \{n_{ij}\} = \begin{bmatrix} n_{11} & \cdots & n_{1m} \\ \vdots & \ddots & \vdots \\ n_{m1} & \cdots & n_{mm} \end{bmatrix},$$

where n_{ij} denotes the number of homologous sites with character *i* in sequence 1 and character *j* in sequence 2 (note that $n = \sum n_{ij}$). The only difference between the alignment of the two sequences and $\mathbf{N}$ is that the order of the homologous sites is present in the alignment and absent in the matrix.

The test statistic for the matched-pairs test of symmetry is computed as follows:

$$X_S = \sum_{i<j} \frac{\left(n_{ij} - n_{ji}\right)^2}{n_{ij} + n_{ji}}. \qquad [1]$$

X_S is asymptotically distributed as a χ^2-variate on υ degrees of freedom, where υ is the number of comparisons done in each test. The number of comparisons done for a pair of sequences depends on the values of the off-diagonal elements of the divergence matrix (e.g., if $n_{ij} = n_{ji} = 0$, the specific comparison is ignored and the value of υ is reduced by 1).

The output of **Homo** includes X_S, υ, and p (i.e., $p = \Pr(X_S|\upsilon)$, that is, the probability of obtaining X_S by chance, given υ).

The compositional distances

Given a divergence matrix, it is possible to obtain two sets of vectors:

$$\begin{cases} \mathbf{Y} = \{y_k\} = (n_{12},\ \cdots,\ n_{1m},\ n_{23},\ \cdots,\ n_{2m},\ \cdots) \\ \mathbf{Z} = \{z_k\} = (n_{21},\ \cdots,\ n_{m1},\ n_{32},\ \cdots,\ n_{m2},\ \cdots) \end{cases},$$

$$\begin{cases} \mathbf{U} = \{u_k\} = (n_{1\bullet},\ \cdots,\ n_{m\bullet}) \\ \mathbf{V} = \{v_k\} = (n_{\bullet 1},\ \cdots,\ n_{\bullet m}) \end{cases}$$

where $\mathbf{Y}$ and $\mathbf{Z}$ are vectors containing the off-diagonal elements of the divergence matrix, and $\mathbf{U}$ and $\mathbf{V}$ are vectors containing the marginal frequencies of the divergence matrix. Here, • denotes the sum over the off-diagonal elements in rows (e.g., $\bullet 1$) or the sum over the off-diagonal elements in columns (e.g., $1\bullet$) (k is an index).

Given these vectors, it is possible to compute four metrics, each of which represents a compositional distance between matching vectors (i.e., $\mathbf{Y}$ versus $\mathbf{Z}$; $\mathbf{U}$ versus $\mathbf{V}$).

Our first metric is based on the Euclidean distance and is obtained using the following equation:

$$d_{EFS} = \sqrt{\sum_{k=1}^{l}\left(\frac{y_k}{n} - \frac{z_k}{n}\right)^2}$$

where l is the number of matching elements in $\mathbf{Y}$ and $\mathbf{Z}$. In this case, l = 6 for nucleotides and 190 for amino acids. Note that $0.0 \leq d_{EFS} \leq 1.0$.

Our second metric is also based on the Euclidean distance and is calculated as follows:

$$d_{EMS} = \frac{1}{\sqrt{2}}\sqrt{\sum_{k=1}^{l}\left(\frac{u_k}{n} - \frac{v_k}{n}\right)^2}$$

where l is the number of matching elements in $\mathbf{U}$ and $\mathbf{V}$. In this case, l = 4 for nucleotides and 20 for amino acids. Note that $0.0 \leq d_{EMS} \leq 1.0$.

Strictly speaking, the two metrics assume Euclidean space, where the data (i.e., the values in $\mathbf{Y}$, $\mathbf{Z}$, $\mathbf{U}$, and $\mathbf{V}$) represent absolute information. However, in our case the elements in $\mathbf{Y}$, $\mathbf{Z}$, $\mathbf{U}$, and $\mathbf{V}$ sum up to n, so the data represent relative information, rather than absolute information, so they are called compositional data. Data of this type lend themselves to compositional data analysis (Aitchison 1986, Egozcue and Pawlowsky-Glahn 2011), where the focus is on the individual proportions and the *ratios* between these proportions. For compositional data, Aitchison's distance is especially useful because it measures the distance between two points within a simplex (Egozcue and Pawlowsky-Glahn 2011). However, when calculating Aitchison's distance, the elements in $\mathbf{Y}$, $\mathbf{Z}$, $\mathbf{U}$, and $\mathbf{V}$ must be larger than

zero, which is not always the case. To bypass this problem, we use Bayesian-multiplicative treatment of count zeros with square root priors (Martin-Fernandéz et al. 2015).

Our third metric is based on Aitchison's distance and is calculated as follows:

$$d_{AFS} = \sqrt{\frac{1}{2l}\sum_{a=1}^{l}\sum_{b=1}^{l}\left(\ln\frac{y_a}{y_b} - \ln\frac{z_a}{z_b}\right)^2}$$

where l is the number of matching elements in $\mathbf{Y}$ and $\mathbf{Z}$. Note that $0.0 \leq d_{AFS} \leq \infty$. The fourth metric is also based on Aitchison's distance and is calculated as follows:

$$d_{AMS} = \sqrt{\frac{1}{2l}\sum_{a=1}^{l}\sum_{b=1}^{l}\left(\ln\frac{u_a}{u_b} - \ln\frac{v_a}{v_b}\right)^2}$$

where l now is the number of matching elements in $\mathbf{U}$ and $\mathbf{V}$. Again, $0.0 \leq d_{AMS} \leq \infty$.

The output of **Homo** also includes estimates of d_{EFS}, d_{EMS}, d_{AFS}, and d_{AMS}.

IMPORTANT NOTES:

- The subscripts 'FS' and 'MS' refers to *full symmetry* and *marginal symmetry*, and were chosen because they relate to statistical tests that evaluate deviations from symmetry (Bowker 1948) or marginal symmetry (Stuart 1955) of $\mathbf{N}$.
- The Euclidean distance has been used previously to measure compositional distances between sequences (e.g., Lockhart et al. 1994, Ho and Jermiin 2004) but the Aitchison distance is better (for further details, see Egozcue and Pawlowsky-Glahn 2011).
- The distances described above are not distances in the evolutionary or phylogenetic sense that the logdet or paralinear distances (Lake 1994, Lockhart et al. 1994, Steel 1994) are.

Compiling and installing Homo

Homo was written in C, and compiled and tested on computers running MacOS, Linux, and Windows. **Homo** is distributed as a licenced source code (homo.c) and will need to be compiled before it can be used. To install **Homo**, put the archive containing **Homo** in your preferred directory of folder, unpack the archive, compile the code, and move the binary to a location on your computer where other line-command programs are kept. On a Mac, for example, you could use the following commands:

```
$ tar -xvjf Homo_1.2.tar.bz2
$ gcc -o homo -O2 -lm -Wall homo.c
$ sudo mv homo /usr/local/bin/.
```

at which point you will be asked to provide your password — enter it! If this does not work, then you may need to speak to your system administrator or install the binary in your home directory.

Using Homo

Homo has its own interface and is easy to use. In a terminal or command window, simply type:

```
$ homo
```

hit the return key, and answer the questions.

Input file

Homo reads FASTA-formatted files with aligned nucleotides or amino acids. Ambiguous nucleotides (–, ?, N, B, D, H, K, M, R, S, V, W, Y) and amino acids (–, ?, X, Z) are allowed in these files.

Input file names may be up to 500 characters long.

Names of sequences may be as long as you like, but only the first nine (9) characters will be included in the output files, so make sure that these first nine characters identify the sequences uniquely. The reason for this constraint is that some of the output files needs to be compatible with the formats of input files for **SplitsTree** (Huson and Bryant 2006) and **Neighbor**, **Fitch**, and **Kitsch** from the PHYLIP program package (Felsenstein 2009).

Sequences may contain up to 1,000,000 sites (I did not bother to use dynamic memory allocation). If longer alignments need to be analysed, contact the author.

Options

Homo expects users to specify whether the data are nucleotides of amino acids. In the former case, users then have the option of recoding the nucleotides to one of 13 possible 3- or 2-state alphabets before the data are analysed.

Workflow

After the name of the input file has been recorded, **Homo** reads the input file to determine whether it conforms to expectations. Sequences are read and analyzed in pairs, and the results are printed to the terminal and six output files. For each pair of sequences, the number of matching unambiguous characters is counted and stored in the divergence matrix before further analysis is done. If recoding of nucleotides is required, then the corresponding rows and columns are added (e.g., if R-recoding is needed, then columns 1 and 3 are added and rows 1 and 3 are added (the order of nucleotides in the divergence matrix is A, C, G and T)). Each comparison involves computing and printing the values of X_S, v, p, d_{EFS}, d_{EMS}, d_{AFS}, d_{AMS}, and the number of sites compared.

When $n_{ij} = n_{ji} = 0$ occurs in a divergence matrix, which is likely to happen when short alignments of amino acids are analysed, **Homo** automatically adjusts the degrees of freedom and the number of matching elements in $\mathbf{Y}$, $\mathbf{Z}$, $\mathbf{U}$, and $\mathbf{V}$ to accommodate this. In other words, if we were analysing data comprising nucleotide sequences and

$$\mathbf{N} = \begin{bmatrix} 33 & 3 & 5 & 0 \\ 1 & 21 & 3 & 6 \\ 4 & 2 & 45 & 5 \\ 0 & 4 & 3 & 27 \end{bmatrix},$$

then v = 5 (because $n_{14} = n_{41} = 0$).

Output

If the input file is called `input.fsa`, the output from **Homo** includes:

- An executive summary of the results printed to the terminal
- A summary file (`input_summary.csv`) with the results from all the comparisons. The file is written in a format that allows it to be opened using Microsoft Excel
- A file (`input_Pvalues.csv`) with the p-values from the matched-pairs tests of symmetry. The file is written in a format that allows it to be opened using Excel
- A file (`input_EMS.txt`) with a matrix of d_{EMS} values. The file is written in a format that allows it to be opened using the above-mentioned phylogenetic programs

- A file (`input_EFS.txt`) with a matrix of d_{EFS} values. The file is written in a format that allows it to be opened using the above-mentioned phylogenetic programs
- A file (`input_AMS.txt`) with a matrix of d_{AMS} values. The file is written in a format that allows it to be opened using the above-mentioned phylogenetic programs
- A file (`input_AFS.txt`) with a matrix of d_{AFS} values. The file is written in a format that allows it to be opened using the above-mentioned phylogenetic programs

When a comparison cannot be completed, usually because of sparse data, **Homo** will print 'NaN' (i.e., Not a Number) instead of the normally expected output. In such cases, the output matrix may not be valid input for the above-mentioned phylogenetic programs.

Example

The merits of **Homo** are demonstrated with a survey of an alignment of small-subunit ribosomal RNA sequences from five species of bacteria: *Thermotoga maritima*, *Thermus thermophilus*, *Deinococcus radiodurans*, *Aquifex pyrophilus*, and *Bacillus subtilis*. The alignment was first analyzed by Galtier and Gouy (1995), who used the data to illustrate the performance of a phylogenetic method that is able to accommodate compositional heterogeneity across the sequences. The alignment has since been subject to similar analyses (Jayaswal et al. 2005, Ababneh et al. 2006a, Jayaswal et al. 2007).

The alignment is included with the program, so that prospective users of **Homo** can try it out using a known data set.

Using **Homo**, we first analyzed the data assuming a 4-letter alphabet. A summary of the results was printed to the terminal. The summary contains the following information:

```
___________ Highlights from the Analysis ___________
                                          Num. Prop.
p-values < 0.05                             6 (0.600)
p-values < 0.01                             6 (0.600)
p-values < 0.005                            6 (0.600)
p-values < 0.001                            6 (0.600)
p-values < 0.0005                           6 (0.600)
p-values < 0.0001                           6 (0.600)
p-values < 0.00005                          6 (0.600)
p-values < 0.00001                          6 (0.600)

Number of tests                            10
Smallest p-value                     2.644790e-12
Family-wise error rate (0.05/tests)  5.000000e-03

WARNING:
At least one pair of sequences is unlikely to have evolved
under the same Markovian process. For further details, see
bacteria_16SrRNA_summary.csv.

Spreadsheet with all results          bacteria_16SrRNA_summary.csv
Matrix with p-values                  bacteria_16SrRNA_Pvalues.csv
Matrix with Euclidean distances (m)   bacteria_16SrRNA_EMS.txt
Matrix with Euclidean distances (f)   bacteria_16SrRNA_EFS.txt
Matrix with Aitchison distances (m)   bacteria_16SrRNA_AMS.txt
Matrix with Aitchison distances (f)   bacteria_16SrRNA_AFS.txt
_____________________________________________________
```

Based on this brief summary, it is evident that six of the 10 comparisons produced low p-values, and that the smallest p-value (2.6×10^{-12}) is smaller than the Bonferroni corrected threshold (5.0×10^{-3}). Consequently, there is strong evidence that evolution has not occurred under stationary, reversible, and globally homogeneous conditions.

A more detailed result is available in the summary file — the content of this file is shown below:

Seq. 1.	Seq. 2.	Chi-square	df	p	d_EMS	d_EFS	d_AMS	d_AFS	Sites
Aquifex	*Thermotoga*	9.837658	6	1.32E-01	1.06E-02	1.20E-02	7.83E-02	1.11E+00	1238
Aquifex	*Bacillus*	61.963248	6	1.79E-11	5.94E-02	4.36E-02	3.75E-01	2.84E+00	1238
Aquifex	*Deinococcus*	55.636806	6	3.45E-10	6.18E-02	4.61E-02	3.95E-01	2.25E+00	1238
Aquifex	*Thermus*	5.278108	6	5.09E-01	7.71E-03	1.19E-02	4.65E-02	8.75E-01	1238
Thermotoga	*Bacillus*	66.040090	6	2.64E-12	5.73E-02	4.31E-02	3.42E-01	3.08E+00	1238
Thermotoga	*Deinococcus*	64.066776	6	6.69E-12	5.90E-02	4.47E-02	3.58E-01	2.85E+00	1238
Thermotoga	*Thermus*	6.077354	6	4.15E-01	1.01E-02	9.79E-03	6.96E-02	9.37E-01	1238
Bacillus	*Deinococcus*	0.686349	6	9.95E-01	4.50E-03	4.35E-03	2.77E-02	1.91E-01	1238
Bacillus	*Thermus*	50.552265	6	3.64E-09	5.28E-02	3.91E-02	3.32E-01	2.48E+00	1238
Deinococcus	*Thermus*	59.389970	6	5.99E-11	5.50E-02	4.18E-02	3.51E-01	3.02E+00	1238

NOTE: df = υ.

The table shows the results for each comparison, with those for the matched-pairs test of symmetry listed in Columns 3, 4 and 5, and the compositional distances listed in Columns 6, 7, 8 and 9. Based on the results in Column 5, it is clear that the observed p-values vary a lot across the 10 comparisons. The table also shows that there is good inverse agreement between the values of p and the values of d_{EFS}, d_{EMS}, d_{AFS}, and d_{AMS}.

Interpretation of results from the matched-pairs test of symmetry

Some knowledge on statistics is useful when interpreting the results from the matched-pairs test of symmetry. For example, it must be recognized that the p-values generated by this test fall on a scale between 0.0 and 1.0, implying that there is not much difference between the following values: 0.049 and 0.051. Yet, many scientists view the first of these p-values as statistical evidence consistent with the null hypothesis that the test was used to evaluate while the second of these p-values is viewed as statistical evidence against this hypothesis. In the present case, it is worth applying a more nuanced approach.

It also must be recognized that when more than two sequences are being surveyed, then that survey involves multiple comparisons of non-independent samples, implying that it would be in appropriate to interpret the individual p-values as outlined above. To address this dilemma, we recommend the following conservative approach:

1. If the smallest p value (i.e., p_{min}) is smaller than the family-wise error rate for the data (i.e., $p_{min} < \alpha/m$, where α is the *a priori* selected threshold used to determine whether the null hypothesis is supported or rejected and m is the number of tests (Holm 1979)), then there is evidence against the null hypothesis for the data as a whole.
2. If $p_{min} < \alpha/m$, it may be useful to know which sequence pairs have evolved under complex conditions. For this, we use the sequential Bonferroni correction (Holm 1979) (see below).

Given these two guidelines, we return to the results tabulated above. Given $m = 10$ and $\alpha = 0.05$, we find that $p_{min} < 0.005$, implying that there is evidence that evolution of the data, as a whole, has not occurred under stationary and globally homogeneous conditions. Accordingly, it would be useful to identify pairs of sequences that violate the null hypothesis. To achieve this, we use the sequential Bonferroni correction:

1. Sort, in ascending order, the rows in the summary file according to the p-values (henceforth called p_{obs})

2. Assign each row a $rank$ ranging from 1 to m (Column 4 in the table below)
3. Check whether p_{obs} for each comparison is smaller than corrected threshold (i.e., whether $p_{obs} < p_{exp} = \alpha/(m - rank + 1)$) (Column 5 in the table below)

The table below shows the results thus obtained:

Seq. 1.	Seq. 2.	p_{obs}	Rank	Check	Answer
Thermotoga	*Bacillus*	2.64E-12	1	$p_{exp} = 0.05/10$?	Yes
Thermotoga	*Deinococcus*	6.69E-12	2	$p_{exp} = 0.05/9$?	Yes
Aquifex	*Bacillus*	1.79E-11	3	$p_{exp} = 0.05/8$?	Yes
Deinococcus	*Thermus*	5.99E-11	4	$p_{exp} = 0.05/7$?	Yes
Aquifex	*Deinococcus*	3.45E-10	5	$p_{exp} = 0.05/6$?	Yes
Bacillus	*Thermus*	3.64E-09	6	$p_{exp} = 0.05/5$?	Yes
Aquifex	*Thermotoga*	1.32E-01	7	$p_{exp} = 0.05/4$?	No
Thermotoga	*Thermus*	4.15E-01	8	$p_{exp} = 0.05/3$?	No
Aquifex	*Thermus*	5.09E-01	9	$p_{exp} = 0.05/2$?	No
Bacillus	*Deinococcus*	9.95E-01	10	$p_{exp} = 0.05/1$?	No

The table shows that for six pairs of sequences there is evidence that evolution has not occurred under stationary and globally homogeneous conditions. The other pairs of sequences are consistent with the assumption of evolution under these conditions. The table also shows that the sequences can be divided into two subsets ([*Aquifex*, *Thermus* and *Thermotoga*] and [*Bacillus* and *Deinococcus*]), with each subset including sequences that are consistent with the assumption of evolution under stationary and globally homogeneous conditions. Consequently, we now have sufficient information to suggest a sensible phylogenetic approach to the analysis of these sequences—for further details, see Jermiin et al. (2017). In some instances, it is useful to analyze the p-values differently.

Phylogenetic data often contain more sequences than the data considered here. In such cases, there are other useful approaches to visualize and analyze the results. One of these involves generating a heat map showing the distribution of significant tests as a function of the sequences being compared. It is easy to generate a heat map from the results stored in the `input_Pvalues.csv` file. Using a spreadsheet program that allows conditional formatting of the cells in a spreadsheet led to the figure below:

	Aquifex	*Thermotoga*	*Bacillus*	*Deinococcus*	*Thermus*
Aquifex	1.00E+00	1.32E-01	1.79E-11	3.45E-10	5.09E-01
Thermotoga	1.32E-01	1.00E+00	2.64E-12	6.69E-12	4.15E-01
Bacillus	1.79E-11	2.64E-12	1.00E+00	9.95E-01	3.64E-09
Deinococcus	3.45E-10	6.69E-12	9.95E-01	1.00E+00	5.99E-11
Thermus	5.09E-01	4.15E-01	3.64E-09	5.99E-11	1.00E+00

In this case, a threshold of 0.0125, obtained using the sequential Bonferroni correction (i.e., $m = 10$, $\alpha = 0.05,\ i = 7$), was used to colour-code the heat map. The benefit of using this approach is that it allows users to identify (1) sequences that have evolved under different conditions (in relation to all other sequence in the data), and (2) sets of sequences that have evolved under the same stationary and globally homogeneous conditions. To facilitate identifying such groups, it is sometimes useful to do row-and-column permutations of the table. This procedure works well for medium-sized data sets

(i.e., 10 - 50 sequences) but can be challenging for larger data sets. The heat map shown in the figure below is the result of such a row-and-column permutation:

	Aquifex	*Thermotoga*	*Thermus*	*Bacillus*	*Deinococcus*
Aquifex	1.00E+00	1.32E-01	5.09E-01	1.79E-11	3.45E-10
Thermotoga	1.32E-01	1.00E+00	4.15E-01	2.64E-12	6.69E-12
Thermus	5.09E-01	4.15E-01	1.00E+00	3.64E-09	5.99E-11
Bacillus	1.79E-11	2.64E-12	3.64E-09	1.00E+00	9.95E-01
Deinococcus	3.45E-10	6.69E-12	5.99E-11	9.95E-01	1.00E+00

The heat map clearly partitions the sequences into two groups depending on the values of p_{obs}.

Another approach involves plotting in a PP plot the values of p_{obs} against the expected values of p. In this case, the expected values of p are uniformly distributed on (0, 1). The PP plot can be generated using the following actions:

1. Sort the rows in the summary file in descending order of p_{obs}
2. Assign each row a $rank$ ranging from 1 to m (Column 4 in the table below)
3. Calculate the expected p value, that is $p_{exp} = 1.0 - rank/(m+1)$ (Column 5 in table below)

Seq. 1.	**Seq. 2.**	p_{obs}	$Rank$	p_{exp}
Bacillus	*Deinococcus*	9.95E-01	1	9.09E-01
Aquifex	*Thermus*	5.09E-01	2	8.18E-01
Thermotoga	*Thermus*	4.15E-01	3	7.27E-01
Aquifex	*Thermotoga*	1.32E-01	4	6.36E-01
Bacillus	*Thermus*	3.64E-09	5	5.45E-01
Aquifex	*Deinococcus*	3.45E-10	6	4.55E-01
Deinococcus	*Thermus*	5.99E-11	7	3.64E-01
Aquifex	*Bacillus*	1.79E-11	8	2.73E-01
Thermotoga	*Deinococcus*	6.69E-12	9	1.82E-01
Thermotoga	*Bacillus*	2.64E-12	10	9.09E-02

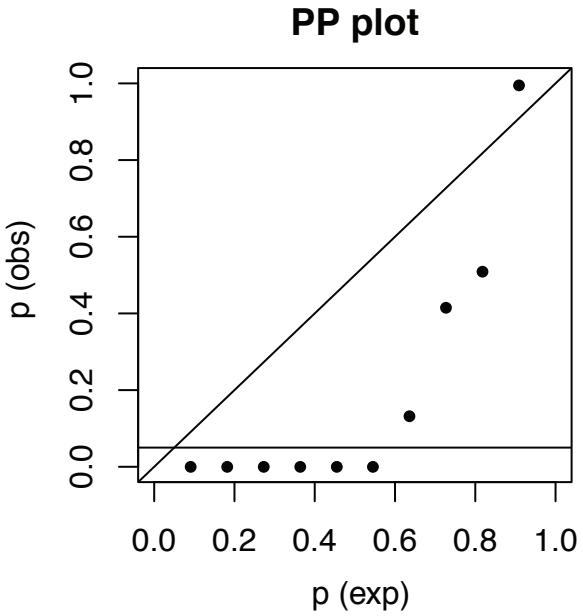
4. Save p_{obs} and p_{exp} in a file called p-values.csv
5. Use the following R script to plot p_{ops} as a function of p_{exp}

R script: PP-plot.R

```
d <- read.csv("p-values.csv", header=T)
pdf(height=4, width=3.5, file="PP-plot.pdf")
plot(d$pe, d$po, ylab="p (obs)", xlab="p (exp)", main="PP plot", xlim=c(0,1), ylim=c(0,1), pch=20)
abline(0, 1, h=0.05, lwd=1)
dev.off()
```

The resulting PP plot is displayed below. If the sequences have evolved under stationary, reversible, and globally homogeneous conditions, then the 10 dots would have fallen near the diagonal of the

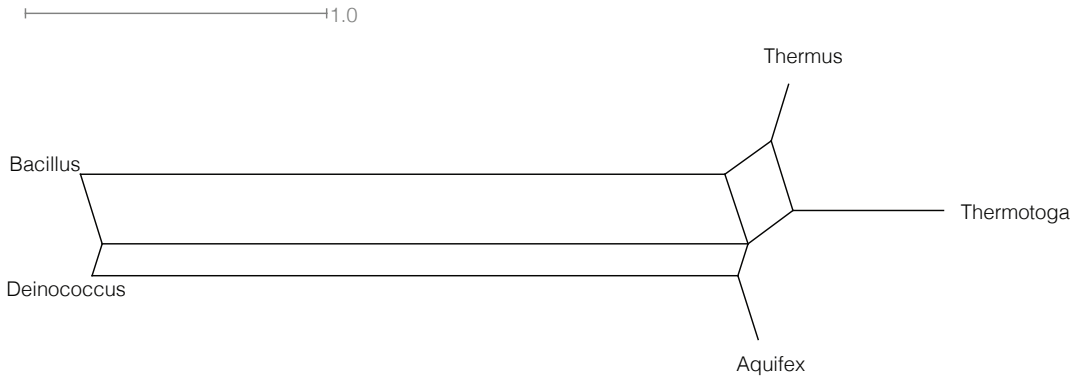
plot. However, in this case, they actually form a J-shaped distribution with a larger-than-expected proportion of dots below the horizontal line (p_{exp} = 0.05). In other words, the plot provides evidence that the data, as a whole, have not evolved under stationary, reversible, and globally homogeneous conditions. Interestingly, it also shows that some sequences are consistent with this assumption.



Visualisation of compositional heterogeneity across sequences

In addition to surveying the alignment using the matched-pairs test of symmetry, it is often useful to visualise the pattern of compositional heterogeneity across the sequences. Occasionally, this pattern is easy to detect in a heat map, but often the patterns are harder to detect because: (*a*) detecting the pattern involves making decisions basis on small differences between different p values, (*b*) the heat maps are produced from alignments with more sequences than analysed here, and (*c*) the data are more complex than those surveyed here. **Homo** facilitates this by generating four files with a matrix of d_{EFS}-, d_{EMS}-, d_{AFS}-, and d_{AMS}-values from the sequences in the alignment. Given each of these files, a network can be inferred using **SplitsTree** (Huson and Bryant 2006), and a tree can be inferred using **Neighbor**, **Fitch**, and **Kitsch** from the PHYLIP program package (Felsenstein 2009). The use of networks to explore phylogenetic data provides many benefits (Morrison 2010), so we will focus on these in the following.

The NeighborNet generated from the matrix in `bacteria_16SrRNA_AFS.txt` looks like this:



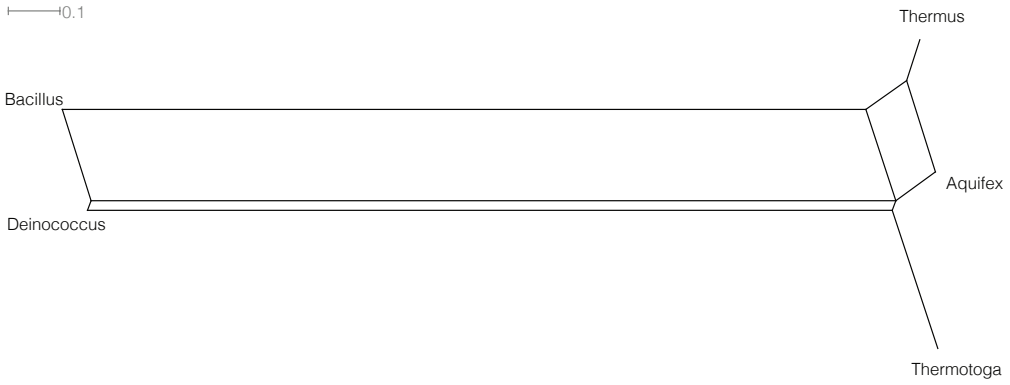
and has a least-squares fit of 95.7% between the pairwise distances in the network and the distances in the matrix.

The network displays evidence of compositional heterogeneity across the sequences in the alignment. If $n_{ij} = n_{ji}$ for all $i < j$, then $d_{AFS} = 0.0$, and the corresponding two sequences will sit next to each other in the network; otherwise, $d_{AFS} > 0.0$, and the sequences will be further apart.

The network divides the sequences into two groups, with *Bacillus* and *Deinococcus* in one group and *Thermotoga*, *Aquifex,* and *Thermus* in the other. Consequently, the test- and network-based methods identified the same pattern. This, however, may not always be the case. For example:

- If the sequences have not evolved under stationary, reversible and homogeneous conditions and they also are short, then the matched-pairs test of symmetry might not detect that the sequences have evolved under complex conditions (because the power of the test increases with the number of sites in the data). The network-based method might not be confounded by this problem—it will cluster the sequences into distinct groups of sequences depending on the conditions under which the sequences have evolved;
- If the alignment contains many sequences, it may be impractical to permute the rows and columns to find the largest groups of sequences that are consistent with evolution under stationary, reversible and homogeneous conditions. On the other hand, the network-based method will cluster the sequences into distinct groups depending on the conditions under which they have evolved;
- If the evolutionary processes are more complex than those assumed to be behind the data analysed here (for details see: Jayaswal et al. 2007), then the heat map may not reveal any groups of sequences that are consistent with the assumption of evolution under stationary, reversible and homogeneous conditions. If this were to happen, it is still possible to apply the network-based method to cluster the sequences into groups of sequences that have evolved under *somewhat* similar conditions.

The NeighborNet generated from the matrix in `bacteria_16SrRNA_AMS.txt` looks like this:



and has a least-squares fit of 100%.

Unlike the previous case, the focus is now on the difference between the vectors $\mathbf{U}$ and $\mathbf{V}$ and, hence, the difference between nucleotide frequencies in pairs of sequences. The network reveals that there is a large difference in the nucleotide composition between the sequences in one group (*Bacillus* and *Deinococcus*) and those in the other group (*Thermus*, *Aquifex* and *Termotoga*). The presence of this difference in nucleotide composition (represented by the edge separating *Bacillus* and *Deinococcus* from the other species) implies that the five sequences are unlikely to have evolved under stationary conditions. This result is consistent with those reported previously (Ababneh et al. 2006a, Jermiin et al. 2017).

Conclusive remarks

The example shown above shows how useful the matched-pairs test of symmetry might be. However, although the test can detect cases where sequences have evolved under different conditions, it is not able to detect under which conditions the sequence might have evolved. Other useful matched-pairs tests are available (Ababneh et al. 2006a). A detailed explanation of these tests is available elsewhere (Jermiin et al. 2017).

It is important to remember that networks inferred from matrices of compositional distances do not represent the evolutionary history of the data. Consequently, if such a network has a strong tree-like appearance (like the long edge in the networks above) and the same splits between sequences also are found in the phylogenetic tree inferred (e.g., using model-based phylogenetic methods) from the same alignment, then there is cause for concern about the phylogenetic tree: did the corresponding bifurcations in the phylogenetic tree actually occur in the past or are they artefacts caused by model misspecification? The answer to this question require further phylogenetic analysis (as in: Jayaswal et al. 2007)—being able to pose the question in the first place simply requires a survey of the data as described above.

One topic not discussed in this manual is what to do if there is statistical evidence that the sequences have not evolved under SRH conditions. Several phylogenetic methods are available for such data (Barry and Hartigan 1987, Reeves 1992, Steel et al. 1993, Lake 1994, Lockhart et al. 1994, Steel 1994, Galtier and Gouy 1995, Steel et al. 1995, Yang and Roberts 1995, Gu and Li 1996, Galtier and Gouy 1998, Gu and Li 1998, Galtier et al. 1999, Tamura and Kumar 2002, Foster 2004, Thollesson 2004, Jayaswal et al. 2005, Blanquart and Lartillot 2006, Jayaswal et al. 2007, Blanquart and Lartillot 2008, Dutheil and Boussau 2008, Jayaswal et al. 2011a, Jayaswal et al. 2011b, Dutheil et al. 2012, Zou et al. 2012, Groussin et al. 2013, Jayaswal et al. 2014) but it is beyond the scope of this manual to review them in the context of the matched-pairs test of symmetry—for a brief comparison and discussion, see Jermiin et al. (2017).

Publishing results obtained using Homo

Please cite the following paper if you publish results generated using **Homo**:

Rouse GW, **Jermiin LS**, Wilson NG, Eeckhaut I, Lanterbecq D, Oji T, Young CM, Browning T, Cisternas P, Helgen LE, Stuckey M, Messing CG. 2013. Fixed, free and fixed: The fickle phylogeny of extant Crinoidea (Echinodermata) and their Permian-Triassic origin. *Mol. Phylogen. Evol.* 66, 161-181.

It was the first paper to rely on **Homo**. A paper describing **Homo** in detail will appear in due course.

Version History

1.3 Code expanded with the inclusion of Aichison's distances based on compositional vectors of full and marginal frequencies and with the inclusion of Euclidean distances based on vectors containing the same type of frequencies.

1.4 Corrected error in code calculating Aichison's distance based on marginal frequencies.

Contact person

Dr Lars Jermiin
E: lars.jermiin@anu.edu.au
A: Research School of Biology, EBL, Australian National University, Canberra, ACT 2601, Australia.

References

Ababneh F, Jermiin LS, Ma C, Robinson J. 2006a. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics, 22:1225-1231.

Ababneh F, Jermiin LS, Robinson J. 2006b. Generation of the exact distribution and simulation of matched nucleotide sequences on a phylogenetic tree. J. Math. Model. Algor., 5:291-308.

Aitchison J. 1986. The Statistical Analysis of Compositional Data. London, Chapman and Hall.

Barry D, Hartigan JA. 1987. Statistical analysis of hominoid molecular evolution. Stat. Sci., 2:191-210.

Blanquart S, Lartillot N. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol. Biol. Evol., 23:2058-2071.

Blanquart S, Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol., 25:842-858.

Bowker AH. 1948. A test for symmetry in contingency tables. J. Am. Stat. Assoc., 43:572-574.

Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. BMC Evol. Biol., 8:255.

Dutheil JY, Galtier N, Romiguier J, Douzery EJP, Ranwez V, Boussau B. 2012. Efficient selection of branch-specific models of sequence evolution. Mol. Biol. Evol., 29:1861-1874.

Egozcue JJ, Pawlowsky-Glahn V. 2011. Basic concepts and procedures. In: Pawlowsky-Glahn V, Buccianti A editors. Compositional Data Analysis. Chicester, John Wiley and Sons, p. 12-28.

Felsenstein J. 2009. PHYLIP (Phylogeny Inference Package). Seattle, Distributed by the author.

Foster PG. 2004. Modelling compositional heterogeneity. Syst. Biol., 53:485-495.

Galtier N, Gouy M. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc. Natl. Acad. Sci. USA, 92:11317-11321.

Galtier N, Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogenous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol., 15:871-879.

Galtier N, Tourasse N, Gouy M. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science, 283:220-221.

Groussin M, Boussau B, Gouy M. 2013. A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. Syst. Biol., 62:523–538.

Gu X, Li W-H. 1996. Bias-corrected paralinear and logdet distances and tests of molecular clocks and phylogenies under nonstationary nucleotide frequencies. Mol. Biol. Evol., 13:1375-1383.

Gu X, Li W-H. 1998. Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. Proc. Natl. Acad. Sci. USA, 95:5899-5905.

Ho SYW, Jermiin LS. 2004. Tracing the decay of the historical signal in biological sequence data. Syst. Biol., 53:623-637.

Holm S. 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat., 6:65-70.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol., 23:254-267.

Jayaswal V, Ababneh F, Jermiin LS, Robinson J. 2011a. Reducing model complexity when the evolutionary process over an edge is modeled as a homogeneous Markov process. Mol. Biol. Evol., 28:3045-3059.

Jayaswal V, Jermiin LS, Poladian L, Robinson J. 2011b. Two stationary, non-homogeneous Markov models of nucleotide sequence evolution. Syst. Biol., 60:74-86.

Jayaswal V, Jermiin LS, Robinson J. 2005. Estimation of phylogeny using a general Markov model. Evol. Bioinf. Online, 1:62-80.

Jayaswal V, Robinson J, Jermiin LS. 2007. Estimation of phylogeny and invariant sites under the General Markov model of nucleotide sequence evolution. Syst. Biol., 56:155-162.

Jayaswal V, Wong TKF, Robinson J, Poladian L, Jermiin LS. 2014. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. Syst. Biol., 63:726-742.
Jermiin LS, Jayaswal V, Ababneh FM, Robinson J. 2017. Identifying optimal models of evolution. In: Keith J editor. Bioinformatics: Volume 1: Data, sequence analysis, and evolution. Totowa, NJ, Humana Press, p. 379-420.
Lake JA. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. USA, 91:1455-1459.
Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol., 11:605-612.
Martin-Fernandéz JA, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J. 2015. Bayesian-multiplicative treatment of count zeros in compositional data sets. Stat. Model., 15:134-158.
Morrison DA. 2010. Using data-display networks for exploratory data analysis in phylogenetic studies. Mol. Biol. Evol., 27:1044-1057.
Reeves J. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by the mitochondrial DNA. J. Mol. Evol., 35:17-31.
Schweder T, Spjotvoll E. 1982. Plots of P-values to evaluate many tests simultaneously. Biometrika, 69:493-502.
Steel MA. 1994. Recovering a tree from the leaf colourations it generates under a Markov model. Appl. Math. Lett., 7:19-23.
Steel MA, Lockhart PJ, Penny D. 1993. Confidence in evolutionary trees from biological sequence data. Nature, 364:440-442.
Steel MA, Lockhart PJ, Penny D. 1995. A frequency-dependent significance test for parsimony. Mol. Phylogenet. Evol., 4:64-71.
Stuart A. 1955. A test for homogeneity of the marginal distributions in a two-way classification. Biometrika, 42:412-416.
Tamura K, Kumar S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. Mol. Biol. Evol., 19:1727-1736.
Thollesson M. 2004. LDDist: a Perl module for calculating LogDet pair-wise distances for protein and nucleotide sequences. Bioinformatics, 20:416-418.
Vera-Ruiz VA, Lau KW, Robinson J, Jermiin LS. 2014. Statistical tests to identify appropriate types of nucleotide sequence recoding in molecular phylogenetics. BMC Bioinformatics, 15 (Suppl. 2):S8.
Yang Z, Roberts D. 1995. On the use of nucleic acid sequences to infer early branches in the tree of life. Mol. Biol. Evol., 12:451-458.
Zou LW, Susko E, Field C, Roger AJ. 2012. Fitting nonstationary general-time-reversible models to obtain edge-lengths and frequencies for the Barry-Hartigan model. Syst. Biol., 61:927-940.