Shannon Heterogeneity In Alignments tool

Version 1.1 Manual

Joe Parker

Evolutionary Biology Group
Department of Zoology
University of Oxford
UK
OX1 3PS

http://evolve.zoo.ox.ac.uk
+44 (0)1865 281 987 (tel)
+44 (0)1865 271 249 (fax)

Contents

Introduction

Welcome to the documentation for this version of the Shannon tool – the first public release of this software. I hope you find the package useful but it is still in the early stages of development. As such please let me know if you find any bugs or have suggestions for improvements. This package has undergone only limited testing so you use it at your own risk – see the disclaimer below.

License & Disclaimer

License

This software is supplied under the GNU Lesser General Public License, Version 3. This is an open-source software liscence, and others are authorised and encouraged to examine and modify code if they see fit, as long as the contribution of previous workers is recognised. For more details see
http://www.gnu.org/licenses/lgpl.html

Disclaimer

No guarantee of the **functionality** of this software, or of the **accuracy of results** obtained using it is made, expressed or implied. The programmers, authors and editors of this documentation and the institutions they represent **will not be held responsible** for any errors of analysis, damage to software or hardware, or other losses incurred as a result of using this programme.

What is it?

This software aims to provide a quick and easy method by which the amount of amino-acid or nucleotide heterogeneity within an alignment may be quantified and compared between alignments.

What can it do?

A number of heterogeneity measures are implemented in this package. Some are scored sitewise along an alignment, giving a position-by-position score of alignment heterogeneity. Others give a summary (Hamming Distance) of heterogeneity in the alignment as a whole that can be compared for similar alignments.

System requirements

Java

Shannon is written and compiled for Java 1.5.0, ("J2SE 5.0"). You will need a computer capable of running this version of Java or higher. For most platforms it is sufficient to download the required version of Java directly from java.sun.com Mac OS X users should note however that on versions 10.4.5 and lower the procedure for upgrading to Java 1.5.0 (from 1.4.2, the default on these systems) differs. They should consult http://www.apple.com/downloads/macosx/apple/macosx_updates/index_abc.html for further instructions.
If unsure, typing ' `java -version` ' from a Terminal session will tell you which version of Java is currently used by the operating system. Max OS X 10.5.x ('Leopard') users are lucky – Java 1.5 is installed on these systems as standard!

Hardware

We have not identified any specific minimum hardware requirements; these in any case scale with the number of taxa and the length of the sequences. Generally speaking, at least 256 Mb of system memory (physical RAM, not virtual memory or swap file cache) should be available for each separate instance of the program that is running. Note that in some rare cases this may not be sufficient and users will need to increase the amount of memory available to the Java Virtual Machine (JVM) using the `–Xms` command; for more information type ' `man java` ' from the command-line or see http://edocs.bea.com/wls/docs70/perform/JVMTuning.html

Important: a 'progress bar' for long operations is not currently implemented. This means that sometimes the package may appear to 'hang' for periods while, for instance, calculating pairwise Hamming distances in large alignments. Please be patient.

Installation

Given the correct JVM (1.5.0 or higher) is available, installation of is simple. This package contains two files: a Macintosh application folder for Mac OS X users, and a java ".jar" file for all other operating systems.

Installation – Mac OS X

Simply drag the 'Shannon Heterogeneity In Alignments v1.1' application into your Applications folder, or wherever you keep phylogenetic software on your machine.

Installation – Other platforms

Drag or copy and paste the .jar file to a location on your hard drive. If you normally need to run Java applications from the command-line, make sure you make a note of the file path.

Usage: input file requirements

Alignment assumptions

Sequences must be aligned and of the same length, though gaps are allowed. Gaps are treated as informative. IUPAC ambiguity codes are included in the Hamming distance calculation but disregarded in the sitewise heterogeneity measures' calculation.

Format

Input files must be in fasta format. Users are discouraged from using non-standard characters, e.g. those other than alphanumerics (a-z, A-Z, 0-9) and some punctuation characters, such as [',', '.', '-', '_']. This is because the behaviour of some of these characters can be hard to predict.

Usage: running an analysis

Mac OS X

Double-click on the application.

Other platforms

Depending on your system settings, you may be able to double-click on the .jar file to launch Java and load the application (you may be prompted to verify that you wish to launch Java.)

If this doesn’t work you will need to launch the application from your command-line (also variously referred to as ‘command-prompt’, ‘MS-DOS’, ‘terminal’, or ‘console’ depending on the platform.) Open a command-line window and navigate to the folder where the application is installed. Then type:

```
java -jar Shannon_v1.1_sealed.jar
```

The application should then launch. Note that since this method opens the Shannon package as a ‘child’ of the command-line window you have opened, closing the command-line window will on some platforms also exit Shannon, and any analyses you have generated will be lost (input alignments should be unaffected though.)

Usage: interpreting analyses & general operations

The heterogeneity measures

Four measures are implemented:

- Shannon information entropy[1] (sitewise by position)
- 1-(consensus frequency)[2] (sitewise by position)
- 1-(heterozygosity index H)[3] (sitewise by position)
- Hamming Distance[4] (summed pairwise across alignment)

The Shannon, consensus and H index scores are also summed across the alignment, but not pairwise. For all these measures, higher scores reflect more heterogeneity; the Shannon entropy naturally scales with the number of sequences in the alignment so as a crude normalization to compare alignments we also calculate the 'normalised' Shannon entropy as the mean entropy per sequence. We have not yet ascertained whether there is a more theoretically sound way to compare entropies in alignments of differing sizes.

Loading an alignment

To load a new alignment, choose 'Add alignment' from the File menu. Shannon will calculate sitewise and summary heterogeneity scores for this alignment and after a brief delay update the output.
On some operating systems the graphs may only refresh when the mouse is moved over them, or need the alignment to be loaded a second time.

[1] This is the Shannon Informtion entropy, *H(x)* as introduced by Shannon (1948.) It is calculated per position as:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_m p(x_i)$$

Where *p(x$_i$)* = observed probability (frequency) of amino acid or nucleotide *x* for all *i* possible amino acids or nucleotides. Note firstly that absent amino acids or nucleotides will be included in the equation (though with an entropy of zero) and secondly that the base of the logarithm taken is *m*, the number of possible 'information states.' For nucleotide data this is taken to be 5 (a,c,g,t plus gaps) and for universal amino acid data this is taken to be 23 (20 amino acids plus start, stop and gaps.)

[2] Simply the frequency of the most common nucleotide or amino acid at that position; an alignment with {"a", "a", "c", "g", "t"} at a particular position for instance has a consensus frequency of 0.4 ("a" is the most common nucleotide observed at 2/5 nucleotides.)

[3] The Heterozygosity index, long used in genetics to quantify the allelic diversity, *h*, in a population. This is given by:

$$h = 1 - \sum_{i=1}^{m} x_i^{\,2}$$

Where x_i^2 is the frequency of each observed amino acid or nucleotide, squared.

[4] Hamming distance is the mean of all pairwise distances in an alignmnent, where the distance of any pair of sequences is defined as the number of positions at which they have differing amino acids or nucleotides.

Output panes

The main application window is divided into two panes – the Graphical View and the Data View. The Graphical View provides a quick visual overview of the alignment heterogeneity as it relates to other previously-loaded alignments analyzed in the same session, while the Data View returns the numerical heterogeneity information for the most recently-loaded alignment.

Data View

The Data View output contains heterogeneity information from the most recently-loaded alignment. You should be able to select, copy and paste this successfully to a variety of other applications, such as Microsoft Excel or Minitab.

Graphical View

This shows the value of the Shannon, consensus, and h index measures at each position in the alignment for all traces alignments loaded since the Shannon tool was started (or since the last 'clear all' command), as well as a graph (‘Compare measures’) summarising the relative values of each index in the last alignment that has been successfully loaded. These graphs copy straight to clipboard in some applicaions; others will need a third-party screen capture program (such as 'File>Grab>Selection...' in Preview on Mac OS.)

Clearing the output

Selecting ‘Clear all’ from the File menu removes all traces from all graph plots (note that on some operating systems axis labels may remain, and the graph axes may not re-size.)

Help

As well as this documentation a summary live help page is available through the ‘Help’ menu.

Quitting

To exit the application, choose ‘quit’ from the File menu.

Usage: caveats and warnings

Computational constraints:

Users should be aware that although this package has received limited testing on our development machines, this is the first public release of the package. As such, the real-world performance of the core packages (which in any case are highly dependent on hardware architecture, and software architecture to a degree) is unknown at this point; in fact, we would appreciate any feedback concerning performance.

As a guide, most alignments should load in a few minutes on a 1.25 GHz Apple PowerPC machine under Java 1.5.0 / Mac OS 10.4.10.

As usual, physical factors increasing compute time performance will include:

- Slower system CPUs
- Slower system bus speeds
- Slower system RAM access

In particular, because Shannon uses a lot of
memory at present heavy reliance on virtual memory coupled with a slow hard drive is likely to adversely affect performance)

Problem parameters that will slow the analysis include:

- Number of taxa - increasing numbers of taxa will increase compute time and memory usage. This particularly affects calculation of the Hamming distance since this is a pairwise measure – the number of possible pairwise comparisons rises much faster than the number of sequences.
- Sequence length – long sequences will take more time.
- Amino acid alignments will take longer to compute than nucleotide ones, since more potential residue probabilities (23 versus 5) must be evaluated (this is unlikely to be a major constraint)

Biological constraints

We have developed this method to analyse heterogeneity information so that researchers can quickly and simply evaluate the relative variability in an alignment as an exploratory analysis, perhaps to identify highly variable regions of a gene, or to compare the variability of a locus in separate patients' viral populations where the same or similar numbers of viral samples have been isolated.
This package is not a substitute for a proper phylogenetic analysis of selection or distance – in particular users should bear in mind that these methods all treat individual sites and sequences as independent pieces of information in the statistical sense. This assumption is violated both along sequences (by epistasis or secondary or tertiary structural constraints may operate) and between sequences (lack of independence due to shared phylogenetic ancestry – see Harvey & Pagel, 1991.)

FAQ

Note: this is just a preliminary list of FAQs; for any persistent problems, or if you have any other comments, please contact the author.

Q: Why can't I open the package?
A: Check you have the appropriate version of Java (1.5.0 or higher) installed. If you are having persistent problems contact the author.

Q: Why can't I load my alignments?
A: Try running the sample alignment included in this package. If the sample files don't run either you may well have the wrong version of java installed; check and install the correct version. If the example alignments load but your data does not, it is likely you have parsed the .fasta files incorrectly. If you are editing your alignments by hand, check they load into a sequence editor program such as BioEdit, Se-Al or Geneious and then try re-exporting the alignment as a .fasta file from there. If problems persist, contact the author.

Q: Why do my alignments take a long time to load / the program crashes when I load my sequences?
A: Your alignment is probably too big. You can try increasing the memory allocated to Shannon (see below) but if this does not solve the problem you may need to analyse a subset of the alignment. Hopefully these issues will be addressed in a future release.

Q: Why do I get an error message saying I've run out of memory?
A: Typically this will manifest itself with a message such as "`Exception in thread 'main': java.lang.OutOfMemoryError: java heap space`" but may also take the form of hangs or crashes.
This error arises then the JVM doesn't have enough system memory ('RAM') available to hold all the data it needs to. Unfortunately increasing your computer's virtual memory allocation will not solve this problem. You can try increasing the default amount of RAM allocated to the JVM with the '`-Xms`' command (see the 'System Requirements > Hardware' section of this documentation for details.)

Q: Why are there three separate sitewise statistics? Which is best?
A: Each statistic is sensitive to slightly different types of diversity. For instance, the consensus frequency measure doesn't pick up on the diversity of non-consensus amino acids or nucleotides – an alignment with sequences displaying bases with {"a", "a", "a", "a", "c", "g", "t"} at a particular position has the same consensus frequency as an alignment displaying bases {"a", "a", "a", "a", "t", "t", "t"}, but different Shannon and heterozygosity index scores.
As to which is 'best' work is ongoing to evaluate the utility of each statistic. For now we suggest users compare all statistics and inspect their alignments where they significantly diverge.

Contact

The author of this documentation and software is Joe Parker. You can contact him at:

Joe Parker
Viral Evolution Group
Department of Zoology, University of Oxford
South Parks Road
OX1 3PS
United Kingdom

Or by email: joe.parker@zoo.ox.ac.uk