Sequence Matrix Manual

Sequence Matrix is designed to ease the assembly of multigene datasets for phylogenetic analysis. In particular, we wanted software that (1) produces a spreadsheet with information on how much gene and sequence information is available for individual taxa, (2) eases the generation of taxon and character sets, (3) allows for an export of data and character sets into NEXUS and TNT format, (4) allows for the exclusion of individual sequences from export (as opposed to entire genes), and (5) does not have narrow limits on dataset size.

1. Installation Instructions

(a) The easiest way to install the current version of Sequence Matrix is to double click on the "SequenceMatrix WebStart.jnlp" file. The first time you run this program, Java will request permission to allow Sequence Matrix to access your hard disk. Please say "Yes" at this prompt. The second time you run this program, Java will prompt you to install the program into your "Applications". Webstart will regularly check the TaxonDNA webpage for updates.

(b) An alternative method of installing Sequence Matrix is to run the "SequenceMatrix.jar" file directly. However, by default, this will start Sequence Matrix with only 64MB of memory.

Platforms: Sequence Matrix runs on Java, which is platform independent.

2. What types of datasets can be combined?

Currently only files in basic FASTA, NEXUS, and MEGA formats can be imported and the data have to be DNA sequences. The files can be dropped into the Sequence Matrix window or can be imported by using the "add sequences" option under "File".

3. Combining data and the data overview spreadsheet

Sequence Matrix uses a spreadsheet to display how much data are available for all genes and taxa. Taxa are listed in rows and genes in columns. One gene after another can be imported by using the "add sequences" function under "File" or by dragging and dropping the files into the Sequence Matrix window. Note that Sequence Matrix uses the names of the files that are imported as column headers (=gene names). We thus recommend to use gene names as file names. Each cell in the spreadsheet provides information on sequence length (excluding leading and trailing gaps). The spreadsheet can be sorted by taxon name, species epithet, number of characters, number of character sets (=usually genes), and total length of sequence data. It can also be exported as a tab-delimited file and thus used to ascertain that all available data are included in the matrix.

The taxon samples in individual gene files do not have to be identical. Only sequences for taxa with identical names are concatenated. All other sequences are added by inserting new taxon rows. Taxon names can be changed and if the change creates a taxon name that is already in the data set, the two rows are fused after alerting the user to this circumstance (when both rows have data for the same gene, the longer sequence is used).

When a NEXUS file is opened that already contains a character set, Sequence Matrix asks whether the dataset should be split back into its components. Note that this option is currently not available for TNT and that it is thus advisable to always export combined matrices in both formats.

4. Exporting the combined matrix

The combined matrix can be exported in NEXUS or TNT format (see above for limitations with regard to importing TNT files). Under "Settings" and "Preferences" the user can specify how to export taxon names and whether the data should be exported as an interleaved matrix (NEXUS). Individual sequences can be excluded from an export by double clicking on their cell in the spreadsheet. An excluded sequence will be marked as "Canceled". This function allows the user to explore the impact of individual sequences on the outcome of the phylogenetic analysis (as opposed to exploring the role of a gene for all taxa). Note that Sequence Matrix will always export missing data as "?" and indels as "-" regardless of what symbol was used for these states in the partition data files. Also, on Macintosh computers the exported file will not be immediately recognized as a NEXUS file. Please open the file in PAUP using the "all files" option in the open menu. Once saved by PAUP, the file will be recognized as a NEXUS file.

5. Creating Taxon Sets for exploring the impact of missing sequences

Multigene data sets are rarely complete and some taxa are data-deficient. It is often desirable to explore whether the missing data has a negative impact on tree resolution and support. In order to facilitate this process, Sequence Matrix automatically creates taxon sets that contain only those taxa that, for example, have data for more than five genes or more than 2000 bp of data. The user can specify what kinds of taxon sets should be exported under "Settings" and "Taxonsets". Note that this option only creates taxonsets. It does not delete any data.

6. How do I cite Sequence Matrix?

Meier, R., S. Kwong, G. Vaidya, and P. K. L. Ng. 2006. *DNA Barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success.* Systematic Biology, 55: 715-728.

7. Can I distribute Sequence Matrix?