Species Identifier Manual

Species Identifier is the software that implements the techniques that are described in Meier et al. 2006. Although the software was mainly designed for data collected from Genbank, it can also be used for new data. Species Identifier is particularly useful for calculating intra- and interspecific distances, exploring different ways to identify DNA sequences to species, and grouping sequences into clusters based on pairwise distances. Species Identifier also includes some tools that are useful for aligning sequences obtained from GenBank.

1. Installation Instructions

(a) The easiest way to install the current version of Species Identifier is to double click on the "SpeciesIdentifier WebStart.jnlp" file. The first time you run this program, Java will request permission to allow Species Identifier 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 Species Identifier is to run the SpeciesIdentifier.jar file directly. However, by default, this will start Species Identifier with only 64MB of memory. To use more memory, you can use the SpeciesIdentifier.bat batch file provided (Windows-only). If this procedure fails, it is most likely due to the batch file not finding your copy of java.exe. Use the search tool in Windows to locate this file, open the batch file in Notepad, and replace "%WINDIR%\System32\java.exe" with the correct location in the batch file. If your PC has less than 1,000 MB of RAM, you will have to lower the value in the batch file.

Platforms: Species Identifier runs on Java, which is platform independent. On Macintosh systems, however, only under (a) will a large amount of memory be available to the program.

2. What types of datasets can be opened?

Species Identifier can open any FASTA and most Mega and Nexus files. TNT support will be added soon. All sequences should be aligned, but Species Identifier has some features that make alignment easier (see below). Please take note of the following conventions and features:

- Species Identifier generally displays the species name of a sequence even if the sequence name in the file is rather convoluted (e.g., the following name: gi|*27884345|gb|AF165466.1| Aburria aburri cytochrome b (cytb) gene, partial cds; mitochondrial*" will be displayed as "*Aburria aburri cytochrome (gi|27884345)*"). Species Identifier guesses the correct name by looking for a string of words that looks like a species name; i.e., a two part name separated by a single space, with the first word having an initial capital letter. While it is generally pretty good at recognizing the species name, we recommend that you use the "Species Summary" tool to look through all the species names that the program has identified. If you find a misidentified sequence, you can generally fix the problem by ensuring that the species name has been typed correctly; i.e., with an initial capital letter on the genus name and *without* a capital letter on the species name. If this fails, add a fake subspecies name after the species name, or mangle the part of the sequence name that Species Identifier is misidentifying as a species name (these changes can be made in the "Sequence Name" window). Note that although Species Identifier only displays the species name in the "Sequences" window, the full name is internally retained and also used when the file is saved.

- Species Identifier automatically links each sequence from GenBank to its webpage at NCBI. Use the "Go to this GI number on NCBI" button to access the page and to resolve problems with names and sequences.

- Genbank sequences have unique GI numbers. Species Identifier will similarly assign a unique number for new sequences in a data set. This unique number code is based on the computer clock and ensures that even multiple sequences with identical names will retain their unique identity.

3. My dataset includes short or ambiguous sequences that I would like to ignore in some analyses.

- Species Identifier has a configuration page that allows the user to (a) specify minimum sequence overlap (b) treatment of ambiguous bases, and (3) treatment of indels. The default is a minimum overlap of 300 bp, using ambiguous bases, and treating internal indels as information and terminal indels as missing data.

4. How do I get started?

- Usually the best starting point is the "species summary" from the "Actions and Views" menu. After clicking the "calculate now" button Species Identifier will display the number of sequences, species, etc... It will also list the species names and the number of sequences per species.

- The "pairwise summary" option provides an overview over the intraspecific and interspecific, congeneric pairwise (uncorrected) distances. The distances can be exported into a file and/or sent to the clipboard. Note, that calculating all pairwise distances can be slow if the dataset is large.

- The "pairwise explorer" allows the user to identify which sequences are providing unusual genetic distances.

5. "I only want to use sequences with complete overlap..."

Many datasets consist of sequences of different length; i.e., the sequences have leading and trailing gaps of different sizes. "Complete Overlap" allows the user to find the largest set of sequences with a defined amount of complete overlap. Species Identifier "walks" along the entire alignment and states how many sequences have complete overlap for a particular section of the sequence. The user can then export this section as a new dataset in fasta format. Note that sequences containing only internal gaps or 'N' nucleotides will be counted as being 'complete' unless specified otherwise in the "ambiguous" option within "Complete Overlap" or in the "configuration" menu.

6. Species Identification

The species identification methods described in Meier et al. (2006) are available under "Best Match/Best Close Match" and "All Species Barcodes". The results can be sent to the clipboard for closer inspection of which sequences are misidentified.

7. Clustering sequences based on pairwise distances

Species Identifier's "Cluster" algorithm clusters sequences according to pairwise distances. Note that three sequences can have two distances complying with a particular pairwise distance threshold while the third distance may exceed the threshold (see Meier et al., 2006). As described in Meier et al. (2006), "Cluster" will produce sets of sequences for which each sequence has at least one sequence match below the specified threshold distances. Threshold violations are reported for each cluster.

8. Consensus Barcode Generator

The barcodes discussed in the barcoding literature are specimen barcodes and not species barcodes because they do not reflect the genetic variability within species. The "Consensus Barcode Generator" produces consensus sequences from all conspecific sequences for the same species in the dataset. The consensus barcodes can be used to determine whether each species has a unique barcode (just run "pairwise explorer" on a dataset composed of consensus barcodes) and to identify unique combinations of characters for diagnosing species as required by some versions of the phylogenetic species concept.

9. How do I prepare the sequences for alignment using ClustalW or AlignmentHelper?

A common problem of Clustal and other alignment programs is that sequence names are truncated. For example, in aligning sequences with full GenBank names in Clustal, only the first 30 characters of the sequence names are retained. The solution is the "Clustal Mapping" function of Species Identifier. It replaces the full Genbank name with the GI number for input into Clustal ("Clustal Input"). After aligning the sequences in Clustal the full GenBank names can be reinstated using Clustal Mapping:

Export with GI numbers: Open the file with the sequences in need of alignment. Choose “Clustal Mapping” in the "Modules" menu, and enter a file name into the "Clustal Input" field. Click “Export now!”. Your data file will be written as a FASTA file with shortened names, which can be changed back into their original names after alignment. Species Identifier uses GI numbers to uniquely identify sequences. If a sequence in your input file does not have a GI number, Species Identifier will create a unique identifier for the sequence. It will need to write these unique identifiers into your original file, however. We thus recommend you keep a backup copy of your original file just in case.

Reinstating original GenBank names: Open the file containing the original sequences with the full names. Under “Clustal Mapping” in the "Modules" menu, enter the name of the FASTA file containing the aligned sequences. Click “Import into this dataset”. Species Identifier will now **replace** all the sequences in the original file with the aligned sequences obtained from the FASTA file containing aligned sequences. Sequences present in the original file but not present in the aligned file will be removed from the original file and displayed in a new Species Identifier window titled "Missing sequences". Sequences present in the aligned file which are not present in the original file will be added to the original file under their abridged name.

Note that while this process will replace the sequences, it will not replace the original file itself. All changes will only be made in the Species Identifier window. You will have to save the file to write the changes to disk.

For aligning protein-encoding genes, we recommend AlignmentHelper which is available from http://inbio.byu.edu/faculty/dam83/cdm/. AlignmentHelper will translate all sequences and use the amino acid sequence for alignment. It will then map the nucleotide sequences back onto the amino acid alignment.

10. How do I export a dataset in a custom format?

The “Export this dataset” module in the “Modules" menu allows you to customize how to export your dataset (supported formats: Nexus, Mega, Fasta). For instance, you can export your dataset as Nexus by selecting the following options:

1. “Which values would you like me to output?”: Select “Genus, species and GI number”.
2. “How would you like me to arrange names and sequences?”: Select “Taxon name, followed by the sequence name on the same line (like Nexus)”.
3. “How long can I make sequence names?”: Select “PAUP* 4.0 maximum allowed [127 letter]”.
4. “Which character should I use to indicate taxon names?”: Select “No characters”.
5. Check “Insert headers/footers for”, and select “NEXUS”.
6. Check “Change spaces to underscore in the taxon name”.
7. Click “Go!” and export to file.

Species Identifier can export Nexus files by using the 'NEXUS' option in the 'Export' menu. However, "Export this dataset" can give you much greater control over how you want your dataset to be exported.

11. I want COI sequences only, but GenBank gives me sequences that also include sequence data from adjacent genes (e.g. t-RNAs).

We are currently working on an additional program (GenBankExplorer) that will be able to excise only the desired sequence by utilizing the annotations in Genbank. A preliminary version of the software can be requested from Rudolf Meier (dbsmr@nus.edu.sg) or Gaurav Vaidya (gaurav@ggvaidya.com).

12. How do I cite Species Identifier?

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.

13. Can I distribute Species Identifier?

You may distribute Species Identifier under the terms of the GNU General Public License 2 (or any later version). The license is available from http://www.gnu.org/licenses/gpl.html.