Sequence Classifier for Geneious R8 or later

Aaron Kennedy (USDA-APHIS-PPQ) and Biomatters

November 10, 2014

Overview

The Classify Sequences plugin classifies a query sequence by aligning it against all sequences in a specified database, and then choosing an appropriate taxon according to user-specified identity levels. Single or multiple genes can be used for classification, and multiple query sequences can be run at once.

The plugin first performs pairwise global alignments (with free end gaps) between the query sequence and each sequence in the database. When sequences from multiple genes are available for classification, pairwise alignments for each gene are performed separately, and these alignments are then concatenated if there are database sequences with the same name for each gene.

The **Overlap Identity** between your query and the database sequences is then used to determine the likely taxon of your query sequence by picking the database sequence with the highest identity to the query. The overlap identity is the pairwise identity in the region in common between the query and database sequence. The sequence will only be classified if it meets a minimum overlap identity that the user specifies, and if multiple database sequences have similar overlap identities, then the query sequence will be classified to the taxonomic level that these sequences have in common.

It is also possible to set cutoffs for classifying sequences at various taxonomic levels. For instance, if 95% is set as a minimum identity to classify at species level, and the top match to the database has an overlap identity of 94.5%, then the query will only be classified to genus level. Thus, you need to know the approximate levels of sequence identity within and between the different taxonomic levels of your database sequences in order to choose the correct settings for classification.

Setting up your database and query sequences

Each sequence in the database must have a unique name, and should be contained in a single folder in your Geneious database (subfolders within this folder are allowed). If only a single locus is used for classification, the locus name is not required.

Formatting sequences when multiple loci are used

If using multiple loci for classification, the locus name must be appended to the sequence name with a specific delimiter in both the query and database sequences. Geneious will use the sequence name before the delimiter to group database and query sequences from the same sample together.

For example, example you may have 4 database sequences named:

Acmopyle pancheri-matK

Acmopyle panceri-rbcL

Acmopyle sahniana-matK

Acmopyle sahniana-rbcL

Your query sequences could be named 'Query1-matK' and 'Query1-rbcL'. With multiple loci, this is treated as a database containing 2 sequences called 'Acmopyle pancheri' and 'Acmopyle sahniana' and a single query sequence called 'Query1'. When the analysis is run, matK and rbcL sequences are aligned separately, and then the alignments are concatenated according to their sequence name with ? used to replace any gaps between alignments.

Alignment options

Options for the pairwise alignments between query and database sequences are specified in the **Searching** panel. To choose the database you wish to align your query sequences to, click the button next to **Database folder** and navigate to the folder containing your database sequences.

The **Sensitivity** setting specifies the parameters that Geneious uses to align the query and database sequences. There is a trade-off between how fast the search runs versus how sensitive it is to finding distantly related matches. A higher sensitivity will align queries that are more distantly related to the database, so if you suspect your query sequence may be only distantly related to your database sequences, or are not sure whether it matches, you should use a higher sensitivity setting.

Classification settings

From each pairwise alignment, an 'Overlap Identity' percentage is calculated for the region in common between two sequences, where data in end gap regions is ignored. The overlap identity is used to classify sequences by picking the database sequence with the highest overlap identity to the query sequence. Parameters that can be set to classify sequences are as follows:

- **Minimum overlap identity to classify**: the query sequence will not be classified if the overlap identity between the query and best database match is below this setting.
- **Minimum % identity higher than next best result to classify**: If other database sequences have an overlap identity within this value of the identity with the closest match, then they will be considered to be equally good matches to the query. In this instance, if "classify using taxonomy" (see below) is not checked, then "Multiple matches" will be

returned as the classification. If ”classify using taxonomy” is checked, then the query sequence will be classified to the taxonomic level that these database sequences have in common.

E.g. for the example dataset given above with ”classify using taxonomy” checked, if this value is set on 0.2%, and the query sequence has 99.3% identity with Acmopyle pancheri and 99.2% identity with Acmopyle sahniana, then it will only be classified as ”Acmopyle” as the identities are within 0.2% of each other.

- **Classify using taxonomy from**: When this setting is checked, you can classify using structured taxonomic levels rather than just the sequence name. You can choose the Name, Description, Organism, or Taxonomy fields of the database sequences to classify by. The taxonomic levels will be delineated according to the **Taxonomic Level Separator** that you choose. If a query sequence matches multiple database sequences with different taxonomies, then the taxonomies are split via this character. The sequence is classified based on the split sections of the taxonomies that are in common, up to the point where they first differ (e.g a query that equally matches Acmophyle sahniana and Acmopyle pancheri will be classified as ”Acmophyle” if the taxonomic separator is a space”).

- **Minimum overlap identity to classify at lowest, second or third taxonomic level:** These settings allow you to set thresholds for classifying to a particular taxonomic level. The overlap identity of the best match to the query must be higher than the threshold for a particular taxonomic level for it to be classified to that level. If the overlap identity of the best result is lower than the threshold for a particular taxonomic level, then the result will be classified at a higher taxonomic level.

  (Note that a query may also be classified at a higher taxonomic level if multiple best matches are above the threshold, but the multiple matches only have a higher taxonomic level in common, as described above in the ”Minimum % identity higher than next best result to classify” section above).

  The lowest taxonomic level does not necessarily have to be ”species” - it could subspecies, locality, or genus, family etc. The classifier will simply interpret the field selected for taxonomic classification according to the taxonomic separator, and the lowest taxonomic level will be the final part of the taxonomy name. E.g. if you choose the sequence name field to classify your sequences by, with ”space” as the separator, and the sequence names are in the format ”Genus species locality”, then ”locality” will be regarded as the lowest taxonomic level, species as the second and genus as the third.

  You may find it useful to perform a multiple alignment of your database sequences and then look at the Distances tab to get an idea of appropriate identity thresholds for differing taxonomic levels.

Multiple loci

If you have sequences from multiple loci to use for classification, check the **Use Multiple Loci** box, and put the delimiter you have used to separate the sequence name and locus name in the box underneath (see ″Setting up your database...″ section). This option will concatenate individual pairwise alignments from different genes if the sequence names before the delimiter match.

When **Replace end gaps with '?' characters when concatenating alignments** is checked, then end gaps from individual pairwise alignments are treated as unknown characters in the concatenated alignment. If this option is unchecked, then the end gaps are regarded as real gaps in the concatenated alignment.

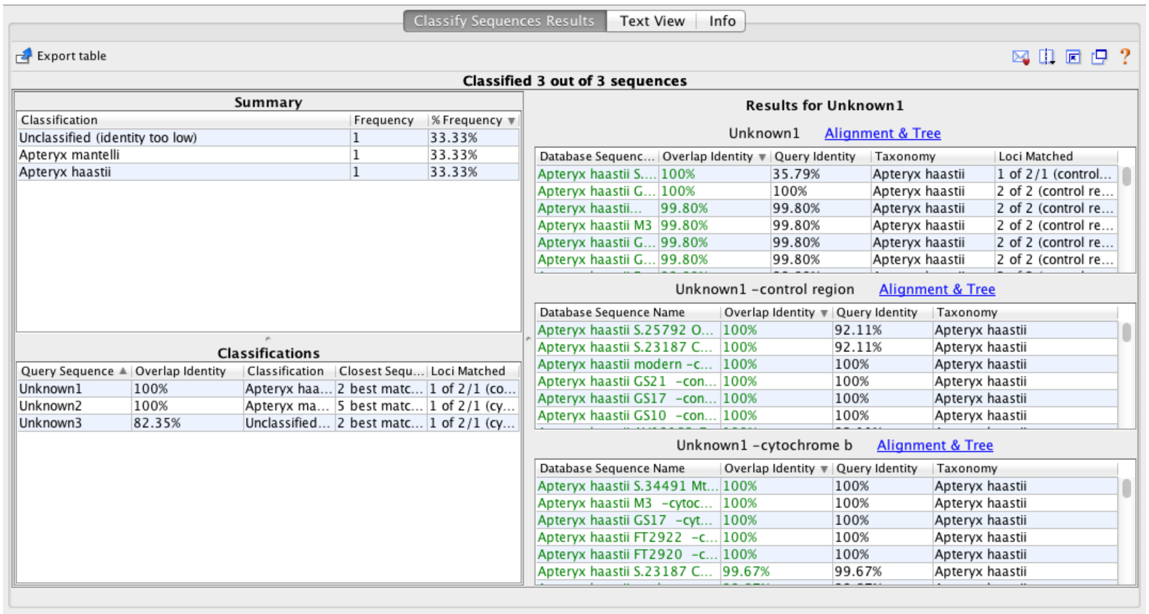
Results

Sequence classifier results are displayed as tables in the sequence viewer (see Figure 1). When multiple queries are run at once, **Summary** and **Classification** tables are produced. The **Summary** table shows how many query sequences are classified into each taxonomic group, and the **Classification** table shows the classification of each query sequence along with its identity to the closest database sequence.

Detailed **Identity** tables for each query can be produced by checking the option **Create table of all hit similarities for each query sequence**. Note that if the sequence classifier is run on a single query sequence, only Identity tables (not Summary or Classification tables) are produced and this option is greyed out.

Other result options can be configured as follows:

- **Save pairwise alignments**: Produces a pairwise alignment document for each query and database sequence and saves them in a separate folder. When a single query sequence is run, double-clicking on a particular row in the identity table will bring up the pairwise alignment for that result. We recommend unchecking this option when running multiple queries as it can potentially produce a large number of alignment documents which can impact performance.
- **Highlight results in green with minimum overlap identity**: Results that exceed the overlap identity entered here will be colored in green in the identity tables. Those below this value will be colored in red. When multiple loci are used, green results will instead be colored orange if all loci for that sequence are not green. However, the concatenated result will always be colored red if none of the loci exceed their green threshold no matter what the concatenated green threshold is.
- **Per Locus Minimums:** Allows you to specify a different minimum value for each locus



Classified 3 out of 3 sequences

Summary

Classification	Frequency	% Frequency
Unclassified (identity too low)	1	33.33%
Apteryx mantelli	1	33.33%
Apteryx haastii	1	33.33%

Classifications

Query Sequence	Overlap Identity	Classification	Closest Sequ...	Loci Matched
Unknown1	100%	Apteryx haa...	2 best matc...	1 of 2/1 (co...
Unknown2	100%	Apteryx ma...	5 best matc...	1 of 2/1 (cy...
Unknown3	82.35%	Unclassified...	2 best matc...	1 of 2/1 (cy...

Results for Unknown1

Unknown1 Alignment & Tree

Database Sequenc...	Overlap Identity	Query Identity	Taxonomy	Loci Matched
Apteryx haastii S....	100%	35.79%	Apteryx haastii	1 of 2/1 (control...
Apteryx haastii G...	100%	100%	Apteryx haastii	2 of 2 (control re...
Apteryx haastii...	99.80%	99.80%	Apteryx haastii	2 of 2 (control re...
Apteryx haastii M3	99.80%	99.80%	Apteryx haastii	2 of 2 (control re...
Apteryx haastii G...	99.80%	99.80%	Apteryx haastii	2 of 2 (control re...
Apteryx haastii G...	99.80%	99.80%	Apteryx haastii	2 of 2 (control re...

Unknown1 -control region Alignment & Tree

Database Sequence Name	Overlap Identity	Query Identity	Taxonomy
Apteryx haastii S.25792 O...	100%	92.11%	Apteryx haastii
Apteryx haastii S.23187 C...	100%	92.11%	Apteryx haastii
Apteryx haastii modern -c...	100%	100%	Apteryx haastii
Apteryx haastii GS21 -con...	100%	100%	Apteryx haastii
Apteryx haastii GS17 -con...	100%	100%	Apteryx haastii
Apteryx haastii GS10 -con...	100%	100%	Apteryx haastii

Unknown1 -cytochrome b Alignment & Tree

Database Sequence Name	Overlap Identity	Query Identity	Taxonomy
Apteryx haastii S.34491 Mt...	100%	100%	Apteryx haastii
Apteryx haastii M3 -cytoc...	100%	100%	Apteryx haastii
Apteryx haastii GS17 -cyt...	100%	100%	Apteryx haastii
Apteryx haastii FT2922 -c...	100%	100%	Apteryx haastii
Apteryx haastii FT2920 -c...	100%	100%	Apteryx haastii
Apteryx haastii S.23187 C...	99.67%	99.67%	Apteryx haastii

Figure 1: Results tables, showing the Summary and Classification tables on the left and Identity tables on the right. Identity tables are given for the overall result and for each gene.

in your database. If any locus has an overlap identity below its specified cut-off (or no match can be found with its corresponding database sequence), all loci for that sequence are excluded from the concatenated alignment and tree unless the option to include orange results is turned on. After building the concatenated alignment, any concatenated sequence which has a combined weighted overlap identity below the general ’minimum overlap for multiple alignments’ will also be excluded.

- **Save multiple alignment of all hits per query**: Creates a multiple alignment of all database hits for each query sequence. Options for this alignment (e.g. choice of alignment program) can be set by clicking the **Alignment Options** button. By default, only the green hits are included in the alignment (see Highlight results in green... above), but red and orange hits can be included by checking these additional options.

- **Save tree of all hits per query**: Builds a phylogenetic tree from the multiple alignment. Parameters for tree building can be set by clicking the **Tree Options** button.

- **Save alignments and trees in subfolder**: Creates a subfolder to save multiple alignments and trees into (recommended when running multiple queries)

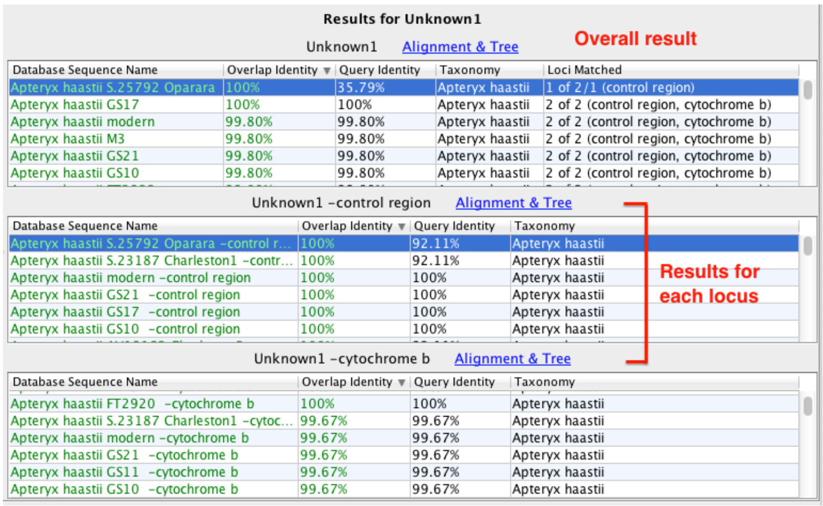
Note that when running the sequence classifier on a large number of query sequences, checking additional results options can cause the result document to become quite large and/or produce many result documents, which can impact performance.

Interpreting Identity tables

The top identity table shows the overall result, when alignments from multiple loci are concatenated. When multiple loci are used, individual identity tables are also provided for each gene (see Figure 2).

The columns in these tables are as follows:

- **Database sequence name:** The name of the database sequence matching your query
- **Overlap Identity:** The percentage identity across the region where database and query sequences overlap, where data in end-gap regions is ignored. In the overall result table produced with multiple loci, the Overlap Identity is a weighted average over the constituent loci, weighted according to the overlap length from each contributing locus. Loci that could not be aligned to the corresponding database sequence are assumed to have a completely mismatching overlap of length equal to the minimum overlap setting. However, if the database and query sequences do not contain some loci, these loci are excluded from the weighted overlap identity without assuming there is a completely mismatching region.
- **Query Identity**: The percentage identity over the entire length of the query sequence, where regions from the query that do not align to or are not covered by the database sequences are counted as mismatches. Thus, where the query identity is lower than the overlap identity it indicates that the query sequence extends outside the database sequence. The overall query identity will be lowered if not all loci have a database match for a particular sequence.
- **Taxonomy**: Taken from the field you specified under ”Classify using taxonomy from”.
- **Loci Matching**: Shows which loci in the database match the query when using multiple loci. E.g. In the example in Figure 2, two loci are used (cytochrome b and control region). The database sequence *Apteryx haastii S.25792 Oparara* does not have a cytochrome b sequence, so this entry reads ”1 of 2/1 (control region)”, indicating that there are 2 query sequences that could potentially have matches (as there is both control region and cytochrome b sequences for the query), and 1 database sequence that could have potentially matched (as there is only a control region sequence for that database sequence).



Results for Unknown1

Unknown1 Alignment & Tree

Database Sequence Name	Overlap Identity ▼	Query Identity	Taxonomy	Loci Matched
Apteryx haastii S.25792 Oparara	100%	35.79%	Apteryx haastii	1 of 2/1 (control region)
Apteryx haastii GS17	100%	100%	Apteryx haastii	2 of 2 (control region, cytochrome b)
Apteryx haastii modern	99.80%	99.80%	Apteryx haastii	2 of 2 (control region, cytochrome b)
Apteryx haastii M3	99.80%	99.80%	Apteryx haastii	2 of 2 (control region, cytochrome b)
Apteryx haastii GS21	99.80%	99.80%	Apteryx haastii	2 of 2 (control region, cytochrome b)
Apteryx haastii GS10	99.80%	99.80%	Apteryx haastii	2 of 2 (control region, cytochrome b)

Unknown1 -control region Alignment & Tree

Database Sequence Name	Overlap Identity ▼	Query Identity	Taxonomy
Apteryx haastii S.25792 Oparara -control r...	100%	92.11%	Apteryx haastii
Apteryx haastii S.23187 Charleston1 -contr...	100%	92.11%	Apteryx haastii
Apteryx haastii modern -control region	100%	100%	Apteryx haastii
Apteryx haastii GS21 -control region	100%	100%	Apteryx haastii
Apteryx haastii GS17 -control region	100%	100%	Apteryx haastii
Apteryx haastii GS10 -control region	100%	100%	Apteryx haastii

Unknown1 -cytochrome b Alignment & Tree

Database Sequence Name	Overlap Identity ▼	Query Identity	Taxonomy
Apteryx haastii FT2920 -cytochrome b	100%	100%	Apteryx haastii
Apteryx haastii S.23187 Charleston1 -cytoc...	99.67%	99.67%	Apteryx haastii
Apteryx haastii modern -cytochrome b	99.67%	99.67%	Apteryx haastii
Apteryx haastii GS21 -cytochrome b	99.67%	99.67%	Apteryx haastii
Apteryx haastii GS11 -cytochrome b	99.67%	99.67%	Apteryx haastii
Apteryx haastii GS10 -cytochrome b	99.67%	99.67%	Apteryx haastii

Figure 2: Identity tables produced by the sequence classifier