TD – Barcoding and integrative taxonomy – M. OLLIVIER

Bumble bees datataset analysis

In BOLD System (https://v4.boldsystems.org/), click on **LOGIN** (top right).

Username : *promoabsv31*, *Mdp : promoABSV31**

1- Click on the dataset **"DS-BBBABSV Dataset of Bumblebee species for a course with ABSV students"**.

What information appears on the screen? Can you describe the dataset? (Number of specimens, sequences, and species, origin, age of sequenced specimens, genetic marker, fragment length, quality, etc.)

2- Click on **"View all records"**.

What fields are included in the table that appears? Explain the meaning of each column. Why are some entries not associated with any sequence? What could be the possible reasons for this?

3- Select all sequences and construct a distance tree.
In the left menu, go to **Sequence Analysis > Taxon ID Tree**.

Enter the Following parameters:

Tree Type = Multipage Classic,
Sequence Data = Nucleotide,
Distance Model = Kimura 2 P,
Align Sequences = MUSCLE,
Select Branch Labels = Sample ID + Taxon + Sex + State/Province + Region + BIN URI
Apply filters = […] >200 bp
Colorize Tree Based on = Barcode Cluster (BIN)
Ambigous Base = Pairwise Deletion
Minimum Complete Overlap = 100 bp
Codon Positions Included = 1, 2, 3
Result Options = View immediately

- By observing the topology of this tree, can you identify specimens that may be misidentified? What do you base this assumption on? What should be done to confirm it?

List of possibly misidentified sequences: …

- By examining the tree topology, can you determine which species the specimens identified morphologically only at the genus level (Bombus sp.) might belong to?

List of specimens and suggested identification: …

4- Another way to verify the consistency of morphological identifications with genetic data is to perform a "**BLAST**" of the sequence, meaning an alignment of the sequence against all available data in the BOLD database.
To do this, go back to the "Records List", click on the sequence page of a specimen you suspect to be misidentified. Copy the sequence and paste it into BOLD’s "Identification Engine": https://v4.boldsystems.org/index.php/IDS_OpenIdEngine.

What are the results? Discuss them. You can also check the BIN page associated with an individual whose identification you doubt to confirm your observations.

5- Return to the **"Records List"**, uncheck the sequences associated with potential identification errors. Then, perform the **"Barcode Gap Analysis"** using the remaining sequences.

Enter the Following parameters:	Distance Model = Kimura 2 P,
	Marker = COI-5P,
	Alignment Options = MUSCLE,
	Apply Filters = […] > 200 bp,
	Ambigous Base = Pairwise Deletion
	Result Options = View immediately

- **Scatter Plots**: Take 5 minutes to carefully read the legend. In pairs, try to understand how the graph is constructed. What does the red line represent? How should it be interpreted?

- **Distance Distribution Histograms**: Explain the histograms presented.

- To differentiate two species based on their respective sequences, how should the genetic distance between these two species (interspecific distance) compare to the intraspecific distance for each species?

- What do the data you are analyzing show? Is this condition met for all Bombus sp. specimens sequenced here?

- **Details for Specimens Comparisons**: Use the data in this table to precisely identify problematic specimens. To clarify the situation, what further steps should be taken?