Taxonomical evaluation of plant genetic markers by Bayesian Classifier

Luisa Matiz¹, Alejandro Reyes², and Juan Manuel Anzola²

¹Affiliation not available

²Universidad de los Andes

November 27, 2020

Abstract

DNA barcodes are standardized sequences that range between 400-800 bp, vary at different taxonomic levels, and make it possible to identify individuals of species that have been previously assigned taxonomically. Several barcodes have been identified in different groups in the tree of life. However, there are groups that lack an accurate DNA marker, and even more so, accurate strategies that enable verification of their taxonomic affiliation. Several DNA barcodes have been postulated for plants, nonetheless, their classification potential has not been evaluated for metabarcoding, and as a result, it would appear as no one of them excels above the others in this area. One tool that has recently gained traction is Naïve Bayesian Classifiers; this type of classifier is based on the independence of attributes and the allocation of categories in each context. The present study aims at evaluating the classification power of several plant genetic markers that have been proposed as barcodes (*trnL*, *rpoB*, *rbcL*, *matK*, *psbA-trnH* and *psbK*) using a Naïve Bayesian Classifier, in order to determine the markers with higher performance at different taxonomic levels for metabarcoding analysis and to identify problematic genera at the time of species assignment. We propose *matK* and *trnL* as potential candidates up to the genus assignment. Some problematic genera (*Aegilops*, *Gueldenstaedtia*, *Helianthus*, *Oryza*, *Shorea*, *Thysananthus* and *Triticum*) within certain families in a sample could lead to misclassification no matter which marker is used. Finally, we propose recommendations when performing taxonomic identification analysis of plants in samples with multiple individuals.

Hosted file

24_11_20_Naive Bayesian Classifier_mainfile.pdf available at <https://authorea.com/users/379396/articles/495617-taxonomical-evaluation-of-plant-genetic-markers-by-bayesian-classifier>

Figure 1. Representation of the percentage of sequences classified by the Bayesian Classifier at species and genus level for 5 to 20 dataset. Shades of green represent the significant ones. Yellow represent misclassifications.

Figure 2. Markers performance by Bayesian Classifier from Phylum to Species based on the "good classification" criterion (the sequences that reach bootstrap score of at least 80) in (a) 5 to 20 dataset and (b) 2 to 20 data set. The y-axis for a) is 80-100% while for b) is 70-100%, this was done due to the number of data for every data set.

Figure 3. Correct and Incorrect classifications for Genus (G) and Species (S) levels for the markers used in the 5-20 dataset.

Figure 4. Accumulative F1 score categorical distribution for the 5-20 dataset at every marker. The percentage of assigned genus for every F1 score represents the accumulation until that number.