609

# High throughput classification of CO1 metabarcodes using a naïve Bayesian classifier

Porter, T.M.[1,2], Hajibabaei M.[2]

[1]Natural Resources Canada, Great Lakes Forestry Centre, Sault Ste. Marie, ON, P6A 2E5, CANADA
[2]Centre for Biodiversity Genomics & Department of Integrative Biology, University of Guelph, Guelph, ON, N1G 2W1, CANADA

## Abstract

**Background:** Groups that have used traditional biomonitoring methods for determining ecosystem status have started to incorporate CO1 metabarcoding in their workflows to facilitate monitoring in a more cost-effective and time-efficient manner. Until now, there has been difficulty with assigning names to CO1 partial barcodes in a rapid, high-throughput manner, while simultaneously providing a statistical measure of confidence for each assignment. **Results:** We have compiled a reference library of 912,253 CO1 sequences mined from the GenBank nucleotide database. This reference set can be used to classify chordates, arthropods, and flag other members of complex eDNA communities as belonging to other major eukaryote groups. We adopted the well-known taxonomic assignment tool, the naïve Bayesian classifier available from the Ribosomal Database Project, to enable high-throughput CO1 taxonomic assignments. We provide statistical support cutoff guidelines for CO1 fragments of different sizes. We also test the coverage and classification accuracy, *in silico*, of a variety of CO1 fragments generated from primers in the literature. We directly compare runtime and false positive rates generated from using the naïve Bayesian classifier or the commonly used top BLAST hit method. **Significance:** We show how the naïve Bayesian classifier can be used to analyze freshwater benthos communities detected by CO1 metabarcoding. We demonstrate the advantage of using a purpose-built taxonomic assignment tool over using the more general, but still widely used, top BLAST hit method to facilitate high throughput taxonomic assignments in a reasonable timeframe and to reduce rates of false positive assignments.

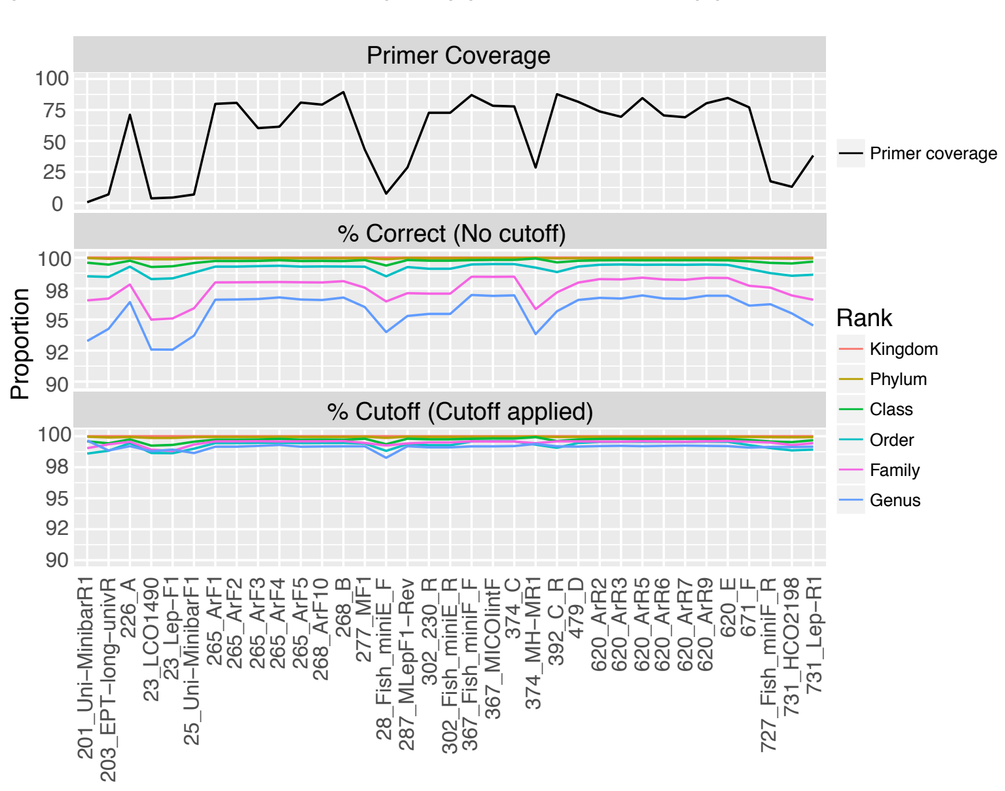
## Objectives

(1) Create a resource to permit high throughput taxonomic assignments for CO1 animal sequences
(2) Provide minimum bootstrap support cutoffs to reduce incorrect taxonomic assignments
(3) Compare the top BLAST hit method with the RDP classifier
(4) Demonstrate performance for taxa important for freshwater biomonitoring

## Methods

Arthropod and Chordate CO1 sequences that were 500 bp or longer were retrieved from GenBank [Oct 2016]. Other Eukaryote CO1 sequences annotated with the 'BARCODE' keyword were retrieved as outgroup sequences. The RDP classifier 2.12 was used to train the classifier and conduct leave-one-out testing (Wang et al., 2007): (1) CO1 Eukaryote v1 makes assignments to the genus rank, (2) CO1 Eukaryote v2 makes assignments to the species rank.

**Figure 1:** The proportion of correct taxonomic assignments is improved when a bootstrap support cutoff is applied.




## Discussion

A bottleneck when working with CO1 sequences from eDNA samples has been making automated, high throughput taxonomic assignments with an associated measure of confidence. We compiled a CO1 reference set of nearly 1 million sequences and show that the RDP classifier is faster than the top BLAST hit method. When bootstrap support is used as a cutoff, the proportion of incorrect assignments is reduced. Compared with BLAST, the RDP classifier reduces false positive assignments by two thirds.

The cost of a false positive taxonomic is high when it leads to an over-estimation of the presence or distribution of a rare threatened or endangered species or if it creates a false alarm for an invasive or harmful species. In such cases, the RDP classifier is a more reliable tool to use than BLAST.

References

Porter et al., 2014. Rapid and accurate taxonomic classification of insect (class Insecta) cytochrome c oxidase subunit 1 (COI) DNA barcode sequences using a naïve Bayesian classifier. *Mol. Ecol. Resour.* **14,** 929–942 (2014).

Wang et al., 2007. Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl. Environ. Microbiol. 73: 5261.

**Figure 2:** The RDP classifier with the CO1 Eukaryota v1 training set classifies significantly more queries per minute than the top BLAST hit method.

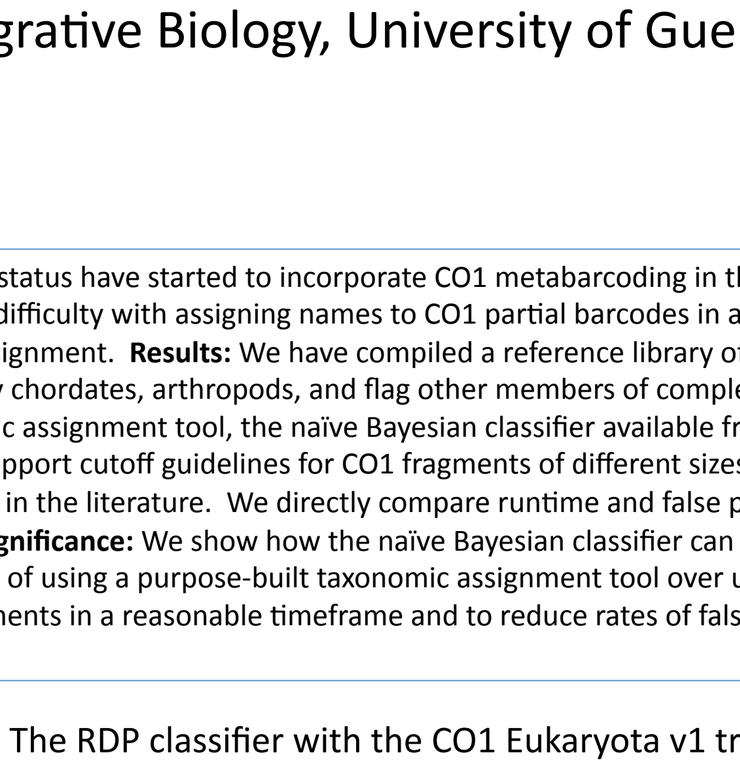



**Table 1:** The RDP classifier produces fewer false positives compared with the top BLAST hit method. The table summarizes taxonomic assignment outcomes from 200 bp primer-anchored CO1 sequences.

| Method | No. False Positives | False positive rate |
|---|---|---|
| Top BLAST hit | 397,820 | ~ 100% |
| RDP classifier | 117,457 | 34% |

**Table 2:** At a coarse scale, taxa important for freshwater biomonitoring are well represented in the CO1 Eukaryote v1 training set. Based on leave-one-out testing of full length barcode sequences, the proportion of incorrect assignments is low when a bootstrap support cutoff is applied.

| Taxa | No. Reference Sequences | % Incorrect (Cutoff applied) |
|---|---|---|
| Bivalvia | 667 | 0.3 |
| Clitellata | N/A | N/A |
| Gastropoda | 1,896 | 0.4 |
| Insecta_Coleoptera | 89,484 | 1.1 |
| Insecta_Diptera | 118,896 | 0.8 |
| Insecta_Ephemeroptera | 6,722 | 0.3 |
| Insecta_Megaloptera | 469 | 1.7 |
| Insecta_Odonata | 3,553 | 1.2 |
| Insecta_Plecoptera | 2,679 | 0.1 |
| Insecta_Trichoptera | 17,277 | 0.3 |
| Malacostraca_Amphipoda | 8,483 | 1.3 |
| Malacostraca_Isopoda | 3,659 | 0.1 |
| Polychaeta | 888 | 0.2 |
| Turbellaria | N/A | N/A |

Funding Sources

We would like to acknowledge funding for T. Porter from the Government of Canada through the Genomics Research Development Initiative as well as office space and computational resources provided by the Hajibabaei lab at the Centre for Biodiversity Genomics, University of Guelph.