

BIOMETRICS

Fall 2008

Assignment 5: DNA Barcoding – the Birds of North America

Due: November 13, 2008



Lab Description:

In this lab you will attempt to identify a North American bird species from an mtDNA sequence. To do this you will be given a file of labeled sequences (DNAtrainfile.txt) with which you can design your classifier. The mtDNA sequences are very discriminative (at least for NA birds), so you don't need much more than one sample per species. The DNA sequencing portion of the lab, i.e., the extraction of the feature vectors, has been done for you and provided by the Biodiversity Institute of Ontario.

For this lab you need to program ONE classifier. This classifier should be tested on a testing file (DNAtestfile.txt) of the same format – except the column in the test file that contains the label for the feature vector will be censored.

As for the previous two labs, you are to consider this lab as a friendly competition between you and the other class members. The winner will get a guaranteed increase in their letter grade by 1/3 – as explained in class.

What to Program:

You can use any of the classification algorithms explained to date or listed in the book. You can even design your own classifier if you like. Please use MATLAB to program and make sure to test your code in MATLAB before submission.

What to Hand In:

Students must hand in a zip file, which should be named "YourUNI_hw5.zip." For example, if one's UNI is xy1234, the filename should be xy1234_hw5.zip. Please send this zip file to jwgu@cs.columbia.edu by the due date.

1. A description of the experimentation done, including the logic behind your choice of classifiers.
2. An edited testfile, where the word **test** is replaced on each line by the results of your classifier, i.e. the bird specie name
3. An edited trainfile, where the true class is replaced on each line by the results of your classifier, i.e. the bird specie name
4. Please submit the MATLAB code as well(i.e. .m files)
5. If your code requires some special input-output, please give instructions of how to run code in a README file.

Additional Information:

1. There are 133 species in total. Some of the species have 2 or 3 training samples, but most of them have 1 training sample. Each row of DNAtrainfile.txt or DNAtestfile.txt corresponds to one sample.
2. Each sample has a DNA barcode of length 753. They are aligned. Missing data is the "NNNN..." at the beginning, and the "-----..." at the end.
3. If you are interested in the common names for the species, you can check out http://www.mumm.ac.be/~serge/birds/search_en.html. It has translations for some but not all. Or if you are interested in visual identification, see www.whatbird.com.