**Training the Ribosomal Database Project naïve Bayesian classifier**

1. Download and install a local copy of the rdp\_classifier from sourceforge.net/projects/rdp-classifier
2. Test your installation with the sample data provided with the installation, according to the instructions in the README file provided with the installation.
3. Instructions for *training* the classifier are available in the README file in the sampledata/ directory.
4. Download the ‘fasta’ and ‘taxonomy’ files for the GenBank-Genus training set from Dryad and follow the instruction provided in the README file from step 3 (above) starting at step 4 ‘Create parsed training files from the raw training data’. *The following text is only slightly modified from the README file provided with the rdp\_classifier\_2.5 installation, and is reproduced here for simplicity but the original documentation provides all the necessary information to train the classifier.*
   1. Assume the two raw files are created in mydir/mydata: GenBankGenus.taxonomy and GenBankGenus.fasta

Create a new directory to hold the trained data:

mkdir /PATH/mydata/mydata\_trained

Run the command to create parsed training files (all on one line):

java –Xmx1g –cp /PATH/rdp\_classifier-version.jar edu/msu/cme/rdp/classifier/train/ClassifierTraineeMaker /PATH/mydata/GenBankGenus.taxonomy /PATH/mydata/GenBankGenus.fasta 1 version1 test /PATH/mydata/mydata\_trained

The first parameter “mydata/GenBankGenus.taxonomy” contains the hierarchical taxonomy information.

The second parameter “mydata/GenBankGenus.fasta” contains the raw training sequences.

The third parameter “1” is the training set number to mark the training files generated.

The fourth parameter “version1” indicates the version of the hierarchical taxonomy.

The fifth parameter “test” holds the modification information of the taxonomy.

The sixth parameter “mydata\_trained” specifies the output directory.

Four parsed training files will be created and saved into the directory mydata\_trained/.

bergeyTrainingTree.xml

genus\_wordConditionalProbList.txt

logWordPrior.txt

wordConditionalProbIndexArr.txt

* 1. Copy the property file /PATH/sampledata/rRNAClassifier.properties to directory /PATH/mydata/mydata\_trained.
  2. Run the command line class to classify sequences. Supported formats are Fasta, GenBank, and EMBL.
     1. Use the default 16S training data to do the classification (all one line):

java –Xmx1g –jar /PATH/rdp\_classifier-version.jar –t /PATH/mydata/rRNAClassifier.properties –q /PATH/sampledata/testQuerySeq.fasta –o /PATH/testquery.out

* + 1. Use the user-provided GenBank-Genus training data to do the classification (all one line):

java –Xmx1g –jar /PATH/rdp\_classifier-version.jar –t /PATH/mydata/rRNAClassifier.properties –q /PATH/sampledata/testQuerySeq.fasta –o /PATH/testquery.out

The option –t specifies the training property file “mydata/rRNAClassifier.properties”.

The option –q specifies the query sequence file “sampledata/testQuerySeq.fasta”.

The option –o specifies the output file “testquery.out”.