rRDP: Interface to the RDP Classifier

Michael Hahsler
Southern Methodist University

Anurag Nagar
Southern Methodist University

Abstract

This package installs and interfaces the naive Bayesian classifier for 16S rRNA sequences developed by the Ribosomal Database Project (RDP). With this package the classifier trained with the standard training set can be used or a custom classifier can be trained.

Keywords: bioinformatics, Bioconductor, Biostrings, sequence classification.

1. Classification with RDP

The RDP classifier was developed by the Ribosomal Database Project which provides various tools and services to the scientific community for data related to 16S rRNA sequences. The classifier uses a Naive Bayesian approach to quickly and accurately classify sequences. The classifier uses 8-mer counts as features Wang, Garrity, Tiedje, and Cole (2007).

1.1. Using the RDP classifier trained with the default training set

RDP is shipped trained with a 16S rRNA training set. The model data is available in the data package **rRDPData**.

For the following example we load some test sequences shipped with the package.

```
R> library(rRDP)
R> seq <- readRNAStringSet(system.file("examples/RNA_example.fasta",
+         package="rRDP"))
R> seq

  A RNAStringSet instance of length 5
    width seq                                          names
[1]  1481 AGAGUUUGAUCCUGGCUC...AGUCGUAACAAGGUAACC 1675 AB015560.1 d...
[2]  1404 GCUGGCGGCAGGCCUAAC...UAAGGUCAGCGACUGGGG 4399 D14432.1 Rho...
[3]  1426 GGAAUGCUNAACACAUGC...GGUAGCCGUAGGGGAACC 4403 X72908.1 Ros...
[4]  1362 GCUGGCGGAAUGCUUAAC...UAGGUGUCUAGGCUAACC 4404 AF173825.1 A...
[5]  1458 AGAGUUUGAUUAUGGCUC...UCGUAACAAGGUAACCGU 4411 Y07647.2 Dre...
```

Note that the name contains the annotation from the FASTA file. In this case the annotation contains the actual classification information and is encoded in Greengenes format. For convenience, we replace the annotation with just the sequence id.

```
R> annotation <- names(seq)
R> names(seq) <-  sapply(strsplit(names(seq), " "), "[", 1)
R> seq
```

```
  A RNAStringSet instance of length 5
    width seq                                        names
[1]  1481 AGAGUUUGAUCCUGGCUC...AGUCGUAACAAGGUAACC 1675
[2]  1404 GCUGGCGGCAGGCCUAAC...UAAGGUCAGCGACUGGGG 4399
[3]  1426 GGAAUGCUNAACACAUGC...GGUAGCCGUAGGGGAACC 4403
[4]  1362 GCUGGCGGAAUGCUUAAC...UAGGUGUCUAGGCUAACC 4404
[5]  1458 AGAGUUUGAUUAUGGCUC...UCGUAACAAGGUAACCGU 4411
```

Next, we apply RDP with the default training set. Note that the data package **rRDPDate** needs to be installed!

```
R> pred <- predict(rdp(), seq)
R> pred
```

```
     rootrank   domain         phylum               class
1675     Root Bacteria Proteobacteria Deltaproteobacteria
4399     Root Bacteria Proteobacteria Alphaproteobacteria
4403     Root Bacteria Proteobacteria Alphaproteobacteria
4404     Root Bacteria Proteobacteria Alphaproteobacteria
4411     Root Bacteria Proteobacteria Alphaproteobacteria
                order            family       genus
1675             <NA>              <NA>        <NA>
4399 Rhodospirillales Rhodospirillaceae Rhodovibrio
4403 Rhodospirillales  Acetobacteraceae Roseococcus
4404 Rhodospirillales  Acetobacteraceae Roseococcus
4411 Rhodospirillales  Acetobacteraceae        <NA>
```

The prediction confidence is supplied as the attribute `"confidence"`.

```
R> attr(pred, "confidence")
```

```
     rootrank domain phylum class order family genus
1675        1      1   0.91  0.91  0.43   0.43  0.42
4399        1      1   1.00  1.00  1.00   1.00  1.00
4403        1      1   1.00  1.00  1.00   1.00  1.00
4404        1      1   1.00  1.00  1.00   1.00  1.00
4411        1      1   1.00  1.00  1.00   1.00  0.39
```

To evaluate the classification accuracy we can compare the known classification with the predictions. The known classification was stored in the FASTA file and encoded in Greengenes format. We can decode the annotation using `decode_Greengenes()`.

```
R> actual <- decode_Greengenes(annotation)
R> actual
```

```
   Kingdom         Phylum               Class             Order
1 Bacteria Proteobacteria Deltaproteobacteria Desulfobacterales
2 Bacteria Proteobacteria Alphaproteobacteria  Rhodospirillales
3 Bacteria Proteobacteria Alphaproteobacteria  Rhodospirillales
4 Bacteria Proteobacteria Alphaproteobacteria  Rhodospirillales
5 Bacteria Proteobacteria Alphaproteobacteria  Rhodospirillales
                          Family       Genus               Species
1                 Nitrospinaceae  Nitrospina               unknown
2              Rhodospirillaceae Rhodovibrio Rhodovibrio salinarum
3               Acetobacteraceae Roseococcus               unknown
4               Acetobacteraceae Roseococcus               unknown
5 Acetobacteraceae; Unclassified     unknown               unknown
   Otu
1 3187
2 2816
3 2785
4 2785
5 2752
                                                                Org_name
1                           AB015560.1_deep-sea_sediment_clone_BD4-10
2                         D14432.1_Rhodovibrio_salinarum_str._NCIMB2243
3 X72908.1_Roseococcus_thiosulfatophilus_str._RB-3_Yurkov_strain_Drews
4                                     AF173825.1_Antarctic_clone_LB3-94
5                           Y07647.2_Drentse_grassland_soil_clone_vii
    Id
1 1675
2 4399
3 4403
4 4404
5 4411
```

Now we can compare the prediction with the actual classification by creating a confusion table and calculating the classification accuracy. Here we do this at the Genus level.

```
R> confusionTable(actual, pred, rank="genus")
```

```
             predicted
actual        Nitrospina Rhodovibrio Roseococcus unknown <NA>
  Nitrospina           0           0           0       0    1
  Rhodovibrio          0           1           0       0    0
  Roseococcus          0           0           2       0    0
  unknown              0           0           0       0    1
  <NA>                 0           0           0       0    0
```

```
R> accuracy(actual, pred, rank="genus")
```

```
[1] 0.6
```

1.2. Training a custom RDP classifier

RDP can be trained using `trainRDP()`. We use an example of training data that is shipped with the package.

```
R> trainingSequences <- readDNAStringSet(
+     system.file("examples/trainingSequences.fasta", package="rRDP"))
R> trainingSequences
```

```
  A DNAStringSet instance of length 20
     width seq                                        names
 [1]  1384 TAGTGGCGGACGGGTGAG...TCGAATTTGGGTCAAGT 13652 Root;Bacter...
 [2]  1386 ATCTCACCTCTCAATAGC...CGAAGGTGGGGTTGGTG 13655 Root;Bacter...
 [3]  1440 ATCTCACCTCTCAATAGC...GCTGGATCACCTCCTTA 13661 Root;Bacter...
 [4]  1421 AATAGCGGCGGACGGGTG...ATCGGAAGGTGCGGCTG 13671 Root;Bacter...
 [5]  1439 ATCTCACCTCTCAATANC...GGTGCGGCTGGATCACC 13677 Root;Bacter...
 ...   ... ...
[16]  1478 TGGCTCAGGACGAACGCT...CGTATCGGAAGGTGCGG 13763 Root;Bacter...
[17]  1507 CCTGGCTCAGGACGAACG...TATCGGAAGGTGCGGCT 13781 Root;Bacter...
[18]  1481 TGGAGAGTTTGATCCTGG...GCAAGGATATAGCCGTC 13797 Root;Bacter...
[19]  1463 CGGCGTGCTTGGACCCAC...GGTCCTAAGGTGGGGGC 13799 Root;Bacter...
[20]  1389 CGAGTGGCAAACGGGTGA...GCAAGGATGCAGCCGTC 13800 Root;Bacter...
```

Note that the training data needs to have names in a specific RDP format:

```
"<ID> <Kingdom>;<Phylum>;<Class>;<Order>;<Family>;<Genus>"
```

In the following we show the name for the first sequence. We use here `sprintf` to display only the first 65 characters so the it fits into a single line.

```
R> sprintf(names(trainingSequences[1]), fmt="%.65s...")
```

```
[1] "13652 Root;Bacteria;Firmicutes;Clostridia;Clostridiales;Peptococc..."
```

Now, we can train a the classifier. The model is stored in a directory specified by the parameter `dir`.

```
R> customRDP <- trainRDP(trainingSequences, dir = "myRDP")
R> customRDP
```

```
RDPClassifier
Location: /tmp/RtmpbT6ZiU/Rbuild7eb56c72d06d/rRDP/vignettes/myRDP
```

```
R> testSequences <- readDNAStringSet(
+     system.file("examples/testSequences.fasta", package="rRDP"))
R> pred <- predict(customRDP, testSequences)
R> pred
```

```
      rootrank  Kingdom     Phylum      Class         Order
13811     Root Bacteria Firmicutes Clostridia Clostridiales
13813     Root Bacteria Firmicutes Clostridia Clostridiales
13678     Root Bacteria Firmicutes Clostridia Clostridiales
13755     Root Bacteria Firmicutes Clostridia Clostridiales
13661     Root Bacteria Firmicutes Clostridia Clostridiales
                                                    Family
13811                                       Veillonellaceae
13813                                       Veillonellaceae
13678                                        Peptococcaceae
13755 Thermoanaerobacterales Family III. Incertae Sedis
13661                                        Peptococcaceae
                    Genus
13811           Selenomonas
13813           Selenomonas
13678      Desulfotomaculum
13755 Thermoanaerobacterium
13661      Desulfotomaculum
```

Since the custom classifier is stored on disc it can be recalled anytime using `rdp()`.

```
R> customRDP <- rdp(dir = "myRDP")
```

To permanently remove the classifier use `removeRDP()`.

```
R> removeRDP(customRDP)
```

Acknowledgments

This research is supported by research grant no. R21HG005912 from the National Human Genome Research Institute (NHGRI / NIH).

References

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007). "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." *Applied and environmental microbiology*, **73**(16), 5261–5267.

Affiliation:

Michael Hahsler
Engineering Management, Information, and Systems
Lyle School of Engineering
Southern Methodist University
P.O. Box 750123
Dallas, TX 75275-0123
E-mail: mhahsler@lyle.smu.edu
URL: http://lyle.smu.edu/~mhahsler

Anurag Nagar
Computer Science and Engineering
Lyle School of Engineering
Southern Methodist University
P.O. Box 750122
Dallas, TX 75275-0122
E-mail: anagar@smu.edu