InvenireSRNA

Carl Tony Fakhry, Ping Chen, Rahul Kulkarni and Kourosh Zarringhalam

2017-08-23

Small RNAs (sRNAs) constitute an important class of post-transcriptional regulators that control critical cellular processes in bacteria. While recent research has led to a dramatic increase in the discovery of bacterial sRNAs, it is generally believed that the currently identified sRNAs constitute a limited subset of the bacterial sRNA repertoire. In several cases, sRNAs belonging to a specific class are already known and the challenge is to identify additional sRNAs belonging to the same class. **InvenireSRNA** is an R package for learning a classification model for a given class of sRNA thus allowing for the discovery of additional sRNAs beloning to the same class. **InvenireSRNA** also provides a pretrained model for predicting RsmA/CsrA regulating sRNAs.

Installation

Before installing this package, make sure you have the latest version of *Rstudio*, *R* and the *devtools* package. You also need to have C++11 available on your machine for the algorithm to run properly. The python package *Biopython* is also required for running the algorithm. Finally, you will need to install the ViennaRNA package.

Usage

The algorithm for the classification model was presented in [1]. We advise the user to read the referenced paper before using this package.

Predicting the Probability of a Sequence Being a CsrA regulating sRNAs

InvenireSRNA contains a pretrained model which computes the probability of an RNA sequence being an RsmA/CsrA regulating sRNA. The following example illustrates this functionality. Note that all sequences used in this example are RsmA/CsrA regulating sRNAs, so we expect them to have high probabilities.

```
library(InvenireSRNA)

# First, indentify a FASTA file
fname = paste(system.file(package = "InvenireSRNA"),
              "extdata/CsrAregulatingsRNAs_seed.txt", sep = "/")

# Second, compute the probability of the sequences in the FASTA
# file being a RsmA/CsrA regulating sRNA
probabilities = predict_fasta(fname)
head(probabilities)
```

```
##                     regions Mean_Probability Standard_Deviation
## 1 AAYV01000020.1/39993-39672        0.9447301       0.0070083610
## 2 CP000826.1/4199485-4199142        0.9970854       0.0011458173
## 3       AF076153.1/1958-2313        0.9977062       0.0007973336
## 4   AL627276.1/270587-270233        0.9974243       0.0012410845
## 5 CP000247.1/2917922-2917563        0.9998024       0.0001072025
## 6 CP000653.1/3538029-3537674        0.9910608       0.0030549092
```

Learning a Classification Model for Difference Classes of sRNAs

For other classes of sRNA, we can use *InvenireSRNA* to train a similar classification model. As an example, we will use seed sequences for RsmA/CsrA regulating sRNAs to train a model however these could have belonged to any target sRNA class. The algorithm makes use of the LASSO regression and a bootstrapping procedure, so the more sequences you train the model with the better the overall performance. However, providing too many sequences may significantly increase the training time. Note that the pretrained model for the RsmA/CsrA class attached to this package was trained with 105 sequences.

```
# First, identify a FASTA file
fname2 = paste(system.file(package = "InvenireSRNA"),
               "extdata/CsrAregulatingsRNAs_seed.txt", sep = "/")
```

We can now train a new model, which is really an ensemble of models as discussed in [1]. The ensemble of models is used to compute different predictions (i.e probabilities) and then the final predictions is an average over all the predictions in the ensemble of models. The parameter for the number of models to train is *n_models* for which the default value is *n_models = 100* which is more statistically valid than the small value of 2 we use in this illustrative example. The default number of bootstrap iterations to be used when training each model in the ensemble is 1000.

```
new_model = TrainModels(fname2, n_models = 2, bootstrap_iterations = 1000)

# Predict probabilities of the sequences using the new model
probabilities = predict_fasta(fname2, new_model)
head(probabilities)
```

Citation

Carl Tony Fakhry, Prajna Kulkarni, Ping Chen, Rahul Kulkarni and Kourosh Zarringhalam (2017). "Prediction of bacterial small RNAs in the RsmA (CsrA) and ToxT pathways: a machine learning approach." BMC Genomics, 18.