Theoretical and practical metagenomic approaches to viral discovery

Practical Session: Random Forest Classifier and viral application

Kevin Lamkiewicz, Manja Marz

23.10.2019

European Virus Bioinformatics Center

How to: Implement the Random Forest Classifier

Random Forest

```
1 from sklearn.ensemble import RandomForestClassifier
2 from sklearn import datasets
3 # we know this already...
4 iris = datasets.load_iris()
5 data = iris.data
6 target = iris.target
7
```

Random Forest

```
1  from sklearn.ensemble import RandomForestClassifier
2  from sklearn import datasets
3  # we know this already...
4  iris = datasets.load_iris()
5  data = iris.data
6  target = iris.target
7
8
9  # create the classifier object
10 # n_estimators: number of trees for the forest
11 # max_depth: maximum depth of one tree
12 rfc = RandomForestClassifier(n_estimators=100, max_depth=2)
13 rfc = rfc.fit(data, target)
```

Training our model to identify viral elements

We want to identify viral pre-miRNAs

Our task:

In the `miRBase` database are currently around 320 viral pre-miRNAs. We want to use them in to train a machine learning model that can distinguish between viral pre-miRNAs and other sequences.

WE WANT TO IDENTIFY VIRAL PRE-MIRNAS

Our task:

In the `miRBase` database are currently around 320 viral pre-miRNAs. We want to use them in to train a machine learning model that can distinguish between viral pre-miRNAs and other sequences.

I already prepared...

- a file with all precursors
- a file with our negative data

```
1 # first we need to translate the sequences into something
2 # that can be used by the machine learning algorithm:
3 # what are possible features for our task?
4
```

```
# first we need to translate the sequences into something
# that can be used by the machine learning algorithm:
# what are possible features for our task?

import numpy as np
```

COFFEE BREAK