**Create Filters from Model**

Input

../data/index/0test\_index.csv

../data/index/0train\_index.csv

../data/index/ 0val\_index.csv

../data/labels.csv

../data/sequences.csv

../data/model/ (has to have a model)

python genCreateFilters.py 1000 12 196 148 21 0 0 0

Sorted Sequences by class

../data/filters/labelsSorted.csv

../data/filters/sequencesSorted.csv

Print Filter Values

../data/filters/filter\_0.csv

…

../data/filters/filter\_11.csv

Print Max pooling Values

../data/filters/maxPool\_0.csv

…

../data/filters/maxPool\_11.csv

cd ../data/filters

python posPool.py

Get most important Positions

./posPool\_0

…

./posPool\_11

python createFeatVector.py

Translate into 21-bps features (posPool\_0)

./featsVector\_0.csv (3,837 features)

Matrix of 21-bps features

./dataDNAFeatures\_0.csv

python getFeatures.py

Reduce to non-repeated sequences

./features\_0.csv (3,827 features)

Create frequency matrix

./data\_0.csv

python getFreqMatrix.py

Create frequency matrix using “features\_0.csv” file

./data\_0.csv

**Create Model**

Input

../data/index/0test\_index.csv

../data/index/0train\_index.csv

../data/index/ 0val\_index.csv

../data/labels.csv

../data/sequences.csv

python genRunTrain.py 1000 12 196 148 21 0 0 0

1000 Evaluations

12 Filters

196 Fully Connected layer

148 window size

21 bps size

0 control variable keep as 0 used for EA

0 fold (0-9)

0 control variable keep as 0 used for EA

Output

../data/model/ (has to have a model)

../data/results/0\_1000\_12\_196\_148\_21.txt Screen information

../data/results/outputVector.txt Test Accuracy (1.00-testAccuracy)

../data/results/results0\_1000\_12\_196\_148\_21.txt Calculated Run ( for comparison )

../data/results/test0\_1000\_12\_196\_148\_21.txt Actual Answer ( for comparison )

./log3.txt Summary of results

0\_0\_1000\_1.0\_1.0\_1.0\_12\_196\_148\_21\_0

1. Generation for EA
2. index for EA

1000 evals

1.0 Test Accuracy

1.0 Val Accuracy

1.0 Train Accuracy

12 Filters

196 Fully Connected layer

148 window size

21 bps size

0 fold (0-9)

**Run Feature Selection**

**More information on this algorithm at**

**https://www.mdpi.com/2072-6694/12/7/1785/htm**  
 Input:

data\_0.csv frequency Matrix

features\_0.csv 21-bps feature vector

labels.csv labels with classes (0-4)

python aBioInf100.py

Output:

./src/run0/

…

./src/run9/

python summaryMulti.py

Output:

./src/best/cMatrix0.csv Confusion Matrix by classifier

…

./src/best/cMatrix7.csv Confusion Matrix by classifier

./src/best/data\_0.csv Reduced frequency Matrix with most important features

./src/best/features\_0.csv Reduced list of most important features

./src/best/results.txt Accuracy Results by classifier in 10-fold cross-validation

**Primer3Plus**

**Finally run candidate 21-bps features in** [**http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi**](http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) **or run** checkPrimer3Plus.py in **Primer3Plus folder**