PEAS Manual

Table of Contents

Dependencies

Feature Extraction Dependencies

Feature Extraction requires the following tools/commands to be installed and available:

JAVA (https://java.com/en/download/)
Command: java -jar

SAMTools (http://samtools.sourceforge.net/)
Command: samtools

MACS2 (https://github.com/taoliu/MACS)
Command: macs2

HOMER (http://homer.ucsd.edu/homer/)
Command: findMotifsGenome.pl
Command: annotatePeaks.pl

In addition install the human hg19 package (or a different package, see note) in HOMER.

Note: If necessary, the genome reference can be changed by modifying the following lines in the PEASFeatureExtraction.sh file, changing **hg19** to, for example, **hg38**:

```
annotatePeaks.pl "${prefix}_peaks.filtered" hg19 -m
"${homermotifs}" -nmotifs > ${prefix}_peaks_annotated.bed

annotatePeaks.pl "${prefix}_peaks.filtered" hg19 -m
"${outDir}/denovo/merge/merged.motifs" -nmotifs >
${prefix}_peaks_denovo.bed
```

In addition, determine the location of the reference genome fasta (.fa) file (or download to keep as future knowledge in order to run the feature extraction shell file.

PEASTools Source File Dependencies

PEASTools requires libraries from htsjdk.samtools
(https://github.com/samtools/htsjdk) and apache math commons(http://commons.apache.org/proper/commons-math/).

Model Training Dependencies

Model training and testing requires the following python libraries to be installed:

1. numpy
2. pandas
3. sklearn
4. matplotlib

Running Feature Extraction

To extract features, locate the PEASFeatureExtraction.sh shell file and execute the following command with 7 arguments:

```
<path to shell file>/PEASFeatureExtraction.sh <arg1>
<arg2> <arg3> <arg4> <arg5> <arg6> <arg7> <arg8>
```

arg1: The path to the paired-end bam file's directory without the training "/". Example: If the path is /user/documents/cd4t.bam, provide: **/user/documents**

arg2: The filename prefix (before .bam). Example: If the filename is cd4t.bam, provide **cd4t**

arg3: The full path to the fasta file location. Example: **/user/documents/hg19.fa**

arg4: The path to the bed file containing error prone regions of the genome to remove. This file is provided in the PEAS Github: https://github.com/UcarLab/PEAS. Example: **/user/documents/hg19.filter.bed**

arg5: The path to the motifs file. This file is provided in the PEAS Github: https://github.com/UcarLab/PEAS. Example: **/user/documents/humantop_Nov2016_HOMER.motifs**

arg6: The path to the conservation bed file. This file is provided in the PEAS Github: https://github.com/UcarLab/PEAS. Example: **/user/documents/phastCons46wayPlacental.bed**

arg7: The path to the CTCF motifs file. This file is provided in the PEAS Github: https://github.com/UcarLab/PEAS. Example: **/user/documents/CTCF.motifs**

arg8: The path to the PEASTools.jar file without the trailing "/". Example: If the PEASTools.jar file is located in /user/documents/peas/, provide **/user/documents/peas**

Example Feature Extraction Command:

```
/user/documents/peas/PEASFeatureExtraction.sh
/user/documents cd4t /user/documents/hg19.fa
/user/documents/hg19.filter.bed
/user/documents/humantop_Nov2016_HOMER.motifs
/user/documents/phastCons46wayPlacental.bed
/user/documents/CTCF.motifs
/user/documents/peas
```

Annotating Features

Features matrices can be annotated with class labels by using the PEASTools.jar file.

```
java -jar <path to jar>/PEASTools.jar annotate
<path to features file> <path to annotation file>
<output file path>
```

The annotation file must be a 4 column tab delimited (without a header) which includes the chromosome (example: “chr1”), start location, end location, and annotation (example: “Enhancer”).

Model Training & Prediction

Model training and prediction is achieved by running the PEASPredictor.py script:

```
python <local path>/PEASPredictor.py
<localpath>/trainfiles.txt
<localpath>/features.txt
<localpath>/classes_train.txt
<random state integer (for example: 929)>
<output directory for output files>
<localpath>/labelencoder.txt
<localpath>/testfiles.txt
<localpath>/classes_test.txt
```

Each of the files in the example above is available at https://github.com/UcarLab/PEAS.

trainfiles.txt
Tab delimited file where each line represents one feature file to use for model training. Features are merged using all files provided in this file.

Column 1: File label
Column 2: File path

features.txt
Tab delimited file specifying which columns in the training/test files are used in the model. Multiple lines can be provided to specify different intervals.

Column 1: Start index (inclusive)
Column 2: End index (exclusive)

For example if the start index is 3 and the end index is 29, all columns from 3 up to column 29 will be included. Indices start from 0.

classes_train.txt
Tab delimited file specifying the class label column and integer representation. One line is required for each class label, otherwise the class label is ignored. This has additional utility in making it easy to combine classes into one.

Column 1: Class label column index
Column 2: Label in feature file
Column 3: Integer label to convert to. (Must start from 0 and increase incrementally)

Indices start from 0.

labelencoder.txt
Used for specifying non-numeric features that need to be converted to integers. One line per column.

Column 1: Index column of the non-numeric feature
Column 2+: List of all non-numeric values. One column for each value.

Indices start from 0.

testfiles.txt
Tab delimited file where each line represents one feature file to use for model testing. Features are merged using all files provided in this file.

Column 1: File label
Column 2: File path

classes_test.txt
Tab delimited file specifying the class label column and integer representation. One line is required for each class label, otherwise the class label is ignored. This has additional utility in making it easy to combine classes into one.

Column 1: Class label column index
Column 2: Label in feature file
Column 3: Integer label to convert to. (Must start from 0 and increase incrementally)

Indices start from 0.

Output Files:

*_confusion.pdf – Confusion matrices of each test file.
ROC.pdf – Combined ROC curves for each test file.
PRC.pdf – Combined PRC curves for each test file.
*_predictions.txt – Prediction for each test file.

Prediction File (*_predictions.txt) Format:

Column 1: Chromosome
Column 2: Start
Column 3: End
Column 4: Integer *prediction* class label
Column 5: Integer *true* class label
Column 6+: Probability for each class label starting from 0

Note: it is assumed that the first three columns of training/test files are the positions of the peak files.

Data

Sample feature matrices are available in
https://github.com/UcarLab/PEAS/tree/master/data

Example results can be found in
https://github.com/UcarLab/PEAS/tree/master/example_results

Results generated were built combining CD4+ T, GM12878, CD14+, PBMC, and Islet16 features and tested on Islet6.