

The cleanUpdTSeq user's guide

Sarah Sheppard, Jianhong Ou, Nathan Lawson, Lihua Julie Zhu*

May 18, 2015

Contents

1	Introduction	1
2	step-by-step guide	2
2.1	Step 1. Load the package cleanUpdTSeq, read in the test dataset and then use the function BED2GRangesSeq to convert it to GRanges.	2
2.2	Step2. Build feature vectors for the classifier using the function buildFeatureVector.	3
2.3	Step 3. Load the training dataset and classify putative polyadenylation sites. . .	3
3	References	4
4	Session Info	4

1 Introduction

3' ends of transcripts have generally been poorly annotated. With the advent of deep sequencing, many methods have been developed to identify 3' ends. The majority of these methods use an oligo-dT primer, which can bind to internal adenine-rich sequences, and lead to artifactual identification of polyadenylation sites. Heuristic filtering methods rely on a certain number of adenines in the genomic sequence downstream of a putative polyadenylation site to remove internal priming events. We introduce a package to provide a robust method to classify putative polyadenylation sites. cleanUpdTSeq uses a naïve Bayes classifier, implemented through the *e1071* [1], and sequence features surrounding the putative polyadenylation sites for classification.

The package includes a training dataset constructed from 6 different Zebrafish sequencing dataset, and functions for fetching surrounding sequences using BSgenome [2], building feature vectors

*sarah.sheppard@umassmed.edu, julie.zhu@umassmed.edu

and classifying whether the putative polyadenylations site is a true polyadenylation site or a mis-primed false site.

A paper has been submitted to Bioinformatics and currently under revision [3].

2 step-by-step guide

Here is a step-by-step guide on using cleanUpdTSeq to classify a list of putative polyadenylation sites

2.1 Step 1. Load the package cleanUpdTSeq, read in the test dataset and then use the function BED2GRangesSeq to convert it to GRanges.

```
> library(cleanUpdTSeq)
> testFile <- system.file("extdata", "test.bed", package="cleanUpdTSeq")
> testSet <- read.table(testFile, sep="\t", header=TRUE)
> peaks <- BED2GRangesSeq(testSet, withSeq=FALSE)
```

If test dataset contains sequence information already, then use the following command instead.

```
> peaks <- BED2GRangesSeq(testSet, upstream.seq.ind=7,
+                           downstream.seq.ind=8, withSeq=TRUE)
```

To work with your own test dataset, please set testFile to the file path that contains the putative sites.

Here is how the test dataset look like.

```
> head(testSet)
```

	chr	start	stop	name	score	strand
1	chr10	2965327	2965327	6hpas-22249	1	-
2	chr10	2966558	2966558	6hpas-22250	1	-
3	chr10	2974251	2974251	6hpas-22251	2	-
4	chr10	2978441	2978441	6hpas-22252	1	-
5	chr11	16772291	16772291	6hpas-33204	1	-
6	chr11	16777848	16777848	6hpas-33205	1	-

	upstream	downstream
1	TCTTCATCATGGTCATCTCGCACCAGAGAGTGTGCCAGGG	CAGGAAGTTTTACCTGTCTGTCAATTATCGT
2	ACCCTGGTGAGGGTATAGAGCTGGTCCAGTGTGCCACGGC	AAAGAGGAAAACAGCATTGTTCCCTCCTGGA
3	TGATTTGTTTGTAAGTGAATTTATCTTTTAATAAAAAAGA	AAAAAGAAAGTCAAGCCAAGAGGCAAATAC
4	GGAGCGCGACCGCATCAACAAAATCTTGCAGGATTATCAG	AAGAAAAAGATGGTGAGTTATTATCATTCA

```

5 AGGGAAATAAATACAAAAGAATAAAAAATATGATTCATTGT AAGAAAAACACTTTAGCTACAAAAGTCCTT
6 ATTTAGTTGGGTATTATTTCAAATAAAGAGAGAGAGAGAC ACAAAAACACTACATCAAATTTGAGGACAAAA

```

2.2 Step2. Build feature vectors for the classifier using the function `buildFeatureVector`.

The zebrafish genome from BSgenome is used in this example for obtaining surrounding sequences. For a list of other genomes available through BSgenome, please refer to the BSgenome package documentation [2].

```

> testSet.NaiveBayes <- buildFeatureVector(peaks, BSgenomeName=Drerio,
+                                         upstream=40, downstream=30,
+                                         wordSize=6, alphabet=c("ACGT"),
+                                         sampleType="unknown",
+                                         replaceNAdistance=30,
+                                         method="NaiveBayes",
+                                         ZeroBasedIndex=1, fetchSeq=TRUE)

```

If sequences are present in the test dataset already, then set `fetchSeq=FALSE`.

2.3 Step 3. Load the training dataset and classify putative polyadenylation sites.

The output file is a tab-delimited file containing the name of the putative polyadenylation sites, the probability that the putative polyadenylation site is false/oligodT internally primed, the probability the putative polyadenylation site is true, the predicted class based on the assignment cutoff and the sequence surrounding the putative polyadenylation site.

```

> data(data.NaiveBayes)
> if(interactive()){
+   predictTestSet(data.NaiveBayes$Negative, data.NaiveBayes$Positive,
+                 testSet.NaiveBayes=testSet.NaiveBayes,
+                 outputFile="test-predNaiveBayes.tsv",
+                 assignmentCutoff=0.5)
+ }

```

Alternatively, instead of passing in a positive and a negative training dataset, set the parameter classifier to a pre-built *PASclassifier* to speed up the process. To build a *PASclassifier* using the training dataset, please use function `buildClassifier`. A *PASclassifier* named as `classifier` is included in the package which is generated using the included training dataset with `upstream=40`, `downstream=30`, and `wordSize=6`. Please note that in order to use this pre-built classifier, you need

to build feature vector using `buildFeatureVector` from your test dataset with the same setting, i.e., `upstream=40`, `downstream=30`, and `wordSize=6`.

```
> data(classifier)
> testResults <- predictTestSet(testSet.NaiveBayes=testSet.NaiveBayes,
+                               classifier=classifier,
+                               outputFile=NULL,
+                               assignmentCutoff=0.5)
> head(testResults)
```

	PeakName	prob False/oligo	dT internally primed	prob True	pred.class
1	6hpas-78439	9.999997e-01	3.137403e-07	0	
2	6hpas-78440	1.000000e+00	7.431591e-14	0	
3	6hpas-78441	4.474675e-06	9.999955e-01	1	
4	6hpas-78442	3.085156e-07	9.999997e-01	1	
5	6hpas-22249	9.899582e-01	1.004181e-02	0	
6	6hpas-22250	1.000000e+00	1.636415e-09	0	

	UpstreamSeq	DownstreamSeq
1	GGTCATTGTCCTGCAAAATGGACTACTTAACCGAACTGGA	GAAGTATAAGAAGTAAGTACATTAAAGCTAC
2	TGGATTTAAATAACAAACAAGTTAAATAAAACGATTTGTA	AAAAAATAAAACAACCTGAAGAAGAAAATGAA
3	ATCTGCTTCAAAATGGATGCTCTGTTGAATCCTGAGCTCA	GGTAATCTTTCAAGTGCTGCTATTGAGCCAA
4	AAATGCTTGCACATAATAAATGTAGGCTTAAAGATTTC	AAACGTTTGTGAGAGACGGATTTTACTTTGC
5	TCTTCATCATGGTCATCTCGCACCAGAGAGTGTGCCAGGG	CAGGAAGTTTTACCTGTCTGTCATTATCGTC
6	ACCCTGGTGAGGGTATAGAGCTGGTCCAGTGTGCCACGGC	AAAGAGGAAAACAGCATTGTTTCCTCCTGGAT

3 References

1. Meyer, D., et al., e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. 2012.
2. Pages, H., BSgenome: Infrastructure for Biostrings-based genome data packages.
3. Sarah Sheppard, Nathan D. Lawson, and Lihua Julie Zhu. 2013. Accurate identification of polyadenylation sites from 3' end deep sequencing using a naïve Bayes classifier. Bioinformatics. Under revision

4 Session Info

```
> toLatex(sessionInfo())
```

- R version 3.2.0 Patched (2015-04-19 r68206), x86_64-apple-darwin10.8.0
- Locale: en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, stats4, utils
- Other packages: ade4 1.7-2, BiocGenerics 0.14.0, Biostrings 2.36.1, BSgenome 1.36.0, BSgenome.Drerio.UCSC.danRer7 1.4.0, cleanUpdTSeq 1.6.1, e1071 1.6-4, GenomeInfoDb 1.4.0, GenomicRanges 1.20.3, IRanges 2.2.1, rtracklayer 1.28.2, S4Vectors 0.6.0, seqinr 3.1-3, XVector 0.8.0
- Loaded via a namespace (and not attached): BiocParallel 1.2.1, BiocStyle 1.6.0, bitops 1.0-6, class 7.3-12, futile.logger 1.4.1, futile.options 1.0.0, GenomicAlignments 1.4.1, lambda.r 1.1.7, RCurl 1.95-4.6, Rsamtools 1.20.2, tools 3.2.0, XML 3.98-1.1, zlibbioc 1.14.0