How to use the ASEB Package

Likun Wang[1,2] and Tingting Li[1,3]

October 17, 2016

[1]Institute of Systems Biomedicine, Peking University Health Science Center.
[2]College of Computer Science and Technology, Jilin University.
[3]Department of Biomedical Informatics, Peking University Health Science Center.

wang.likun@gmail.com

Contents

1 Introduction

Lysine acetylation is a well-studied posttranslational modification on kinds of proteins. About four thousand lysine acetylation sites and over 20 lysine (K)-acetyl-transferases (KATs) have been identified. However, which KAT is responsible for a given protein or lysine site acetylation is mostly unknown. In our previous study, we found that different KAT families acetylate lysine sites with different sequence features (**?**). Based on these differences, we developed a computer program, Acetylation Set Enrichment Based (ASEB) method to predict which KAT-families are responsible for acetylation of a given protein or lysine site.

2 Getting started

To load the *ASEB* package, type `library(ASEB)`. Total six methods are presented in this package. They are `readSequence`, `asebSites`, `asebProteins`, `drawStat` and `drawEScurve`.

3 Methods

In this package, we use a GSEA-like method to make predictions. Gene Set Enrichment Analysis (GSEA) method was developed and successfully used to detect coordinated expression changes (**??**) and find the putative functions of the long non-coding RNAs (**?**). In our study (**?**), we treated the (acetylated) lysine sites and their surrounding amino acids (8 on each side) as (acetylated) peptide sequences. We first define all the validated acetylated peptide sequences from one KAT family as a KAT-specific set, and define 10000 random selected peptide sequences from the whole proteome as a background set. When given a

new query peptide sequence, similarity scores are calculated according to the BLOSUM62 matrix between this query peptide and peptides in the KAT-specific set and background set. A list is then created by ranking the scores. Similar with GSEA method, a running sum score (enrichment score) was calculated by walking down the list. To estimate the significance of the enrichment score for a query peptide, at first, a certain number of peptide sets with the same size as KAT-specific set was randomly generated. Secondly, treating each randomly generated set as a pseudo predefined KAT-specific set, an enrichment score can be calculated for each randomly generated set. At last, rank all the enrichment scores (from high to low) and a nominal P-value could be calculated. The nominal P-value is defined as the rank of enrichment score for the KAT-specific set divided by the total number of random selected sets. The whole process is similar with the GSEA method (permuting gene sets). Please see **?** for details.

4 Data

We provide the KAT-specific set for CBP/P300 and GCN5/PCAF family in this package. The file predefined_sites.fa under extdata contains the KAT-specific set for CBP/P300 family (total 267 sites). While the file predefined_sites2.fa contains the KAT-specific set for GCN5/PCAF family (total 82 sites). The two sets were generated by searching the PubMed literature. The file background_sites.fa contains 10000 randomly selected sites and the KAT-specific set for CBP/P300 family (total 10000+267 sites). While the file background_sites2.fa contains 10000 randomly selected sites and the KAT-specific set for GCN5/PCAF family (total 10000+82 sites).

5 Examples

5.1 Example for readSequence

This function return an object of `SequenceInfo` that contains sequences and identifiers from FASTA format input file.

```
> library(ASEB)
> ff <- system.file("extdata", "background_sites.fa", package="ASEB")
> readSequence(ff)
```

```
object of SequenceInfo
total 10267 sequences
first 10 sequences
Slot "ids":
 [1] "O60313_K698" "Q96H12_K26"  "P35227_K233" "Q9BX69_K747" "Q5T619_K231"
 [6] "Q9NYD6_K324" "Q08043_K227" "Q9Y305_K346" "Q5XPI4_K813" "Q9Y3Q7_K370"
Slot "sequences":
 [1] "KEHDDIFDKLKEAVKEE" "SILLALVEKYKYVLECK" "KYRVQPACKRLTLATVP"
 [4] "GNFNHVSLKASWVMGRP" "LAKARNSRKVQNQAGRR" "QNRRMKLKKMNRENRIR"
 [7] "NTAFEVAEKYLDIPKML" "AVDDIMFQKPVEVGSLL" "KSQKVFSEKLDHLSRRL"
[10] "HDYRYFVSKFETKCLQK"
```

5.2 Example for asebSites

This function is used to predict lysine sites that can be acetylated by a specific KAT-family.

```
>   backgroundSites <- readSequence(system.file("extdata", "background_sites.fa", package="
>   prodefinedSites <- readSequence(system.file("extdata", "predefined_sites.fa", package="
>   testSites <- readSequence(system.file("extdata", "sites_to_test.fa", package="ASEB"))
>   resultList <- asebSites(backgroundSites, prodefinedSites, testSites, permutationTimes=1
```

```
Please wait patiently!
background sites: /tmp/RtmpDNkEGB/backgroundSitesFile
predefined sites: /tmp/RtmpDNkEGB/prodefinedSitesFile
sites to test: /tmp/RtmpDNkEGB/testSitesFile
output: /tmp/RtmpDNkEGB/outputFile
Permutation times: 100

processed 2 sites
```

```
>   resultList$results[1:2,]
```

```
         site       ES p-value
1 A0AVK6_K550 0.004100    0.82
2 A0AVK6_K555 0.111549    0.01
```

This method can also process FASTA format files directly without loading all the sequences to the workspace of R. In this case, it can process huge number of lysine sites each time.

```
>   backgroundSitesFile <- system.file("extdata", "background_sites.fa", package="ASEB")
>   prodefinedSitesFile <- system.file("extdata", "predefined_sites.fa", package="ASEB")
>   testSitesFile <- system.file("extdata", "sites_to_test.fa", package="ASEB")
>   asebSites(backgroundSitesFile, prodefinedSitesFile, testSitesFile, permutationTimes=100
```

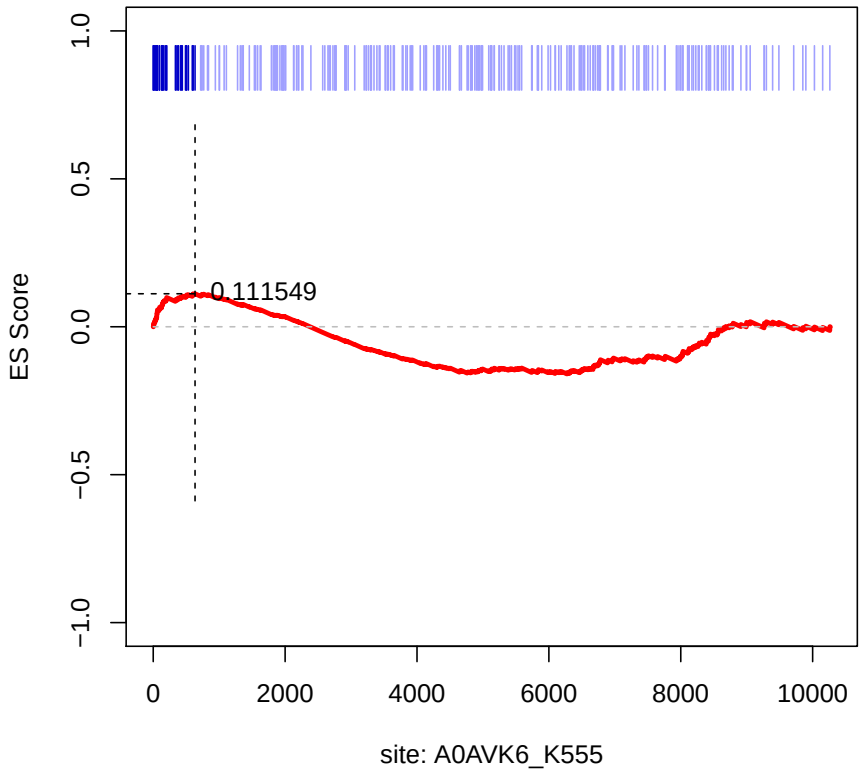
```
Please wait patiently!
background sites: /private/tmp/RtmphtPPC4/Rinst14f87501135e4/ASEB/extdata/background_sites.
predefined sites: /private/tmp/RtmphtPPC4/Rinst14f87501135e4/ASEB/extdata/predefined_sites.
sites to test: /private/tmp/RtmphtPPC4/Rinst14f87501135e4/ASEB/extdata/sites_to_test.fa
output: /tmp/RtmpDNkEGB/outputFile
Permutation times: 100

processed 2 sites
```

5.3 Example for drawEScurve

This method can be used to draw enrichment score curve for a specific site. Please see **?** for details about enrichment score curve.

```
>   drawEScurve(resultList$curveInfo, max_p_value=0.1, min_es=0.1)
```



These curves show running-sum process for calculating enrichment score. Users can find detail algorithm from **?**. The data.frame object contains curve information is given by `asebSites`, or `asebProteins`.

5.4 Example for asebProteins

This function is used to predict all lysine sites on a specific protein that can be acetylated by a specific KAT-family.

```
> backgroundSites <- readSequence(system.file("extdata", "background_sites.fa", package="
> prodefinedSites <- readSequence(system.file("extdata", "predefined_sites.fa", package="
> testProteins <- readSequence(system.file("extdata", "proteins_to_test.fa", package="ASE
> resultList <- asebProteins(backgroundSites, prodefinedSites, testProteins, permutationT
> resultList$results[1:2,]
```

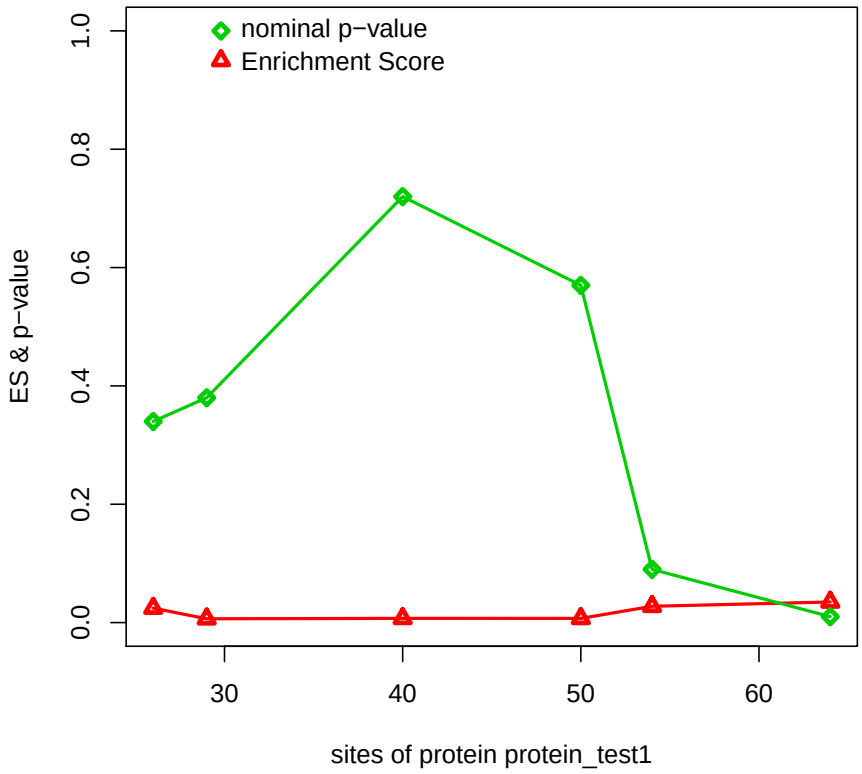
For processing huge number of lysine sites on proteins, this method can be used as below.

```
> backgroundSitesFile <- system.file("extdata", "background_sites.fa", package="ASEB")
> prodefinedSitesFile <- system.file("extdata", "predefined_sites.fa", package="ASEB")
> testProteinsFile <- system.file("extdata", "sites_to_test.fa", package="ASEB")
> asebProteins(backgroundSitesFile, prodefinedSitesFile, testProteinsFile, permutationTim
```

5.5 Example for drawStat

This function is used to show P-values and enrichment scores for all lysine sites on a specific protein. The X-axis shows positions of all lysine sites on a specific protein, and Y-axis shows the enrichment scores (0 1) and P-values (0 1) for each lysine site.

```
> drawStat(curveInfoDataFrame=resultList$curveInfo);
```



The sites with less P-values are more significant. For sites that have similar P-values, the ones with higher enrichment scores are more likely to be acetylated. These P-values are nominal (**??**), hence it is hard to give an threshold to predict significant sites. However, users can order all the sites and pay more attention to the ones with relatively less P-values. Please see **?** for details.