Genome-wide assessment of differential translations with ribosome profiling data
– the xtail package

Zhengtao Xiao[1−3], Qin Zou[1,3], Yu Liu[1−3], and Xuerui Yang[1−3]

[1]MOE Key Laboratory of Bioinformatics,
[2]Tsinghua-Peking Joint Center for Life Sciences,
[3]School of Life Sciences, Tsinghua University, Beijing 100084, China.

July 6, 2015

1 Introduction

This package is for identifing genes undergoing differential translation with ribosome profiling data. A simple assumption of our method is that for a gene undergoing translational dyresgulation under certain experimental or physiological condition, the change of its RPF abundance should be discoordinated with the change of its mRNA abundance. Based on the this assumptions, Xtail estimates log2 fold changes (log2FC) of RPF and mRNA accross two conditions, separately, and uses ratio of these two fold changes (ratio of fold changes, ROF) as the magnitude of differential translation. It also estimates log2 ratio (log2R) of RPF over mRNA in two conditions, this is mathematically equal to log2FC, and then taking fold change of these two ratios (fold change of ratios, FOR) across two conditions. Then, Xtail tests for each gene whether the log2FC of RPF and log2FC of mRNA differ significantly, and also compares the log2R of RPF over mRNA between the two conditions. Finally, the two P-values by comparing log2FC and log2R are return, and larger one was select as the final P-values for differential translations.

The general outline of Xtail is illustrated in Figure 1. First, we implement the negative binomial (NB) distribution as the model for count variability of both mRNA and RPF. We adapt the strategy of DESeq2 [1] to normalize read counts of mRNA and RPF in all samples, and fit NB distributions with dispersions α and μ. Then, with the NB distributions of RPF and mRNA estabished in both conditions, we derived probability density distributions of the log2FC of mRNA and RPF, by comparing two experimental groups. Similarly, probability density distributions of the log2R of RPF over mRNA were also derived in each of the two experimental groups. For each gene, Xtail tests wherther log2FC of RPF and log2FC of mRNA different significntly. This is done by crossing the probability density distributions of log2FC for mRNA and RPF to generate a joint probability matrix, and taking summation of the upper or lower triangle, whichever the smaller, of the matrix as the P-value. We also compare two logR of RPF over mRNA between the two conditions. With the similar method, a P-value of two probability distributions of log2R is also return. Finally, the larger P-value is selected as the final P-value for differential translations. The overlap of two log2FC distribution of RPF and mRNA, OVL, is also used to measure similarity of RPF and mRNA changes. Smaller overlaps suggest uncoordinated changes of RPF and mRNA across different conditions, supporting differential translations.

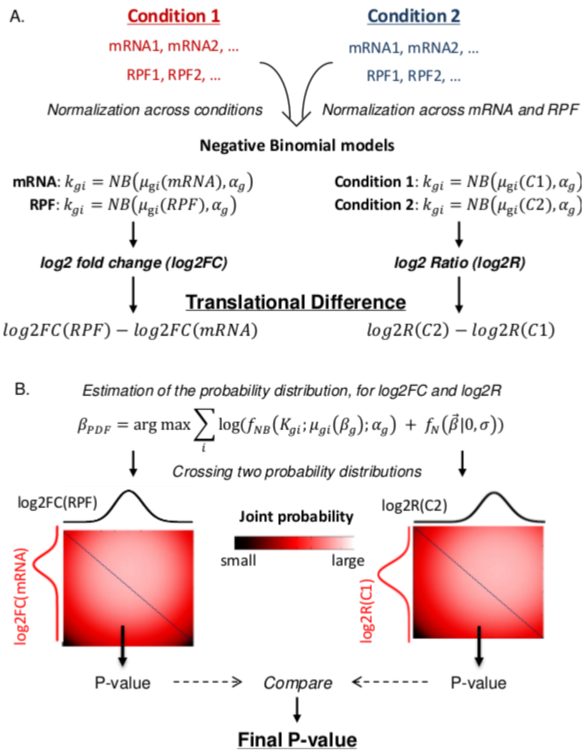
A.
Condition 1
mRNA1, mRNA2, ...
RPF1, RPF2, ...
Condition 2
mRNA1, mRNA2, ...
RPF1, RPF2, ...
Normalization across conditions
Normalization across mRNA and RPF
Negative Binomial models
mRNA: $k_{gi} = NB(\mu_{gi}(mRNA), \alpha_g)$
RPF: $k_{gi} = NB(\mu_{gi}(RPF), \alpha_g)$
Condition 1: $k_{gi} = NB(\mu_{gi}(C1), \alpha_g)$
Condition 2: $k_{gi} = NB(\mu_{gi}(C2), \alpha_g)$
log2 fold change (log2FC)
log2 Ratio (log2R)
Translational Difference
$log2FC(RPF) - log2FC(mRNA)$
$log2R(C2) - log2R(C1)$
B.
Estimation of the probability distribution, for log2FC and log2R
$\beta_{PDF} = \arg\max \sum_i \log(f_{NB}(K_{gi}; \mu_{gi}(\beta_g); \alpha_g) + f_N(\vec{\beta}|0,\sigma))$
Crossing two probability distributions
log2FC(RPF)
log2FC(mRNA)
Joint probability
small
large
log2R(C2)
log2R(C1)
P-value
Compare
P-value
Final P-value

Figure 1: **Schematic description of Xtail.** Xtail is designed to identify differential translations from mRNA and RPF read counts across two conditions, condition 1(C1) and 2(C2). By comparing log2 fold change (log2FC, left) of RPF and mRNA, or log2 ratios (log2R, right) across conditons, Xtail infers the magnitude of differential translations (A) and evaluates the statistical significance with P-value (B).

2 Data Preparation

As input, the `Xtail` package uses raw read counts of RPF and mRNA, in the form of rectangular table of integer values. The rows and the columns of the table correspond the genes and samples. Each cell in the *i-th* row and the *j-th* columns is the count number of reads mapped to gene *i* in sample *j*. The mRNA count data and RPF count data are stored in two text files. We can read them in using **R**'s standard function. For example,

```
mrna <- read.table("mRNA_counts.txt", header=TRUE, row.names=1)
rpf <- read.table("RPF_counts.txt", header=TRUE,row.names=1)
```

Here, header=TRUE indicates that the first line contains columns names and row.names=1 means that the first column should be used as row names (genes names).

In this vignette, we selected a published ribosome profiling dataset from yeast after amino acid starvation [2]. This dataset consisting of 4952 genes from two conditions ("rich" vs. "starvation") with each condition having two replicates. We can load this dataset by,

```
data(testdata)
```

3 An Example

Here we perform an analysis on the ribosome profiling data described above. First we load the library and data.

```
library(xtail)
data(testdata)
```

Next we can have a look at the first five lines of the mRNA (`mrna`) and RPF (`rpf`) elements of `testdata`.

```
mrna <- testdata$mrna
rpf <- testdata$rpf
head(mrna,5)
```

```
##         rich1 rich2 starvation1 starvation2
## LSR1     1294  3548        1592        3223
## NME1       72   151          33          90
## R0020C   1732  3596         650        2110
## R0030W    290   891         135         431
## R0040C    948  1896         243         895
```

```
head(rpf,5)
```

```
##         rich1 rich2 starvation1 starvation2
## LSR1       29    40          86         112
## NME1     1503  1853        1344        1741
## R0020C    176   263         113         413
## R0030W     29    55          23          88
## R0040C    223   401          89         358
```

We assign condition labels corresponding to the columns of the mRNA and RPF inputs.

```
condition <- c("rich","rich","starvation","starvation")
```

We run the main function, `xtail()`. By default, the `baseLevel` is set to the first condition (here is "rich"). This control the log2FC or log2R as the expected comparison against the baseLevel. The argument `minMeanCount` is set to 10. This is the minimum average expression of mRNA counts and RPF counts acrossing all experiments for a gene to be used. The argument threadsNo is the number of CPU cores for using. By default, `threadsNo` is set to 'NA', so the `detectCores` function in `parallel` library is used to determine the available cores. The argument `bins` is the number of bins used for calculate the probability densities of log2FC and log2R. This paramater will determine how accurate the final pvalue. Here, in order to keep the run-time of this vignette small, we will set `bins` to '1000'.

```
test.results <- xtail(mrna,rpf,condition,bins=1000)
```

```
## 1.  Estimate the log2 fold change in mrna
## normalize counts by size factors
## estimate dispersion paramater of NB model
## estimate the log2 fold change
## 2.  Estimate the log2 fold change in rpf
## normalize counts by size factors
## estimate dispersion paramater of NB model
## estimate the log2 fold change
## 3.  Estimate the log2FC difference between mrna and rpf
## 4.  Estimate the log2 ratio in first condition
## normalize counts by size factors
## estimate dispersion paramater of NB model
## estimate the log2 fold change
## 5.  Estimate the log2 ratio in second condition
## normalize counts by size factors
## estimate dispersion paramater of NB model
## estimate the log2 fold change
## 6.  Estimate the log2R difference between two conditions
```

Now we can examine the first five lines of the results produced by the `xtail()` run.

```
head(test.results,5)
```

```
##         mRNA_log2FC RPF_log2FC   OVL_log2FC pvalue_log2FC rich_log2R starvation_log2R
## LSR1      0.6429113  2.0213945 0.0007839313  2.985325e-06  -4.862788       -3.0358339
## NME1     -0.3503434  0.4244706 0.1998081144  7.585921e-02   5.014454        5.2420773
## R0020C   -0.5098046  0.5134972 0.0010729626  3.635477e-06  -2.461468       -1.3746036
## R0030W   -0.5066380  0.6747171 0.0086142957  2.558489e-04  -2.575015       -1.3102852
## R0040C   -0.9155999 -0.2121955 0.0366386801  3.033306e-03  -1.098382       -0.3403117
##           OVL_log2R pvalue_log2R final_pvalues          FDR
## LSR1   0.0019287458 1.909045e-05  1.909045e-05 0.0004799884
## NME1   0.4880421159 5.028056e-01  5.028056e-01 0.7385880175
## R0020C 0.0003011577 3.039952e-07  3.635477e-06 0.0001298931
## R0030W 0.0056944582 8.849840e-05  2.558489e-04 0.0038112658
## R0040C 0.0176398707 8.125405e-04  3.033306e-03 0.0264968217
```

The elements `mRNA_log2FC` and `RPF_log2FC` are log2 fold changes of mRNA and RPF comparson with baseLevel conditon. The `OVL_log2FC` is the overlap coefficence between mRNA and RPF log2FC distributions, which suggest coordinated RPF and mRNA changes. Smaller OVL suggest uncoordinated changes of RPF and mRNA across different coditions, supporting differential translations. `pvalue_log2FC` tells us the statistical significance of differential translations. The `rich_log2R`, `starvation_log2R` represent log2 ratio of RPF over mRNA in each condition. The `final_pvalues` is the final pvalue for differential translations. The `FDR` is the estimated false discovery rate corresponding to the gene for the `final_pvalue`.

Finally, the plain-text file of the results can be exported using the base **R** functions *write.csv* or *write.table*.

```
write.table(test.results,"test_results.txt",quote=F,sep="\t")
```

Session Info

```
sessionInfo()
```

```
## R version 3.2.0 (2015-04-16)
## Platform: x86_64-redhat-linux-gnu (64-bit)
## Running under: Fedora 20 (Heisenbug)
## 
## locale:
##  [1] LC_CTYPE=zh_CN.UTF-8       LC_NUMERIC=C              LC_TIME=zh_CN.utf8        
##  [4] LC_COLLATE=zh_CN.UTF-8     LC_MONETARY=zh_CN.utf8    LC_MESSAGES=zh_CN.UTF-8   
##  [7] LC_PAPER=zh_CN.utf8        LC_NAME=C                 LC_ADDRESS=C              
## [10] LC_TELEPHONE=C             LC_MEASUREMENT=zh_CN.utf8 LC_IDENTIFICATION=C       
## 
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets  methods  
## [9] base     
## 
## other attached packages:
##  [1] xtail_0.0.1                DESeq2_1.6.3              RcppArmadillo_0.5.200.1.0
##  [4] Rcpp_0.11.6                GenomicRanges_1.18.4      GenomeInfoDb_1.2.5       
##  [7] IRanges_2.0.1              S4Vectors_0.4.0           BiocGenerics_0.12.1      
## [10] knitr_1.10.5              
## 
## loaded via a namespace (and not attached):
##  [1] genefilter_1.48.1    locfit_1.5-9.1       reshape2_1.4.1       splines_3.2.0       
##  [5] lattice_0.20-31      colorspace_1.2-6     base64enc_0.1-2      survival_2.38-2     
##  [9] XML_3.98-1.2         foreign_0.8-64       DBI_0.3.1            BiocParallel_1.0.3  
## [13] RColorBrewer_1.1-2   foreach_1.4.2        plyr_1.8.3           stringr_1.0.0       
## [17] munsell_0.4.2        gtable_0.1.2         codetools_0.2-11     evaluate_0.7        
## [21] latticeExtra_0.6-26  Biobase_2.26.0       geneplotter_1.44.0   AnnotationDbi_1.28.2
## [25] highr_0.5            proto_0.3-10         acepack_1.3-3.3      xtable_1.7-4        
## [29] scales_0.2.5         checkmate_1.6.0      formatR_1.2          Hmisc_3.16-0        
## [33] annotate_1.44.0      XVector_0.6.0        sendmailR_1.2-1      gridExtra_0.9.1     
## [37] brew_1.0-6           BatchJobs_1.6        BiocStyle_1.4.1      ggplot2_1.0.1       
## [41] fail_1.2             digest_0.6.8         stringi_0.5-5        BBmisc_1.9          
## [45] grid_3.2.0           tools_3.2.0          magrittr_1.5         RSQLite_1.0.0       
## [49] Formula_1.2-1        cluster_2.0.2        MASS_7.3-41          iterators_1.0.7     
## [53] rpart_4.1-10         nnet_7.3-10
```

References

[1] Love MI, Huber W, Anders S: *Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2*. 2014.

[2] Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS: *Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling*. Science 2009, 324:218,223.