The *Rnits* package for normalization and inference of differential expression in time series microarray data.

*Dipen P. Sangurdekar* ¹

May 1, 2024

Contents

1 Introduction .................................................................................. 1
2 Method ....................................................................................... 2
  2.1 Data preprocessing and normalization. ..................................... 2
  2.2 B-spline basis model based hypothesis testing. .......................... 2
  2.3 Clustering ............................................................................... 3
  2.4 Candidate model evaluation. .................................................... 3
  2.5 Model Selection ...................................................................... 4
3 Expression profiling in Yeast in a chemostat perturbation experiment ................................................................................. 4
  3.1 Data preparation ........................................................................ 5
  3.2 Running Rnits ....................................................................... 6
    3.2.1 Build Rnits object from *ExpressionSet* or data matrix ........ 6
    3.2.2 Fitting the model .............................................................. 9
    3.2.3 Get top genes and other fit summary statistics .................. 9
4 Summary ..................................................................................... 13

1 Introduction

Gene expression studies interrogating various specific aspects of cellular physiology – cell cycle regulation, dynamic response to perturbations or natural circadian rhythms – involve samples being collected over a time series. A number of methods have been developed to infer differentially expressed genes from temporal transcriptional data. One class of methods involves extending static differential expression methods like ANOVA ([1]) and empirical Bayes ([2, 3]) to temporal data sets. Another approach is to use functional data analysis methods to model the time series data as linear combinations of basis functions (splines) ([4, 5, 6, 7, 8, 9]). Differential expression is then inferred by hypothesis testing on the basis coefficients, empirical Bayes analysis ([10, 11]) or by hypothesis testing between two groups ([5, 9, 12, 8, 13]).
The *Rnits* package for normalization and inference of differential expression in time series microarray data.

The method proposed by [8] and [9] utilizes a model driven approach to infer differential expression. The null hypothesis is that for a given gene, the difference in profiles between two treatments (or their deviation from the population average for the gene) is random. They model this scenario by fitting a single average curve, constructed from natural cubic spline basis curves, onto the two profiles. Under the alternative hypothesis, the two profiles arise from distinct underlying biological processes and this is modeled by using different models for each profile. If the gene is significantly DE, the quality of fit (as measured by the sum of squares of the regression residuals) under the alternative model is better than under the null model. To get the significance of the fit statistic, the method instead uses a bootstrapping based approach to simulate expression profiles under the null model and empirically estimates the frequency of the simulated fit statistic being higher than the observed one. A critical parameter in this modeling approach is the dimensionality of the spline basis $p$. [8] propose a method to pick an optimal model that minimizes the generalized cross-validation error on the top eigenvectors inferred from the data. In the current work, we extend the previous work by proposing an alternative approach to model selection for inference in complex temporal data. We identify distinct optimal models for clustered subsets of genes sharing complex patterns based on the statistical power to detect genes with differential trajectories. We implement this approach in a statistical package written in R, *Rnits*, to assist the research community in the end-to-end genomic analysis of time course experiments, from raw data import normalization, inference of DE and visualization of results.

## 2 Method

### 2.1 Data preprocessing and normalization

Raw or pre-processed expression data can be imported into *Rnits* either as a data matrix, an RGLList object created by the package *limma*, an AffyBatch object (package *affy*) or an eSet object downloaded from NCBI's Gene Expression Omnibus by package *GEOquery*. For all data formats, *Rnits* offers probe filtering and normalization options that are native to the respective packages. Processed data, along with probe and sample phenotype data, is stored as an *Rnits* class object, which inherits all methods from the *ExpressionSet* object in R that is commonly used for storing expression data and metadata. This allows expression and metadata retrieval from the object using the standard methods defined for *ExpressionSet* objects. In addition to the standard metadata requirements for the *ExpressionSet* object, building *Rnits* objects requires columns labeled "Sample" and "Time" in the phenotype data matrix, and a column labeled "GeneName" for the probe data matrix (for gene level summarization). One of the time series experiments may be labeled as ‘control’ for the purpose of clustering data. Inference can be done at the probe level or at the gene level if multiple probes represent a single gene. For gene level analysis, probe data is collapsed into a gene level summary using the robust Tukey Biweight estimator, which penalizes outlier values. For analysis, the distinct time series experiments must have been sampled at the same points.

### 2.2 B-spline basis model based hypothesis testing

The analysis approach is to model the time course data under the null hypothesis that individual gene expression trajectories from distinct sets is a realization of the same underlying basis trajectory modeled as a B-spline curve. Under the alternative hypothesis, each series is modeled as a distinct underlying basis. [8] et al describe a method of selecting the optimal B-spline basis model $L$. For each data set, the top eigenvectors are computed and the set of models $\hat{L}$ are used to compute the generalized cross-validation error using the eigenvectors
The \textit{Rnits} package for normalization and inference of differential expression in time series microarray data.

as predicted variables. For each data set, the model that minimizes the cross-validated error across all top components is chosen as optimal and the largest optimal model among all data sets is chosen (\textit{Rnits.gcv}). We extend this work to develop a parallel approach for selecting the optimal model based on the power of inferring differentially expressed genes. \textit{Rnits.power}. A series of candidate models $\mathcal{L}$ are evaluated on the entire data set and optimal models are selected based on the distribution of the $p$-values of fit.

2.3 Clustering

In the first step of this approach, the genes (or probes) may be clustered. Clustering allows models of varying complexity to be selected to represent the diverse expression patterns observed within the data. Each gene/probe is assigned to one of $K$ distinct clusters using k-means clustering on gene centered data (either all the series or based on a single "control" series). The total number of clusters may be provided by the user as an argument or is determined empirically as twice the number of eigenvectors required to explain 70\% of variation in the time series labeled as control or the entire data set. If any of the resulting clusters have less than 500 genes, clustering is iteratively repeated with one less initial cluster centroid.

2.4 Candidate model evaluation

The size $r$ of the individual time series experiment determines the set of candidate B-spline models that can be investigated. Each model $\mathcal{L}(c, p, r)$ is defined by its degree of curvature $c$ and the degrees of freedom $p$ i.e. the number of basis splines used. The rank constrains the curvature and the number of basis splines as follows. For a given matrix with rank $r$, the set of candidate models $\mathcal{L}$ were:

$$\mathcal{L}(c, p, r) \begin{cases} 3 \leq p \leq \min(8, r - 2) \\ 2 \leq c \leq \min(5, p - 1) \end{cases}$$

where each model was constrained to have a maximum degree of curvature of 5 and maximum number of basis splines at 8. Whenever sufficient degrees of freedom were available to allow manual placement of knots ($p - c - 1 > 0$), knots were chosen based on which time points had the maximum inflection. For each cluster $k$ (or for the entire data), each model $\mathcal{L} \in \mathcal{L}$ or the optimal model obtained by generalized cross-validation was evaluated as follows:

- The gene expression data matrix for series $s$ can be represented as $Y_{mn}$ where $m$ represents the number of probes (or genes) in cluster $k$, and $n$ equals size of the data series. The candidate model $\mathcal{L}$ is represented by its B-spline basis $X^\mathcal{L}_{pn}$. For models with larger basis splines, knot placement is guided by regions of higher curvature.

- Under the null model of no differential expression between the different time series, the distinct profiles are assumed to be generated from the same underlying curve, and the expression matrix for all series $Y_{mn}$ is fit as follows:

$$Y_{mn} = \hat{\beta}_{mn} X^\mathcal{L}_{pn} + E_{mn} = \hat{Y}_{null} + E_{null}$$

- Under the alternate hypothesis, for each individual series $s \in (1, S)$, we fit a distinct curve and the individual data matrix is fit as follows

$$Y_{mr} = \hat{\beta}_{mr} X^\mathcal{L}_{pr} + E_{mr} = \hat{Y}_{alt} + E_{alt}$$
The Rnits package for normalization and inference of differential expression in time series microarray data.

- The two models are evaluated using a ratio statistic $F^c_k$

$$F^c_k = \frac{\sum_{s} E^2_{null} - \sum_{s} \sum_{r} E^2_{alt}}{\sum_{s} E^2_{alt} + s_0}$$

where $\sum_{r} E^2_{null}$ is the vector of sum of square of residuals under the null hypothesis for all genes in cluster $k$ and $\sum_{s} \sum_{r} E^2_{alt}$ is the sum of squared residuals for each gene in cluster $k$ added over all series $s$. We add a variance stabilizing factor to the denominator $s_0$ computed as described by [14].

To compute p-values for the ratio statistic, we use the bootstrap approach as described by [8]. First the residuals under the alternate model $E_{alt}$ are standardized to correct for heteroskedasticity for each series. This penalizes the models with larger basis splines by increasing the effective size of their residuals.

$$E'_{alt} = \frac{E_{alt}}{\sqrt{1 - X(X^T X)^{-1}X^T}}$$

For each bootstrap iteration $b$, a synthetic null data set is constructed by adding the bootstrapped studentized residuals to the null fit.

$$Y^*_{null} = \hat{Y}_{null} + E'_{alt}^{*}$$

Eqn (2)-(4) are repeated for each iteration and ratio statistics $F^c_{k*}$ under the null hypothesis of no differential expression are computed. Empirical p-values are computed for the observed ratio statistic based on the distribution of the simulated ones under the null hypothesis.

$$P^c_k = \frac{\#(F^c_{k*} > F^c_k)}{B}$$

where $B$ are the number of bootstrap iterations.

2.5 Model Selection

For each cluster $k$, p-values are similarly calculated for each model $\mathcal{L} \in \mathcal{L}$. The significance of each model fit is a function of the size of the studentized residuals under the null and alternative hypotheses. While more complex models (larger number of basis functions) may yield smaller residuals during the alternative fit, these may get heavily penalized during correction for heteroskedasticity, leading to low statistical power. At the other extreme, an insufficiently complex model may not effectively capture the complexity of the expression trajectories that may lead to larger residuals and lack of power. To alleviate this effect, we select the model $\mathcal{L}$ which results in the maximum power, as determined by the number of genes inferred at an estimated False Discovery Rate $< 5\%$.

3 Expression profiling in Yeast in a chemostat perturbation experiment

We applied the Bspline methods to time course expression data tracking the glucose perturbation responses of a wild type yeast strain grown at steady state in a chemostat with galactose as the carbon source ([15]). Two glucose concentrations (effective concentration of
The *Rnits* package for normalization and inference of differential expression in time series microarray data.

0.2 g/l and 2 g/l) were used as pulses and responses were tracked at 10, 15, 20, 30, 45, 90, 120 and 150 minutes post each treatment. We obtained the published microarray data from NCBI Gene Expression Omnibus (GEO accession GSE4158) and analyzed it using *Rnits*.

### 3.1 Data preparation

First, we download the data from GEO and format it.

```r
# Download NCBI GEO data with GEOquery
library(knitr)
rm(list = ls())
library(GEOquery)
## Loading required package: Biobase
## Loading required package: BiocGenerics
##
## Attaching package: 'BiocGenerics'
## The following objects are masked from 'package:stats':
##     IQR, mad, sd, var, xtabs
## The following objects are masked from 'package:base':
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, table,
##     tapply, union, unique, unsplit, which.max, which.min
## Welcome to Bioconductor
##
## Vignettes contain introductory material; view with
## 'browseVignettes()'. To cite Bioconductor, see
## 'citation("Biobase")', and for packages 'citation("pkgname")'.
## Setting options('download.file.method.GEOquery'='auto')
## Setting options('GEOquery.inmemory.gpl'=FALSE)
library(stringr)
library(Rnits)
## Loading required package: ggplot2
## Loading required package: limma
##
## Attaching package: 'limma'
## The following object is masked from 'package:BiocGenerics':
##     plotMA

gds <- getGEO("GSE4158", AnnotGPL = FALSE)[[1]]
```
The *Rnits* package for normalization and inference of differential expression in time series microarray data.

```r
## Found 1 file(s)
## GSE4158_series_matrix.txt.gz
class(gds)
## [1] "ExpressionSet"
## attr("package")
## [1] "Biobase"
gds
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 10752 features, 26 samples
## element names: exprs
## protocolData: none
## phenoData
## sampleNames: GSM94988 GSM94989 ... GSM95013 (26 total)
## varLabels: title geo_accession ... data_row_count (36 total)
## varMetadata: labelDescription
## featureData
## featureNames: 1 2 ... 10752 (10752 total)
## fvarLabels: ID METACOLUMN ... TYPE (10 total)
## fvarMetadata: Column Description labelDescription
## experimentData: use 'experimentData(object)'
## pubMedIds: 16381818
## Annotation: GPL3415

# Extract non-replicate samples
pdata <- pData(gds)
filt <- pdata$characteristics_ch2 %in% names(which(table(pdata$characteristics_ch2) == 2))
gds <- gds[, filt]
pdata <- pData(gds)
time <- as.numeric(str_extract(pdata$characteristics_ch2, '\\d+'))
sample <- rep("2g/l", length(time))
sample[grep("0.2g/l", gds["title"])] <- "0.2g/l"

# Format phenotype data with time and sample information
gds["Time"] <- time
gds["Sample"] <- sample
dat <- gds
tData(dat)["Gene Symbol"] <- tData(dat)$ORF
```

### 3.2 Running Rnits

#### 3.2.1 Build Rnits object from *ExpressionSet* or data matrix

Next we build the *Rnits* object

```r
# Build rnis object from formatted data (samples can be in any order) and
# between-array normalization.
rmitsobj = build.Rnits(dat[, sample(ncol(dat))], logscale = TRUE, normmethod = "Between")
```
The *Rnits* package for normalization and inference of differential expression in time series microarray data.

```r
# Between Array normalization
# Data set with 10752 features and 2 samples in 1 replicates
rnitsobj
# Rnits (storageMode: lockedEnvironment)
# assayData: 10752 features, 16 samples
# element names: exprs
# protocolData: none
# phenoData
# sampleNames: GSM95010 GSM95011 ... GSM94998 (16 total)
# varLabels: title geo.accession ... Sample (38 total)
# varMetadata: labelDescription
# featureData
# featureNames: 1 2 ... 10752 (10752 total)
# fvarLabels: ID METACOLUMN ... GeneName (12 total)
# fvarMetadata: labelDescription
# experimentData: use 'experimentData(object)'
# Annotation:
# Alternatively, we can also build the object from just a data matrix, by
# supplying the 'probedata' and 'phenodata' tables

datdf <- exprs(dat)
rownames(datdf) <- fData(dat)$ID
probedata <- data.frame(ProbeID = fData(dat)$ID, GeneName = fData(dat)$ORF)
phenodata <- pData(dat)
rnitsobj0 = build.Rnits(datdf, probedata = probedata, phenodata = phenodata, logscale = TRUE,
   normmethod = "Between")
# Log scale is TRUE
# Between Array normalization
# Data set with 2 time series
# Dataset with 10752 features and 2 samples in 1 replicates
# Extract normalized expression values
lr <- getLR(rnitsobj)
head(lr)
```

```r
# GSM95010 GSM95011 GSM95012 GSM95013 GSM95002 GSM95004
# 1  0.99269970 -0.84741081 -0.22867060  0.5941574 -1.18590131  0.5137818
# 2  0.12448779  0.76758977  1.43044492  0.9482546 -0.04944478  0.4212989
# 3  0.05362765  0.69975822 -0.78887206  0.2560923  0.62460336  0.3294558
# 4  0.59017581 -0.47129159 -0.785572  -1.03973077  0.8060511
# 5  0.05032467  0.04912668  0.4690431  -1.59731454  0.4243162
# 6  0.84482623  1.0519264  1.25240820  1.6704265  0.22307705  0.307783
```

```r
# GSM95006 GSM95008 GSM94998 GSM94989 GSM949991 GSM949992
# 1  0.31852729 -0.19303436 -0.20592568 -0.0797288 -0.05466009  0.00567193
# 2  0.31219698  0.13219700  0.62636186  0.6992206  0.78594481  0.684897162
# 3  0.05362765 -0.69975822 -0.78887206  0.2560923  0.62460336  0.3294558
# 4  0.59017581 -0.47129159 -0.785572  -1.03973077  0.8060511
# 5  0.05032467  0.04912668  0.4690431  -1.59731454  0.4243162
# 6  0.84482623  1.0519264  1.25240820  1.6704265  0.22307705  0.307783
```
The \textit{Rnits} package for normalization and inference of differential expression in time series microarray data.

```
## 1  0.3418725 -2.487601  0.1680312  0.2977183
## 2  0.8925079 NA  1.1047615  0.7014654
## 3 -1.0909498 NA  0.1363547  0.7593684
## 4 -0.1664416 NA -0.2371354 NA
## 5  0.1791928 NA  0.3288013  0.5037860
## 6  1.0160011 -1.292664  0.8986091  0.7248260

# Impute missing values using K-nn imputation
lr <- getLR(rnitsobj, impute = TRUE)

```

```
## Warning in knnimp(x, k, maxmiss = rowmax, maxp = maxp): 1635 rows with more
## than 50 % entries missing;
## mean imputation used for these rows

## Cluster size 9117 broken into 2854 6263
## Cluster size 2854 broken into 2608 846
## Cluster size 2608 broken into 1305 703
## Done cluster 1305
## Done cluster 703
## Done cluster 2008
## Done cluster 846
## Done cluster 2854
## Cluster size 6263 broken into 3939 2324
## Cluster size 3939 broken into 2055 1884
## Cluster size 2055 broken into 1059 996
## Done cluster 1859
## Done cluster 996
## Done cluster 2055
## Cluster size 1884 broken into 798 1086
## Done cluster 798
## Done cluster 1086
## Done cluster 1884
## Done cluster 3939
## Cluster size 2324 broken into 855 1469
## Done cluster 855
## Done cluster 1469
## Done cluster 2324
## Done cluster 6263

head(lr)

```

```
##   GSM95010 GSM95011 GSM95012 GSM95013 GSM95002 GSM95004
## 1  -0.99269970 -0.84741081 -0.22867060  0.5941574 -1.18590131  0.5137818
## 2   0.12448779  0.76758977  1.43044492  0.9482546 -0.04944478  0.4212989
## 3   0.05362765 -0.69975822 -0.70887206  0.2560923  0.62460336  0.3294558
## 4  -0.59017581 -0.47129159 -0.08102022  0.78850923  0.62460336  0.3294558
## 5  -1.05032467  0.04912668  0.09847670 -0.4690431 -1.59731454  0.4243162
## 6   0.84482623  1.10519264  1.25240820  1.6704265 -0.22307705  0.3077783

```

```
##       GSM95006 GSM95008 GSM94988 GSM94989 GSM94991 GSM94992
## 1 -0.31826977  0.49126680 -0.09847670 -0.4690431 -1.59731454  0.4243162
## 2  0.31219698  0.13219700  0.62536186  0.6922206  0.78594811  0.68497162
## 3  0.09733628 -0.39948994 -0.26945472 -0.7516256 -1.01418075 -1.06631505
## 4  0.57356735  0.43055165 -0.37012699 -0.3610244 -0.97771053 -0.84130862
```
```
The Rnits package for normalization and inference of differential expression in time series microarray data.

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 3.2.2 Fitting the model

Next fit the model using gene-level summarization and by clustering all genes.

```r
# Fit model using gene-level summarization
rnitsobj <- fit(rnitsobj0, gene.level = TRUE, clusterallsamples = TRUE)
```

### Warning: did not converge in 10 iterations

The model may be fit using no clustering of samples, or by precomputing the optimal model using the method of [8].

```r
rnitsobj_nocl <- fit(rnitsobj0, gene.level = TRUE, clusterallsamples = FALSE)
```

```r
opt_model <- calculateGCV(rnitsobj0)
```

```r
rnitsobj_optmodel <- fit(rnitsobj0, gene.level = TRUE, model = opt_model)
```

### 3.2.3 Get top genes and other fit summary statistics
The *Rnits* package for normalization and inference of differential expression in time series microarray data.

```r
# Get p-values from fitted model
pval <- getPval(rnitsobj)
head(pval)
## YAL001C  YAL003W  YAL005C  YAL008W  YAL010C  YAL012W
## 0.343467949 0.015498982 0.004038462 0.517227564 0.555160256 0.046776986

# Get ratio statistics from fitted model
stat <- getStat(rnitsobj)
head(stat)
## YAL001C  YAL003W  YAL005C  YAL008W  YAL010C  YAL012W
## 0.019622677 1.197300806 0.598136354 -0.007683565 -0.012834775 0.638745445

# If clustering was used, check the cluster gene distribution
table(getCID(rnitsobj))
##
## 1  2  3  4
## 1964 3120  415  843

# P-values, ratio statistics and cluster ID's can be retrieved for all genes
# together
fitdata <- getFitModel(rnitsobj)
head(fitdata)
## Ratio.statistic  p.value  clusterID
## YAL001C     0.019622677 0.343467949 2
## YAL003W     1.197300806 0.015498982 1
## YAL005C     0.598136354 0.004038462 2
## YAL008W    -0.007683565 0.517227564 2
## YAL010C    -0.012834775 0.555160256 2
## YAL012W     0.638745445 0.046776986 1

# View summary of top genes
summary(rnitsobj, top = 10)
## Rank  Name    Statistic  p.value  FDR.percent  ClusterID  Model.degree
## 1  YFR053C 15.835 0.000 0.000 3 4
## 2  YDL163W  6.288 0.000 0.000 4 2
## 3  YDR423C  2.743 0.353 0.353 2 2
## 4  YDR419W  2.066 0.792 0.792 2 2
## 5  YPL250C  1.988 0.792 0.792 2 2
## 6  YLR108C 11.325 0.792 0.792 1 5
## 7  YNR019W  1.845 0.792 0.792 2 2
## 8  YMR081C 14.090 0.792 0.792 3 4
## 9  YCLX02C  1.785 0.792 0.792 2 2
## 10 YDL181W  1.705 0.792 0.792 2 2

Model.df
## YFR053C  6
## YDL163W  5
## YDR423C  3
## YDR419W  3
## YPL250C  3
## YLR108C  6
```
The *Rnits* package for normalization and inference of differential expression in time series microarray data.

```r
## Extract data of top genes (5% FDR)
td <- topData(rnitsobj, fdr = 5)
## 1039
head(td)
```

```
# YNR019W 3
# YMR081C 6
# YCLX02C 3
# YDL181W 3

# Extract data of top genes (5% FDR)
td <- topData(rnitsobj, fdr = 5)
## 1039
head(td)
```

Next we plot the results for the top 16 differentially expressed genes by FDR

```r
# Plot top genes trajectories
plotResults(rnitsobj, top = 16)
```

```
## 'geom_smooth()' using formula = 'y ~ x'
## Next: 1 of 1
```
The *Rnits* package for normalization and inference of differential expression in time series microarray data.

### Session Info

```r
sessionInfo()
## R version 4.4.0 beta (2024-04-15 r86425)
## Platform: x86_64-pc-linux-gnu
## Running under: Ubuntu 22.04.4 LTS
##
## Matrix products: default
## BLAS: /home/biocbuild/bbs-3.19-bioc/R/lib/libRblas.so
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0
##
## locale:
## [1] LC_CTYPE=en_US.UTF-8   LC_NUMERIC=C
## [3] LC_TIME=en_US.UTF-8    LC_COLLATE=en_US.UTF-8
## [5] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
## [7] LC_PAPER=en_US.UTF-8   LC_NAME=C
## [9] LC_ADDRESS=C           LC_TELEPHONE=C
##
## time zone: America/New_York
tzcode source: system (glibc)
##
## attached base packages:
## [1] stats  graphics grDevices utils  datasets methods base
##
## other attached packages:
## [1] Rnits_1.38.0  limma_3.60.0  ggplot2_3.5.1
## [4] stringr_1.5.1  GEOquery_2.72.0  Biobase_2.64.0
## [7] BiocGenerics_0.50.0 knitr_1.46
##
## loaded via a namespace (and not attached):
## [1] gtable_0.3.5  impute_1.78.0   xfun_0.43
## [4] lattice_0.22-6  tzdb_0.4.0   vctrs_0.6.5
## [7] tools_4.4.0   generics_0.1.3  curl_5.2.1
```
The `Rnits` package for normalization and inference of differential expression in time series microarray data.

4 Summary

Dynamic changes in transcriptional programs to changing stimulus or cell cycle stages necessitate an experimental design that tracks expression changes over a period of time. There are often constrains on the replication levels in such designs, posing a challenging inference problem. A number of methods have been developed to infer differential expression between conditions. We extend a previously developed method by [8] of using B-splines to perform hypothesis testing. We extended the method to include a model selection step that optimizes for distinct subsets within the expression data, corresponding to diverse transcriptional regulatory patterns in cells. We present this method as an comprehensive analysis R package `Rnits` to enable the data import, pre-processing, normalization, analysis and visualization of gene expression data from a variety of sources (GEO, raw or normalized data tables).

References


The *Rnits* package for normalization and inference of differential expression in time series microarray data.


