Model-based analysis of DNA methylation data

Tao Wang, Mengjie Chen, Hongyu Zhao

Oct 2014

1 Introduction

This guide provides a tour of the Bioconductor package `MBAmethy`, a R package implements a region-based DNA methylation analysis method. This method utilize both the information from biological replicates and neighboring probes by explicitly modeling the probe-specific effect and encouraging the neighboring similarity by a group fused lasso penalty. When multiple biological replicates availale, this method can be applied as an alternative to single probe analysis. Our method is applicable to densely methylated data from platforms such as Illumina 450k arrays, CHARM arrays, and bisulfite sequencing.

2 Overview of capabilities

`MBAmethyl()` is the main function in the package. It takes a window of multiple probes (we use 200 probes on real data) from multiple samples as input and returns smoothed methylation values.

Let us simulate a simple dataset with 80 probes from 40 samples.

```
>   p <- 80
>   n <- 40
>   K <- 2
>   k <- K - 1
>   cp <- numeric()
>   L <- c(0, floor(p / K) * (1 : k), p)
>   cp <- floor(p / K) * (1 : k) + 1
>   ## phi0: probe effects; theta0: true methylation values; part: partition of prob
>   phi0 <- runif(p, 0.5, 2.0)
>   theta0 <- matrix(0, p, n)
>   part <- list()
>   for (s in 1 : K) {
+     part[[s]] <- (L[s] + 1) : L[s + 1]
+     phi0[part[[s]]] <- phi0[part[[s]]] / sqrt(mean(phi0[part[[s]]]^2))
```

```
+   }
>   theta0[part[[1]], ] <- rep(1, length(part[[1]])) %x% t(runif(n, 0.1, 0.6))
>   theta0[part[[2]], ] <- rep(1, length(part[[2]])) %x% t(runif(n, 0.4, 0.9))
>   error <- matrix(runif(p * n, 0, 0.1), p, n)
>   Y <- theta0 * phi0  + error
```

The input matrix Y is a $p \times n$ matrixof methylation values (beta values), where p is the number of probes and n is the number of samples. To get smoothed the methylation values, just apply function `MBAmethyl()`:

```
>  library(MBAmethyl)
>  fit <- MBAmethyl(Y, steps = 30)
```

The function will return two lists of results using AIC and BIC as model selection criteria, respectively. To check BIC result,

```
>   str(fit$ans.bic)
```

```
List of 5
 $ theta   : num [1:80, 1:40] 0.478 0.478 0.478 0.478 0.478 ...
 $ phi     : num [1:80] 0.868 0.583 1.275 0.585 1.197 ...
 $ change.p: num 41
 $ rss     : num 0.000419
 $ df.bic  : num 160
```

```
>   theta <- fit$ans.bic
```

`theta` stores the smoothed values.

SessionInfo

- R version 3.3.1 (2016-06-21), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=C`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: MBAmethyl 1.8.0
- Loaded via a namespace (and not attached): tools 3.3.1