polyaPeak: a tool for reranking ChIP-seq peaks with peak shape information

Hao Wu
Department of Biostatistics and Bioinformatics
Emory University
Atlanta, GA 303022
`hao.wu@emory.edu`

May 10, 2012

Abstract

`polyaPeak` is an R package for reranking transcription factor binding sites (TFBS), or peaks, detected from ChIP-seq data. ChIP-seq data for mapping TFBSs have a characteristic pattern: around each binding site, sequence reads aligned to the forward and reverse strands of the reference genome form two separate peaks shifted away from each other, and the true binding site is located in between. Although this pattern has been used previous in various way, few method fully incorporate the information in a rigorous statistical model to improve peak detection. `polyaPeak` describes peak shapes using a flexible Pólya mixture model. The shapes are automatically learnt from the data using a hybrid Expectation-Maximization (EM) and Minorization-Majorization (MM) algorithm. The peak shape information is then integrated with the read count information via a hierarchical mixture model to distinguish true binding sites from background noises. Analyses of real data show that PolyaPeak is capable of robustly improving the state-of-the-art peak calling algorithms MACS and CisGenome, and it also outperforms PICS which is another peak calling method that uses the peak shape information.

1 Quick start

`polyaPeak` depends on two Bioconductor packages `Rsamtools` and `GenomicRanges` for handling BAM files and obtain window counts. Follow the Bioconductor installation guide to install these packages.

To get start, `polyaPeak` requires following inputs:

- A ranked peak list from another peak caller (such as MACS or CisGenome). The list must be saved as a data frame with 4 columns: chr, start, end, summit for the chromosome number, start/end positions and summits for all peaks.

- The aligned sequence read file in BAM format.

To rerank the peak, simple call `rerank.peak` function. Follow the function help page (e.g., type `?rerank.peak` in R command console) for detailed instruction of function syntax.

2 Session Info

```
>   sessionInfo()
```

```
R version 2.14.1 (2011-12-22)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.iso885915       LC_NUMERIC=C
 [3] LC_TIME=en_US.iso885915        LC_COLLATE=C
 [5] LC_MONETARY=en_US.iso885915    LC_MESSAGES=en_US.iso885915
 [7] LC_PAPER=C                     LC_NAME=C
 [9] LC_ADDRESS=C                   LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.iso885915 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
[1] tools_2.14.1
```