MultiChIPmixHMM: ChIP-chip data analysis modeling spatial dependencies and multiple replicates

Caroline Bérard

Modified: June 13, 2013. Compiled: June 14, 2013

Abstract

This vignette illustrates the use of the `MultiChIPmixHMM` package to analyse data from ChIP-chip experiments. Based on a mixture model of regressions, we propose two methods to take into account or not dependencies between adjacent probes along the genome. It is also possible to analyse multiple replicates simultaneously.

Contents

1 Introduction

Chromatin immunoprecipitation coupled with hybridization to a tiling array (ChIP-chip) is a cost-effective and routinely used method for identifying target genes of transcription factors, for analyzing histone modifications or for studying the methylome on a genome-wide scale. In a ChIP-chip experiment, a chromatin immunoprecipitation sample (IP) is compared against a reference sample of genomic DNA (Input).

The `MultiChIPmixHMM` package implements a linear regression mixture model, designed to perform a joint modeling of IP and Input measurements. This two-component mixture model discriminates the population of enriched probes from non-enriched ones. Moreover, this package enabling modeling of spatial dependencies and a simultaneous analysis of replicates to further improve the identification of enriched probes.

2 Model

The method is based on a mixture model of regressions [2]. To be specific, let (x_t, Y_t) be the log-Input and log-IP intensities of probe t, respectively. The (unknown) status of the probe is characterized through a label Z_t which is 1 if the probe is enriched and 0 if it is normal (not enriched). We assume the Input-IP relationship to be linear whatever the population, but with different parameters. More precisely, we have:

$$\begin{aligned} Y_t &\sim \mathcal{N}(a_0 + b_0 x_t\ ,\ \sigma^2) \qquad \text{if } Z_t = 0 \text{ (normal probe)} \\ Y_t &\sim \mathcal{N}(a_1 + b_1 x_t\ ,\ \sigma^2) \qquad \text{if } Z_t = 1 \text{ (enriched probe)} \end{aligned}$$

The `MultiChIPmixHMM` R package presents two different functions based on this model:

- The first function (`multiChIPmix`) allows us to analyse one sample, with a generalization to simultaneously analyse multiple biological replicates;
- The second function (`multiChIPmixHMM`) takes into account the spatial dependence between neighboring probes with a hidden Markov model, and also makes it possible to analyse multiple biological replicates.

To account for the dependence between adjacent probes, we assume that the latent variable $\{Z_t\}$ is a first order Markov chain.
To simultaneously analyse multiple biological replicates, we note (x_{tr}, Y_{tr}) the log-Input and log-IP intensities of probe t for the replicate r. For each biological replicate r, the model is the following:

$$\begin{aligned} Y_{tr} &\sim \mathcal{N}(a_{0r} + b_{0r} x_{tr}\ ,\ \sigma_r^2) \qquad \text{if the probe is normal} \\ Y_{tr} &\sim \mathcal{N}(a_{1r} + b_{1r} x_{tr}\ ,\ \sigma_r^2) \qquad \text{if the probe is enriched} \end{aligned}$$

The replicates are assumed to be independent. This model makes it possible to determine a single status by probe for all replicates.

The marginal distribution of $\log(Y)$ for a given $\log(x)$ is:

$$(1-p)\prod_r f_{0r}\{\log(Y_r)|\log(x_r)\} + p\prod_r f_{1r}\{\log(Y_r)|\log(x_r)\}\ ,$$

where f_{jr} stands for the probability density function of a Gaussian distribution with mean $a_{jr} + b_{jr}\log(x_r)$ and variance σ_r^2.

All parameters of the HMM are estimated using the Baum-Welch algorithm [3] representing a special case of the EM algorithm [1]. To obtain relevant initial values of the mixture parameters, we applied a Principal Component Analysis to each biological replicate and used the first axis to derive the intercept and slope of the regression.

Identification of enriched probes is based on conditional probabilities. A probe is declared enriched if its enriched conditional probability (state-posterior probability of the enriched state) is higher than $1 - \alpha$, where α is chosen by the user.

The mixture model with two linear regressions is adapted if there are two populations. But if there is no enriched probes in the dataset, there is only one population to find. In order to choose between one or two populations, a linear regression mixture model is applied with one and two components, and the best model is selected with BIC criterion [4].

3 ChIP-chip dataset analysis

To begin, we load the package:

```
> library(MultiChIPmixHMM)
```

The input data can be either a file or a RData object. Each file/object must have exactly 3 columns named ID, INPUT and IP (ID=probe identifier, INPUT=log2(INPUT), IP=log2(IP)). The field separator character is tabs if input data files are used.
For example, we load the datasets provided in the package, which is a small toy example:

```
> data(data_Rep1)
> data(data_Rep2)
> head(data_Rep1)
```

```
  ID     INPUT        IP
1  1  9.949472  9.257767
2  2 10.072620 11.763969
3  3 10.471866 10.074208
4  4 11.978731 12.051185
5  5  9.840943 10.110992
6  6 10.220290  8.266811
```

We have now to choose if we want to take into account the dependence between neighboring probes.
If yes, the data have to be ordered according to the position of the probes, and we use the `multiChIPmixHMM` function:

```
> res = multiChIPmixHMM(files=c(data_Rep1,data_Rep2),init.by.PCA=TRUE,
+ alpha=0.01,proba=0.5,eps=1e-06,fileOUT="multiChIPmixHMM-results.txt",
+ fileOUTgraph="multiChIPmixHMM-results.pdf")
```

If not, we use the `multiChIPmix` function:

```
> res = multiChIPmix(files=c(data_Rep1,data_Rep2),init.by.PCA=TRUE,
+ alpha=0.01,proba=0.5,eps=1e-06,fileOUT="multiChIPmix-results.txt",
+ fileOUTgraph="multiChIPmix-results.pdf")
```

The parameter `files` is a vector of the names of the replicates. The size of the vector corresponds to the number of replicates. If you want to load .txt files instead of RData objects, you have to write : `files=c("data1.txt","data2.txt")`.

After that, we obtain the values of the estimated parameters (displayed on the screen and also available in the `res$out` object), a .pdf graph showing the probe classification and the regression lines for each replicate, and a .txt file containing the output informations, that is the ID, the INPUT and IP values, the posterior probabilities and the predicted status for each probe.

If we want to directly plot the graph (for the replicate 1 for example), the command lines are:

```
> couleur = c()
> couleur[res$status == 1] = "red"
> couleur[res$status == 0] = "black"
> rep=1
> plot(data_Rep1$INPUT,data_Rep1$IP,xlab="INPUT",ylab="IP",col=couleur)
> x<-seq(min(data_Rep1$INPUT),max(data_Rep1$INPUT),length=100)
> lines(x,y=res$out$a[1,rep]+res$out$slope[1,rep]*x,col="blue",lwd=2)
> lines(x,y=res$out$a[2,rep]+res$out$slope[2,rep]*x,col="red",lwd=2)
```

In order to concatenate adjacent probes with the same status into regions, we can use the `makeRegions` function. The first argument of the function is a file with a column named ID (for the probe identifier) and a column named status for the status of each probe. The second argument is the gap parameter, which corresponds to the number of probes with status "normal" allowed between two probes with status "enriched". It returns a file named "regions.txt" with the detected enriched regions. Each region is defined by the ID of the first and last probes covering the region.

```
> makeRegions(fileIN="multiChIPmixHMM-results.txt", gap=1)
```

4 Session Info

```
> sessionInfo()
```

```
R version 2.15.1 (2012-06-22)
Platform: i486-pc-linux-gnu (32-bit)

locale:
 [1] LC_CTYPE=fr_FR.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=fr_FR.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=fr_FR.UTF-8    LC_MESSAGES=fr_FR.UTF-8
 [7] LC_PAPER=C                 LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=fr_FR.UTF-8 LC_IDENTIFICATION=C

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

other attached packages:
[1] MultiChIPmixHMM_4.0

loaded via a namespace (and not attached):
[1] tools_2.15.1
```

References

[1] A.˜P. Dempster, N.˜M. Laird, and D.˜B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society, series B*, 39(1):1–38, 1977.

[2] M.L. Martin-Magniette, T.˜Mary-Huard, C.˜Bérard, and S.˜Robin. Chipmix: mixture model of regressions for two-color chip-chip analysis. *Bioinformatics*, 24:i181–i186, 2008.

[3] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. pages 257–286, 1989.

[4] G.˜Schwarz. Estimating the dimension of a model. *Annals of Statistics 6*, pages 461–464, 1978.