An Introduction to the BETR Package

Martin Aryee[*]

March, 2009

Department of Oncology
Johns Hopkins School of Medicine
Baltimore, MD, USA

1   Introduction

The betr package package implements the BETR (Bayesian Estimation of Temporal Regulation) algorithm to rank differentially expressed genes in microarray time-course data. It calculates the probability of differential expression for each feature (gene) in a data set. The algorithm takes the correlations between time points into account, resulting in increased sensitivity compared to analyzing each time point in isolation. It can be used to make comparisons between two conditions (e.g. two treatments) or within a single condition where the first baseline measurement is compared to subsequent time points (e.g. a developmental time-course).

A key step in the algorithm involves fitting two different models to each gene's data. The first assumes no differential expresson, whereas the second allows for correlated differential expression between time points. By considering which of the two models better fit the data we can calculate the probability of differential expression for each gene. The algorithm is described in detail in [1].

2   Example

We will use the simulated 1000-gene dataset included in the betr package for this demonstration. The data represents an experiment comparing gene expression in *StrainA* and *StrainB*. There are 3 replicates from each strain and each replicate is sampled at 4 time points. The samples are hybridized to one-color arrays giving a total of 24 arrays.

```
R> library(betr)
R> library(Biobase)
```

[*]aryee@jhu

```
R> data("timeEset")
R> pData(timeEset)[1:15, ]

              strain time rep
strainA.t0.r1 strainA    0   1
strainA.t0.r2 strainA    0   2
strainA.t0.r3 strainA    0   3
strainA.t1.r1 strainA    1   1
strainA.t1.r2 strainA    1   2
strainA.t1.r3 strainA    1   3
strainA.t2.r1 strainA    2   1
strainA.t2.r2 strainA    2   2
strainA.t2.r3 strainA    2   3
strainA.t3.r1 strainA    3   1
strainA.t3.r2 strainA    3   2
strainA.t3.r3 strainA    3   3
strainB.t0.r1 strainB    0   1
strainB.t0.r2 strainB    0   2
strainB.t0.r3 strainB    0   3
```

We load the betr package and run the *betr* function to calculate the probability of differential expression for each gene.

```
R> prob <- betr(eset = timeEset, cond = pData(timeEset)$strain,
     timepoint = pData(timeEset)$time, replicate = pData(timeEset)$rep,
     alpha = 0.05)

....

R> head(prob)

1420852_a_at 1425036_a_at   1438882_at 1436951_x_at
  0.99979548   0.03449886   0.06014775   0.02508742
  1456574_at   1422484_at
  0.03496866   0.06720107
```

The parameters used include:

- *eset*: an *ExpressionSet* object containing the experimental data

- *cond*: a vector indicating the experimental condition (strain in this example) for each sample

- *timepoint*: a vector indicating the time point for each sample

- *replicate*: a vector indicating which replicate each sample belongs to

- *alpha*: the acceptable false positive rate

The function returns a vector containing the probabilities of differential expression for each feature (row) in the data set. Genes with the best evidence for differential expression will have values close to 1.

References

[1] Aryee et al. An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (bayesian estimation of temporal regulation). *submitted*.

3   Details

This document was written using:

```
R> sessionInfo()

R version 2.11.0 (2010-04-22)
x86_64-unknown-linux-gnu

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

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

other attached packages:
[1] Biobase_2.8.0 betr_1.4.0

loaded via a namespace (and not attached):
[1] limma_3.4.0   mvtnorm_0.9-9
```