Bayesian Inference of Regulatory influence on Expression (biRte)

Holger Fröhlich

April 2, 2015

1 Introduction

Expression levels of mRNA is regulated by different processes, comprising inhibition or activation by transcription factors (TF) and post-transcriptional degradation by microRNAs (miRNA). *biRte* (Bayesian Inference of Regulatory influence on Expression (biRte)) uses the regulatory networks of TFs and miRNAs together with mRNA and miRNA expression data to infer the influence of regulators on mRNA expression. Furthermore, *biRte* allows to consider additional factors such as CNVs. *biRte* has the possibility to specify Bayesian priors for the activity of each individual regulatory factor. Moreover, interaction terms between regulators can be considered. *biRte* relies on a Bayesian network model to integrate data sources into a joint likelihood model. In the model mRNA expression levels depend on the activity states of its regulating factors via a sparse Bayesian linear regression using a spikes and slab prior [**?**]. Moreover, miRNA expression levels depend on miRNA activity states. *biRte* uses Markov-Chain-Monte-Carlo (MCMC) sampling to infer activity states of regulatory factors. During MCMC, switch moves - toggling the state of a regulator between active and inactive - and swap moves - exchanging the activitiy states of either two miRNAs or two TFs - are used [8].

biRte is meant as a replacement for the earlier package *birta*. *biRte* offers several advantages compared to *birta*.

- possibility to include additional regulatory factors and data apart from TFs and miRNAs
- possibility to include target specific regulation strength values
- possibility to define a prior probabilities for activity of each individual regulator and even regulator pairs.
- significantly faster inference (about 15 fold speed-up)
- significantly higher accuracy of inference due to improved likelihood calculation
- inference of regulatory networks as a follow-up step
- possibility to work with arbitrarily complex statistical designs, if log fold changes are used.

The package can be loaded by typing:

```
> rm(list=ls())
> library(birte)
```

2 Usage of biRte

The two main functions of the package are `birteRun` and `birteLimma`. `birteLimma` is a convenience function, which passes the output of `limmaAnalysis` to `birteRun`. The most important input arguments to `birteRun` are

- **dat.mRNA**. Matrix of mRNA expression data with row names indicating genes.
- **affinities**. A weighted regulator-target graph. This is a list with at most three components (TF, miRNA, other). Each of these lists again contains a weighted adjacency list representation. See `affinities` for more information and `humanNetworkSimul` for an example. Per default weights are ignored in the inference process. IMPORTANT: gene names used in this network have to match with row names of dat.mRNA.
- **nrep.mRNA** is an integer vector, which specifies the number of replicates per condition for mRNA data

3 Applying biRte to RNAseq Data

biRte relies on the assumption that data are (multivariate) normally distributed. Application to RNAseq data is thus not immediately possible. Data should thus be transformed appropriately, e.g. via the voom + limma mechanism [5].

4 Example: Aerobic vs. anaerobic growth in E. Coli

To demonstrate the use of *biRte* we here show a most basic application to a microarray dataset by [2] together with a filtered TF-target graph [1]. The gene expression data comprises three replicates from E. Coli during aerobic growth and four replicates during anaerobic growth. The TF-target graph contains annotations for 160 transcription factors. Expression values are stored in an *ExpressionSet*.

```
> library(Biobase)
> data(EColiOxygen)
> EColiOxygen

ExpressionSet (storageMode: lockedEnvironment)
assayData: 4205 features, 7 samples
  element names: exprs
protocolData: none
phenoData
  rowNames: GSM18261 GSM18262 ... GSM18289 (7 total)
  varLabels: Strain GrowthProtocol GenotypeVariation Description
  varMetadata: labelDescription
featureData
  featureNames: 1 2 ... 4205 (4205 total)
  fvarLabels: symbol Entrez
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
  pubMedIds: 15129285
Annotation: org.EcK12.eg.db

> head(exprs(EColiOxygen))
```

```
        GSM18261 GSM18262  GSM18263  GSM18286  GSM18287  GSM18288  GSM18289
947315 10.277125 10.22119 10.410919 10.208393 10.179176 10.186009 10.009045
945490 10.138638 10.17328 10.215396 10.170649  9.993040 10.277822  9.968522
944896 11.016805 11.28574 11.308092 11.287854 11.582083 11.632015 11.463312
945321  8.726455  9.00633  8.973156  9.149897  9.245039  9.298647  9.113609
944895 11.179725 11.09959 11.270414 10.792218 10.750200 11.289802 10.960788
947758 12.399980 12.50940 12.043803 12.460848 12.531210 12.440010 12.510939
```

Before starting our biRte analysis we try to simplify the TF-target by clustering regulators with highly overlapping target gene sets. Then we determine possible interactions between regulators by looking for regulators, which have an overlap that is large enough to be considered, but not as large that the effect is indistinguishable from main effects by individual regulators.

Afterwards, differentially expressed genes are calculated using `limmaAnalysis`. The result is then passed to `biRteLimma`, together with the TF-target graph `EColiNetwork`. As a final step we use biRte to look for regulator activities that can explain differential gene expression between anaerobic and aerobic growth. In a real application the number of MCMC iterations should be increased significantly:

```
> # prepare network
> affinities = list(TF=sapply(names(EColiNetwork$TF), function(tf){w = rep(1, length(EColiNetwork
> affinities = simplify(affinities)
> affinities$other = proposeInteractions(affinities)
> # prepare data
> colnames(exprs(EColiOxygen)) = make.names(paste(pData(EColiOxygen)$GenotypeVariation, pData(ECo
> limmamRNA = limmaAnalysis(exprs(EColiOxygen), design=NULL, "wild.type.anaerobic - wild.type.aer
> mydat = cbind(exprs(EColiOxygen)[,colnames(exprs(EColiOxygen)) =="wild.type.aerobic"],  exprs(E
> ecoli_result = birteLimma(dat.mRNA=mydat, limmamRNA=limmamRNA, affinities=affinities, niter=500

> plotConvergence(ecoli_result, title="E. Coli")
```

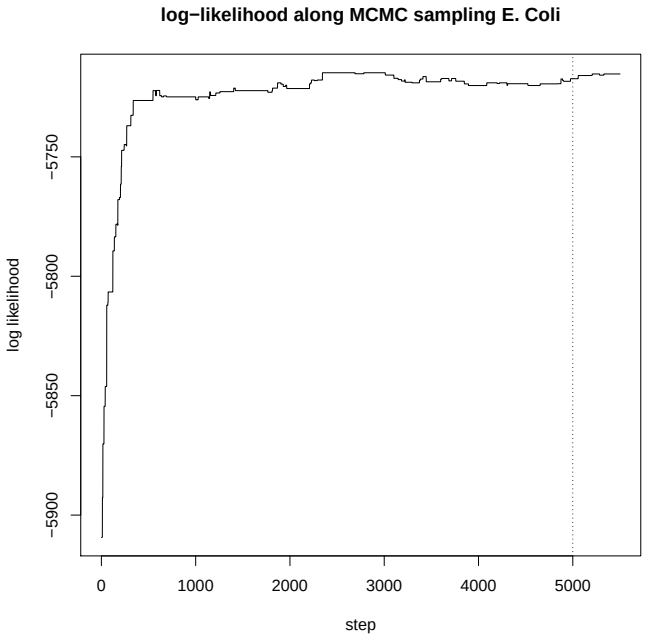

Figure 1: Log-likelihood during MCMC sampling for the E. Coli data set.

The log-likelihood is shown in Figure 1. Below we show those TFs, who reveal a marginal activity probability of larger than a cutoff corresponding to an expected false positve rate of 0.001. We look at the total number of target genes together with the number of differentially expressed target genes for the predicted TFs:

```
> tau = suggestThreshold(ecoli_result$post[,1])

start values: (alpha, beta) =  1.5 5 lambda =  0.5 0.5
logLik =  -1032.839 (alpha, beta) =  149.4132 149.4132 lambda =  0.08564836 0.9143516
logLik =  -1032.853 (alpha, beta) =  149.4132 149.4132 lambda =  0.08264463 0.9173554
[1] "converged!"

> activeTFs = rownames(ecoli_result$post)[ecoli_result$post[,1] > tau]
> activeTFs

 [1] "caiF"             "fhlA"             "fruR"             "fur"
 [5] "galS"             "gatR"             "gcvA"             "glpR"
 [9] "hipB"             "narL"             "uidR"             "arcA_crp"
[13] "argP_dnaA"        "crp_fnr"          "gadW_gadX"        "arcA_ihfA_U_ihfB"
[17] "fnr_ihfA_U_ihfB"  "narL_narP"        "argP_nrdR"        "dnaA_nrdR"
[21] "gadE_pdhR"        "gadE_torR"

> if(length(activeTFs) > 0){
+         DEgenes = rownames(limmamRNA$pvalue.tab)[limmamRNA$pvalue.tab$adj.P.Val < 0.05 & abs(li
+         genesetsTF = c(sapply(affinities$TF, names), sapply(affinities$other, names))
+         DEgenesInTargets = sapply(genesetsTF[intersect(activeTFs, names(genesetsTF))],
+         function(x) c(length(which(x %in% DEgenes)), length(x)))
+         rownames(DEgenesInTargets) = c("#DEgenes", "#targets")
+         DEgenesInTargets[,order(DEgenesInTargets["#targets",], decreasing=TRUE)]
+ }

          narL fnr_ihfA_U_ihfB fur crp_fnr arcA_crp narL_narP arcA_ihfA_U_ihfB
#DEgenes    31              15   1      10        4        12                2
#targets   101              85  78      77       57        43               40
          fruR fhlA caiF glpR galS gatR gadW_gadX gadE_pdhR gadE_torR argP_dnaA
#DEgenes     3   15    0    0    0    0         2         0         1         0
#targets    36   29   10    9    8    6         5         5         5         4
          gcvA uidR argP_nrdR dnaA_nrdR hipB
#DEgenes     0    0         0         0    0
#targets     3    3         2         2    1
```

We can ask, how well log fold changes predicted by our biRte model agree with obseved log fold changes:

```
> pred = birtePredict(ecoli_result, rownames(mydat))
> cor(pred[[1]][[1]]$mean, limmamRNA$pvalue.tab[rownames(mydat), "logFC"])

[1] 0.4306204
```

Once again it should be noted that in a real application the MCMC sampler should run much longer and hence better results are expected.

5 Using Regulator Expression Data

One of the strength of *biRte* is that measurements of regulators can be integrated smoothly into the inferenceprocess.

In our example situation no miRNA expression data is available, but some transcription factors have been measured on the microarray. In accordance with published results [7], *biRte* does not suppose that the mRNA expression levels of a TF and its (putative) target genes are correlated. However, differential TF expression on mRNA level might still give a hint on activity differences on protein level. Thus, *biRte* allows to integrate expression data of differentially expressed TFs.In our case `TFexpr` contains an excerpt of `EColiOxygen`. It comprises mRNA expression for all 160 TFs in `EColiNetwork`. The row names of the expression matrix were converted to the corresponding TF identifiers in `EColiNetwork`.

```
> head(exprs(TFexpr))
```

```
      GSM18261  GSM18262  GSM18263  GSM18286  GSM18287  GSM18288  GSM18289
acrR  8.277473  8.309069  8.504610  7.857166  7.686808  8.111077  7.915678
ada   9.277946  9.540328  9.186303  9.578132  9.646316  9.444881  9.384217
adiY  6.330554  6.555999  6.686157 10.801038 10.986309  8.788498  8.612713
agaR 10.854649 10.726303 10.782988 10.936007 11.041171 11.200971 11.130323
allR 11.324718 11.193124 11.389784 11.102606 11.274170 11.273160 10.936546
allS  8.520564  8.764251  8.693574  8.806117  8.806997  8.705715  8.272239
```

Differential expression of these TFs can be assessed by subsetting our previous `limmamRNA` object. We use the obtained results to define an informative prior for each regulator and regulator-regulator interaction, before running a biRte analysis, and to set up a reasonable initial state for the sampler:

```
> limmaTF = limmamRNA
> limmaTF$pvalue.tab = limmaTF$pvalue.tab[rownames(limmaTF$pvalue.tab) %in% fData(TFexpr)$Entrez,
> names(limmaTF$lm.fit$sigma) = as.character(fData(EColiOxygen)$symbol[match(names(limmaTF$lm.fit
> rownames(limmaTF$pvalue.tab) = as.character(fData(EColiOxygen)$symbol[match(rownames(limmaTF$pv
> diff.TF = rownames(limmaTF$pvalue.tab)[limmaTF$pvalue.tab$adj.P.Val < 0.05 & abs(limmaTF$pvalue
> theta.TF = rep(1/length(affinities$TF), length(affinities$TF))
> names(theta.TF) = names(affinities$TF)
> theta.other = rep(1/length(affinities$other), length(affinities$other))
> names(theta.other) = names(affinities$other)
> theta.other[unique(unlist(sapply(diff.TF, function(tf) grep(tf, names(theta.other)))))] = 0.5 #
> init.TF = theta.TF
> init.TF = (init.TF >= 0.5)*1
> init.other = theta.other
> init.other = (init.other >= 0.5)*1
> # note that niter and nburnin are much too small in practice
> ecoli_TFexpr = birteLimma(dat.mRNA=mydat, data.regulators=list(TF=exprs(TFexpr)), limmamRNA=lim

> tau = suggestThreshold(ecoli_TFexpr$post[,1])

start values: (alpha, beta) =  1.5 5 lambda =  0.5 0.5
logLik =  -888.6431 (alpha, beta) =  10.51383 149.4132 lambda =  0.1977083 0.8022917
logLik =  -888.6604 (alpha, beta) =  10.51383 149.4132 lambda =  0.2024793 0.7975207
[1] "converged!"

> activeTFs = ecoli_TFexpr$post[ecoli_TFexpr$post[,1] > tau,1]
> activeTFs

    adiY     appY     arcA     caiF
   1.000    1.000    1.000    1.000
    dicA     gadE     lrhA     narP
   0.608    1.000    1.000    1.000
```

nikR	rstA	tyrR	uhpA
0.866	1.000	0.924	1.000
zntR	betI	bolA	cusR
0.872	1.000	1.000	1.000
fnr	feaR	fur	glcC
1.000	1.000	1.000	1.000
hcaR	iscR	lldR	marA
1.000	1.000	1.000	1.000
mhpR	mntR	narL	dhaR
1.000	1.000	1.000	0.048
fhlA	fruR	galS	gatR
1.000	1.000	1.000	1.000
gcvA	glpR	iclR	arcA_crp
1.000	1.000	1.000	1.000
cbl_cysB	arcA_fnr	crp_fnr	gadE_gadW
0.334	1.000	1.000	1.000
gadE_gadX	appY_iscR	iscR_narL	appY_narP
0.760	0.442	0.696	0.420
fnr_narP	iscR_narP	dnaA_nrdR	iscR_oxyR
0.048	0.680	1.000	1.000
gadE_pdhR	ompR_rstA	rcsA_U_rcsB_rstA	gadE_torR
1.000	0.078	1.000	1.000

6 Network Inference

After having determined active regulators one may ask, in which way these regulators influence each other. Bayesian Networks are a principal possibility, but would usually require direct measurements of regulators, which is difficult to obtain for TFs. Moreover, the typically small sample size imposes a principal limitation. We thus restrict ourselves to subset relationships between differentially expressed target genes. These subset relationships can have two possible interpretations: One possibility is that egulator A acts upstream of regulator B, if differential targets of B are a subset of those of A. Another possibility is that A and B jointly co-regulate certain target genes. The idea of (noisy) subset relationships has striking similarities to Nested Effects Models (NEMs) [**?**, 4], which have been introduced for causal network inference from perturbation data. Although in our case we do not have targeted perturbations of individual regulators, probabilistic inference of subset relationships between differentially expressed targets of regulator pairs can be effectively solved via NEM inference. *biRte* uses the pair-wise inference algorithm discussed in [6] as default.

biRte offers a convenience function `estimateNetwork` for this purpose. The function decomposes clusters of active regulators into individual regulators and performs appropriate calls to functions from *nem* [3]. The output is a network indicating subset relationships between differential targets of active regulators. In our example this would be done as follows:

```
> DEgenes = rownames(limmamRNA$pvalue.tab)[limmamRNA$pvalue.tab$adj.P.Val < 0.05 & abs(limmamRNA$
> net = estimateNetwork(ecoli_TFexpr, thresh=tau, de.genes=DEgenes)
> library(nem)
> if(require(Rgraphviz)){
+   plot(net, transitiveReduction=TRUE)
+ }
```

This yields the network shown in Figure 2. In addition to the network structure we can investigate the estimated dependencies regulator-gene dependencies in more depth. This may give additional insights whether a particular gene is a direct target of a particular transcription factor or not and hence allow for filtering out false positive target predictions:

```
> net$mappos
```

In our case there are several totally unspecific target genes (assigned to "null"), which may indicate false positive target gene predictions.

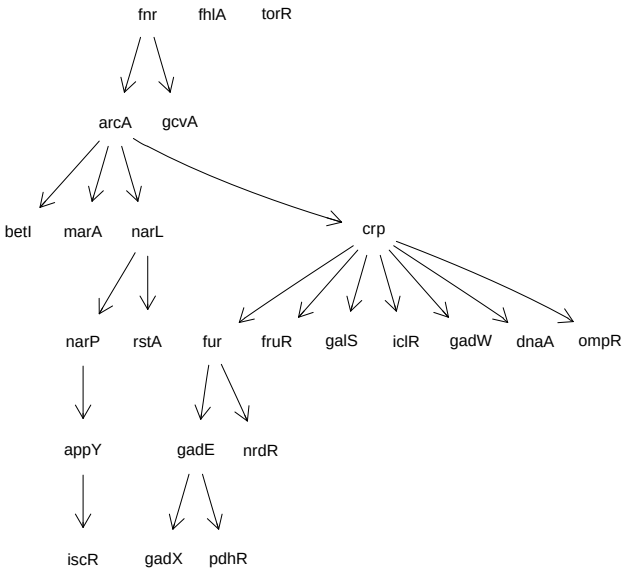

Figure 2: Inferred network between active TFs.

7 Conclusion

biRte integrates regulator expression and mRNA data into a probabilistic framework to make inference on regulator activities. It is a step towards the important goal to unravel causal mechanisms of gene expression changes under specific experimental or natural conditions. A unique feature is the combination with network inference.
This vignette was generated using the following package versions:

- R version 3.2.0 alpha (2015-03-20 r68043), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, grDevices, graphics, grid, methods, parallel, stats, utils
- Other packages: Biobase 2.27.3, BiocGenerics 0.13.11, Rcpp 0.11.5, RcppArmadillo 0.4.650.1.1, Rgraphviz 2.11.2, birte 1.1.0, graph 1.45.3, nem 2.41.2
- Loaded via a namespace (and not attached): MASS 7.3-40, RBGL 1.43.0, RColorBrewer 1.1-2, boot 1.3-16, class 7.3-12, e1071 1.6-4, evaluate 0.5.5, formatR 1.1, knitr 1.9, limma 3.23.11, plotrix 3.5-11, ridge 2.1-3, statmod 1.4.21, stats4 3.2.0, stringr 0.6.2, tools 3.2.0

References

[1] R. Castelo and A. Roverato. Reverse engineering molecular regulatory networks from microarray data with qp-graphs. *J Comput Biol*, 16(2):213–227, Feb 2009.

[2] M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429(6987):92–96, May 2004.

[3] H. Fröhlich, T. Beißbarth, A. Tresch, D. Kostka, J. Jacob, R. Spang, and F. Markowetz. Analyzing gene perturbation screens with nested effects models in R and bioconductor. *Bioinformatics*, 24(21):2549–2550, Nov 2008.

[4] H. Fröhlich, A. Tresch, and T. Beissbarth. Nested effects models for learning signaling networks from perturbation data. *Biom J*, 51(2):304–323, Apr 2009.

[5] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biol*, 15(2):R29, 2014.

[6] F. Markowetz, D. Kostka, O. Troyanskaya, and R. Spang. Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*, 23:i305 – i312, 2007.

[7] M. Wu and C. Chan. Learning transcriptional regulation on a genome scale: a theoretical analysis based on gene expression data. *Brief Bioinform*, May 2011.

[8] B. Zacher, K. Abnaof, S. Gade, E. Younesi, A. Tresch, and H. Fröhlich. Joint bayesian inference of condition-specific mirna and transcription factor activities from combined gene and microrna expression data. *Bioinformatics*, 28(13):1714–1720, Jul 2012.