Uncovering gene regulatory relationships from knockdown expression data using BayesKnockdown

William Chad Young, Ka Yee Yeung, and Adrian E. Raftery
Department of Statistics (WCY and AER) and Institute of Technology (KYY)
University of Washington

This document illustrates the use of the BayesKnockdown R package to calculate posterior probabilities of relationships between a single predictor and multiple potential targets. The package was developed specifically for gene expression datasets in the form of knockdown experiments, but can be applied more generally to other over-expression data and to infer differential expression.

1 Posterior Probabilities

Given a predictor x and a set of possible targets y, the BayesKnockdown function can be invoked to estimate the posterior probabilities of a relationship between x and each individual target in y [2]. The BayesKnockdown function allows specification of a prior probability of regulation via the prior argument, and it can be a constant for all targets or unique to each target. This is useful particularly when an informative prior is available to incorporate additional knowledge. The prior is set to 0.5 by default, which corresponds to an uninformative prior.

Additionally, the method allows specification of Zellner's g-prior via the g argument [3]. The g-prior specifies the expected strength of the signal relative to noise, with larger values corresponding to a larger expected signal. It is recommended that g be set to a value between 1 and the number of observations in the data. The default value is $\sqrt{n}$, which we have found to be a good compromise between the extremes.

1.1 Simulated Data Example

As a simple example of using the BayesKnockdown function, we generate random data for the knockdown gene as well as the potential targets. We then introduce a relationship between the knockdown gene and target number 3. The BayesKnockdown function takes this data and produces the posterior probability of a relationship between x and each target. Figure 1 shows the posterior probabilities calculated for each target.

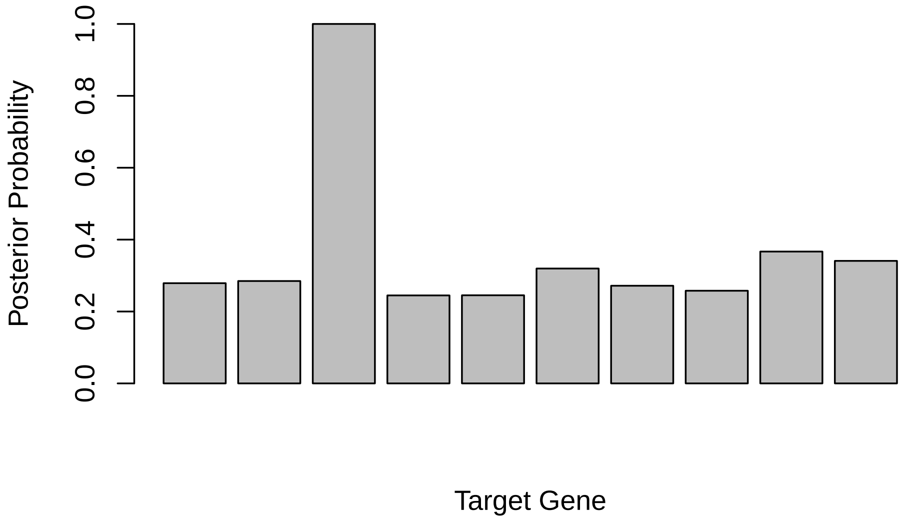

Figure 1: Bar plot showing the posterior probabilities of a relationship between the knockdown gene and each target in simulated data. Gene 3 is the only true relationship.

```
> library(BayesKnockdown);
> set.seed(1618);
> n <- 100;
> p <- 10;
> x <- rnorm(n);
> y <- matrix(nrow=p, data=rnorm(n*p));
> y[3,] <- y[3,] + 0.5*x;
> simResult <- BayesKnockdown(x, y);
> simResult;

 [1] 0.2788326 0.2847752 1.0000000 0.2448188 0.2451588 0.3195648 0.2717004
 [8] 0.2578031 0.3667949 0.3408614

> barplot(simResult, names.arg="", xlab="Target Gene",
+         ylab="Posterior Probability", ylim=c(0,1));
```

1.2 Knockdown Data Example

A more realistic example uses data from the National Institute of Health (NIH) Library of Integrated Network-based Cellular Signatures (LINCS) program (`http://lincsproject.org`) [1]. The aim of this program is to generate genetic and molecular signatures of cells in response to various perturbations. To support this endeavor, many large datasets have been made available, including proteomic and imaging data.

The LINCS L1000 data capture gene expression levels of 1,000 genes in human cell lines under a variety of conditions. The `lincs.kd` data is a 21 by 27 matrix containing data from knockdown experiments targeting gene PPARG in cell line A375. Cell line A375 is a human skin melanoma cell line with over 100,000 experiments in the L1000 data. The first row is the expression levels of PPARG in the 27 experiments targeting PPARG for knockdown, while the other 20 rows are a subset of the measured genes in the same experiments. The data have been normalized to account for differences in the experimental settings, as described in [2]. The full LINCS L1000 data is available at `http://lincscloud.org`.

Given the L1000 data, the `BayesKnockdown` function can be invoked to calculate the posterior probabilities of a relationship between gene PPARG and the other genes in the dataset. In this case, we specify a prior probability of 0.0005, reflecting the belief that there are very few relationships relative to the total possible number. Figure 2 shows the range of values returned for the different target genes.

```
> data(lincs.kd);
> kdResult <- BayesKnockdown(lincs.kd[1,], lincs.kd[-1,], prior=0.0005);
> kdResult;
```

```
        ATF1     SERPINE1        CEBPA         MUC1         EZH2        SNX13 
0.9959445881 0.0271544446 0.0007644977 0.0090443118 0.9637199199 0.0504674678 
      ELOVL6        CASC3       MRPL12        KIF2C        BCL7B        PRAF2 
0.0002019284 0.8210955027 0.0048439740 0.9986842921 0.9997559484 0.0005646142 
        NET1       ATP1B1        H2AFV      TIMM17B       ZNF586         RFNG 
0.0002563962 0.0046399699 0.0222102583 0.0004304130 0.0153763269 0.0003694639 
       CDK19       SFMBT1 
0.9990571273 0.9053986408 
```

```
> barplot(kdResult, names.arg="", xlab="Target Gene",
+         ylab="Posterior Probability", ylim=c(0,1));
```

1.3 ExpressionSet Example

The `BayesKnockdown.es` function allows calculation of posterior probabilities using an `ExpressionSet` object from the `bioBase` library. The function works similarly to the `BayesKnockdown` function, except that one of the features of the `ExpressionSet` is identified to be the predictor variable, and all other features are used as response variables.

```
> library(Biobase);
> data(sample.ExpressionSet);
```

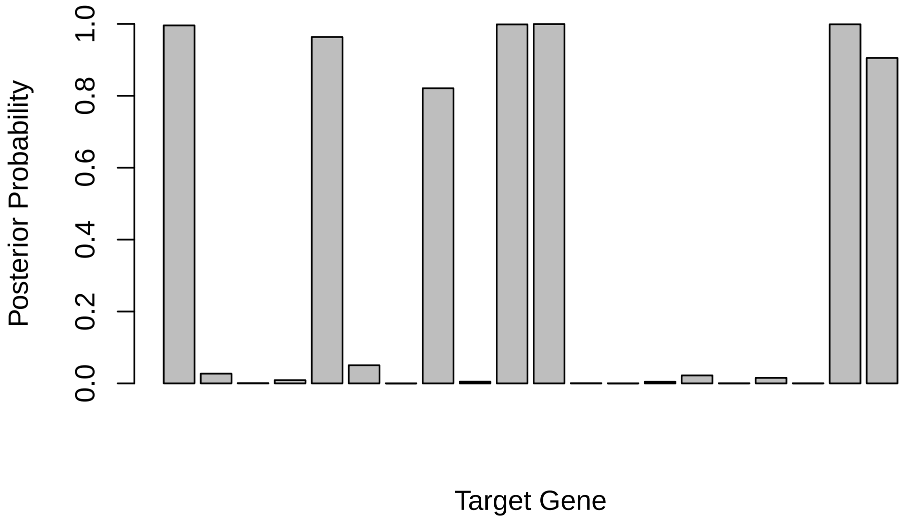

Figure 2: Bar plot showing the posterior probabilities of a relationship between the knockdown gene PPARG and each target in LINCS L1000 data.

```
> subset <- sample.ExpressionSet[1:10,];
> BayesKnockdown.es(subset, "AFFX-MurIL10_at");
```

```
 AFFX-MurIL2_at  AFFX-MurIL4_at  AFFX-MurFAS_at  AFFX-BioB-5_at  AFFX-BioB-M_at
      0.3418659       0.3361832       0.7430095       0.3940327       0.4110827
 AFFX-BioB-3_at  AFFX-BioC-5_at  AFFX-BioC-3_at AFFX-BioDn-5_at
      0.6523990       0.3267071       0.3181790       0.7516404
```

2 2-Class Data

The `BayesKnockdown.diffExp` function tests for differential expression in a set of variables between two experimental conditions. In gene expression data, this often takes the form of comparing the effects of a drug perturbation compared to a baseline. Of interest is the set of genes which show different expression levels between the two conditions. The `BayesKnockdown.diffExp` function takes two matrices of observations for a set of variables, one matrix for each condition, and gives posterior probabilities that the variables are different between the two conditions.

As an example, we generate two random datasets for 10 genes, corresponding to different experimental conditions. The first has 25 observations and the second has 30. We add an offset for gene 3 in the second dataset, reflecting a change of expression between the two conditions. The `BayesKnockdown.diffExp` function produces posterior probabilities for each gene reflecting how likely they are to be expressed differently between the two conditions.

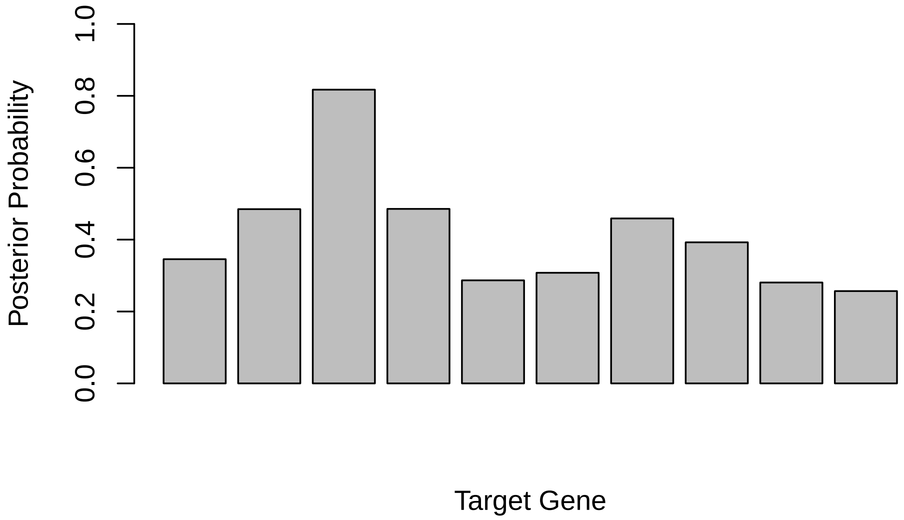

Figure 3: Bar plot showing the posterior probabilities that each gene is differentially expressed between two conditions. Gene 3 is the only gene which is actually differentially expressed.

Figure 3 shows the posterior probabilities that each gene is differentially expressed between the two conditions.

```
> n1 <- 25;
> n2 <- 30;
> p <- 10;
> y1 <- matrix(nrow=p, data=rnorm(n1*p));
> y2 <- matrix(nrow=p, data=rnorm(n2*p));
> y2[3,] <- y2[3,] + 1;
> diffExpResult <- BayesKnockdown.diffExp(y1, y2);
> barplot(diffExpResult, names.arg="", xlab="Target Gene",
+         ylab="Posterior Probability", ylim=c(0,1));
```

3 Acknowledgements

Funding: This research was supported by National Institutes of Health grants [R01 HD054511 and R01 HD070936 to A.E.R., U54 HL127624 to A.E.R. and K.Y.Y.]; Microsoft Azure for Research Award to K.Y.Y.; and Science Foundation Ireland ETS Walton visitor award 11/W.1/I207 to A.E.R.

References

[1] Q. Duan, C. Flynn, M. Niepel, M. Hafner, J. L. Muhlich, N. F. Fernandez, A. D. Rouillard, C. M. Tan, E. Y. Chen, T. R. Golub, and others. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Research*, page gku476, 2014.

[2] W. C. Young, K. Y. Yeung, and A. E. Raftery. A posterior probability approach for gene regulatory network inference in genetic perturbation data. *arXiv preprint arXiv:1603.04835*, 2016.

[3] A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, 6:233–243, 1986.