Bayesian Variable Selection Of Pathway Gene Sets Applied To Gene Expression Microarray Data

A. Skarman[*], M.M. Shariati[*], P. Sørensen[*]

Introduction

For some diseases or other phenotypes it is not only interesting to identify which genes are involved but also to find the involved processes or pathways. This is often done using sets of genes such as genetic pathways found in KEGG: Kyoto Encyclopedia of Genes and Genomes (Kanehisa, M., and Goto, S. (2000)). Typically a gene set analysis is performed, for example using methods like the gene set enrichment analysis (Mootha, V.K, Lindgren C.M., and Eriksson K.F. *et al.* (2003) Subramaniana, A., Tamayoa, P., and Moothaa, V.K. *et al.* (2005)).

Often genes are involved in many different pathways and biological processes. This potentially leads to a substantial overlap between the gene sets. This is often ignored in the current methods used for gene set analysis. Gene set overlap could lead to correlations of the gene sets that is against the assumptions of the gene set tests. Another potential problem caused by the overlap is that it may be difficult to identify which of the pathways are key player(s). It is therefore important to consider methods that could reduce the gene set lists by finding gene sets that are redundant or identify gene sets that are only identified because they overlap with gene sets that are realy important. This could be done using linear models (Hahne, F., Huber, W., and Gentleman, R. *et al.* (2008)).

The method presented is a Bayesian variable selection method (George E.I., and McCulloch R.E. (1993)). This method takes into account the overlap between the gene sets. Instead of first finding the differentially expressed gene sets and then try to find out how the gene sets overlap and thereby correcting for the correlations between the gene sets this method is supposed to take this in just one step. The method is presented on data from a microarray study on bovine Mastitis expression data from udder samples.

Material and methods

The idea was that one could use measures of the differential expression for probe sets belonging to different KEGG pathways and use the Bayesian variable selection method, implemented in the iBay software (Janss, L.L.G. (2009)) to rank the pathways. Pathway annotations where taken from KEGG pathway database. The genes were associated to the

[*] Faculty of Agricultural Sciences, Department of Genetics and Biotechnology, Aarhus University, Blichers Alle 20, 8830 Tjele, Denmark

different gene sets. Gene sets containing less than 5 genes were removed to avoid noise in the further analysis.

Model. For ranking gene sets involved in gene expression data, a stochastic search variable selection (SSVS, George and Maccolluch, 1993) in a linear model is introduced. The linear model relating expression data (in this case t-test statistics) to different gene sets is

$$y_i = \mu + \sum_j X_{ij}\beta_j + e_i \quad (1)$$

Where y_i is the t-test statistic for probe *i*, μ is a general mean, β_j is the effect of *j*th gene set and is assumed to be randomly distributed with $\beta_j \sim N(0, \varphi_i^2)$, X_{ij} is a known indicator variable associating effects of gene sets β_{ij} to the expression data y_i, $\boldsymbol{e}_i$ is residual with $e_j \sim N(0, \varphi_e^2)$, and φ_i is scaling factor that models the magnitude of variance of the *j*th gene set. It is assumed that the prior distribution of φ_i is a two-component mixture of truncated normal distributions as follows:

$$\varphi_i \sim \begin{cases} \pi_0 N(0, \sigma_{s_0}^2) \\ \pi_1 N(0, \sigma_{s_1}^2) \end{cases} \qquad \varphi_2 > 0 \text{ and } \pi_1 = 1 - \pi_0 \qquad (2)$$

where π_1 determines which portion of gene sets are "switched on" with variance $\sigma_{S_1}^2$ that is much larger than the predefined variance of "switched off" gene sets $\sigma_{S_0}^2$. It should be noticed that if one sets $\pi_1 = 1$, the mixture prior distribution in (2) becomes a common prior truncated normal distribution. A Gibbs sampling procedure was used with a chain length of 40000 in total and 5000 as burn in.

Alternative approach. We compared the results from our Bayesian variable selection procedure described above with a simple method commonly used for gene set analysis. This method tested pathways gene sets by the use of ANOVA: To ensure only biological relevant results we used a stringent p-value cutoff of 0.0001 adjusted for multiple testing using a the bonferroni correction.

Expression Data. Infection of the udder by *Escherichia coli* (*E.coli*) is usually causing a severe, short term immune response, i.e. acute phase response. To study the genes involved in the acute immune response against *E.coli* infection, 16 Danish HF cows were challenged intra-mammarily with *E.coli* 4 to 6 weeks after parturition. Udder biopsies were collected from the infected quarter and a healthy control quarter of each cow at 24 h (acute stage) post

challenge, and the gene-expression was measured using the Affymetrix Bovine Genome array. Differential expression of each gene was assessed using linear modelling and empirical Bayes methods, which were implemented using the R (R Development Core Team (2009))package Limma (version 2.18.3) (Gentleman, R., Carey, V., Dudoit, S., *et al.* (2005)). The linear models allowed for changes for time points. The time points were 24h and 192h. The contrast tested was 24h vs 192h post infection of the infected quarters. Each transcript targeted by a probe was tested for its expression change using a modified t-test.

Results and discussion

From Figure 1 below one can se some overlapping pathways are overlapping with others but only one or a few have a posterior equals to one. This could be because variable selection method takes the overlaps into account

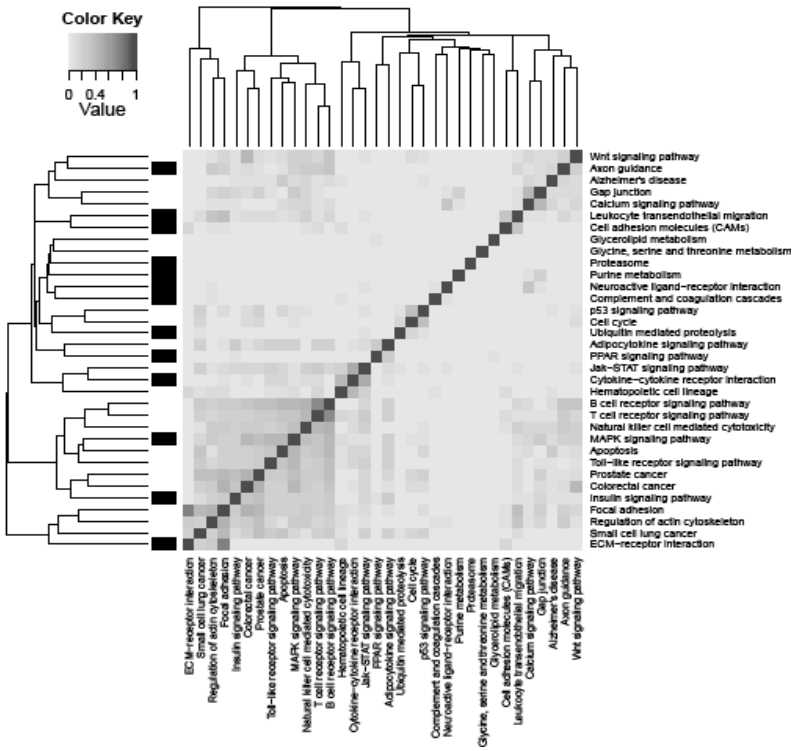

Figure 1: Heatmap showing the overlap between each pathway with a corrected p-value below 0.0001. The black bars on the left side show which pathways that got a posterior probability of one.

One illustrative example is the pathway "Focal adhesion" which contains 131 genes. It is considered to be highly significant in the ANOVA test which ignores the overlap between gene sets. From the results of the Bayesian variable selection procedure it has a very low posterior probability (p=0.012) of being included in the model. Therefore it does not appear to be an important player in the acute phase response in the udder It turns out that it has large overlap with the pathway "ECM-receptor interaction" by 41 genes. In total "ECM-receptor interaction" has 53 genes. The big overlap is likely the main reason why "Focal adhesion" has a low posterior.

Conclusion

We presented a variable selection method that takes the overlaps between gene sets into account. This potentially leads to better identification of the key pathways involved in the studies diseases.

References

Gentleman, R., Carey, V., Dudoit, S., *et al.* (2005) *Bioinformatics and Computational Biology Solutions using R and Bioconductor,* ISBN: 978-0-387-25146-2 page 397-420

George E.I., and McCulloch R.E. (1993) *Journal of the American Statistical Association,* 88:881-889.

Hahne, F., Huber, W., and Gentleman, R. *et al.* (2008) *Bioconductor Case Studies,* ISBN: 978-0-387-77239-4 page 193-205.

Janss, L.L.G. (2009) **"***iBay manual version 1.47***"**

Kanehisa, M., and Goto, S. (2000) *Nucleic Acids Res.* 28:27-30

Mootha, V.K, Lindgren C.M., and Eriksson K.F. *et al.* (2003) *Nature Genetics* 34:267 – 273.

R Development Core Team (2009). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Subramaniana, A., Tamayoa, P., and Moothaa, V.K. *et al.* (2005) *PNAS,* 102:15545-15550.