Prediction of Gene Expression in Human Using Rat *in Vivo* Gene Expression in Japanese Toxicogenomics Project

Martin Otava[1], Ziv Shkedy[1]

[1] *Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Center for Statistics, Universiteit Hasselt, Belgium*
martin.otava@uhasselt.be

Abstract

Motivation

Japanese Toxicogenomics Project (TGP) represents unique source of information for toxicology and safety challenges. The main topic that we address in this paper is related to the prediction of drug-induced liver injury (DILI) in humans using rat data. Successful prediction enables to stop the trial before even reaching human patients which would have high economical impact together with saving patients of side effects. Our aim is to explore connection between human data and rat data in even broader sense.

Core part of the rat data is gene expression level information across multiple compounds with multiple time points and dose levels. A subset of genes is common for rat and human and lot of the genes are already connected with some biological processes or diseases. The analysis presented in this paper is focused on the question if there exist a subset of genes that their response to the treatment is similar in rat and human. In this case, we would be able to predict human gene expression level using *in vivo* rat experiment and, similarly as in DILI case, predict properties of drug if used in human patients.

Data sets

The data considered for the analysis presented in this paper consists of 93 compounds that are common in rat *in vivo* and human experiment and have DILI information for possible use of the DILI indicator as covariate. In total, 4440 arrays are available for rat (91 compounds with 48 arrays and 2 compound with 36 arrays) and 1116 arrays are available for human (12 arrays per compound). We focus on genes that are common for rat and human (i.e. their gene names are same) and are filtered using the I/NI calls criterion

(Kasim *et al.* 2010). The final data set consists of 4359 genes. Response is computed as log ratio of the gene expression level against mean of expression levels under control dose (vehicle). The gene expression values are based on FARMS (Hochreiter *et al.* 2006) summarized data.

For each compound were arrays for rat measured in 4 doses (including control), each in 4 different time points (2 compounds with 36 arrays miss highest dose). In human, for each compound were arrays measured in 3 doses and 2 time points. The particular values of time points and doses vary among compounds. For the analysis presented in this paper we use the ordinal dose levels, i.e., low, middle or high that is provided in original data set as well. Time points are treated as factor, i.e., with respect to their ordering.

Methodology

We consider two different analyses for the TGP data. The first analysis is based on two-way ANOVA model and the goal is to detect genes with significant response to the treatment in both human and rat. The second analysis consists of a trend analysis at each time point and the goal of the analysis is to detect genes in rat that can be used to predict gene expression is human.

For the first analysis, a gene specific, linear model with dose and time as covariates is used. Interaction between covariates is included as well. Significance of covariates and overall F-test significance is considered and multiplicity adjustment is applied. Group of genes significant for both rat and human are identified under several settings (overall significance, significant interaction, any dose effect, etc.). Family wise error rate (FWER, Hochberg and Tamhane 1987) using the Bonferroni method is used for multiplicity adjustment. Resulting gene lists can be compared across compounds. Indicator of significance of particular gene can be compared with DILI status of compound.

A trend analysis is a common analysis in toxicology. The aim of such analysis is to identify a subset of genes for that a monotone relationship with dose can de detected (Lin *et al.* 2012). Hence, within the second modeling approach the null hypothesis of no dose effect is tested against an ordered alternative. The analysis was done per compound and per time point. Multiple contrast test with Marcus' contrast (MCT, Mukerjee *et al.* 1987) is used to identify significant genes and multiplicity adjustment is conducted using FWER approach (with Bonferroni correction). For a particular gene, isotonic means in each dose are estimated and their values are compared between human and rat. Hence, we can identify genes in rat that can be used in order to predict the gene expression level in human. Especially, we focus on last time point in both rat and in human.

Results

Figure 1 shows the number of genes with significant interaction effects in both rat and human and reveals a heterogenous pattern across compounds. For example, for the compound sulindac there are 201 genes with significant interaction in both rat and human while for compound perhexiline there is only 1 gene in common. Example of one significant gene is shown on Figure 2. There exists a subset of genes significant both in rat and human consistently across multiple compounds, even in case of strict multiplicity corrections. These genes are usually present only in DILI connected compounds. Hence, their significance in rat *in vivo* could emphasize danger of DILI in human. Naturally, these genes are typically connected with the liver processes.

As mentioned in previous section, the second analysis consists of trend analysis per time point. As the first stage of the analysis we identify the time point of rat with the strongest signal. Figure 3 present the number of genes with significant dose-response relationship per time point and clearly shows that there are much more significant genes in the last time point for rat and human than in any other time point. Hence, for prediction, the dose effect of rat in the last time point are used. The dose effect of both rat and human can be estimated using isotonic regression (Robertson *et al.* 1988). Only 91 compounds having high dose are considered for the analysis and we use the isotonic mean of the rat in order to predict human isotonic means. The results for the compound omeprazole are shown in Figure 4. We note that the correlation between the rat and human dose effects is higher when we consider only genes that are found to be significant in both rat and human.

References

Hochberg, Y. and Tamhane A.C. (1987). *Multiple comparison procedures*. New York: Wiley.

Hochreiter, S., Clevert D.-A., Obermayer K. (2006). A new summarization method for Affymetrix probe level data. *Bioinformatics*, **22(8)**, 943–949

Kasim, A., Lin, D., Van Sanden, S., Clevert, D.-A., Bijnens, L., Goehlmann, H.W., Amaratunga, D., Hochreiter, S., Shkedy, Z., and Talloen, W. (2010). Informative or Noninformative Calls for Gene Expression: A Latent Variable Approach. *Statistical Applications in Genetics and Molecular Biology*, **9(1)**, Article 4.

Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D., and Bijnens, L. (Ed.) (2012). *Modeling Dose-Response Microarray Data in Early Drug Development Experiments Using R: Order-Restricted Analysis of Microarray Data*. Springer

Mukerjee, H., and Robertson, T., and Wright, F. T. (1987). Comparison of Several Treatments with a Control Using Multiple Contrasts. *Journal of the American Statistical Association*, **82 (399)**, 902-910

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. John Wiley & Sons Ltd.

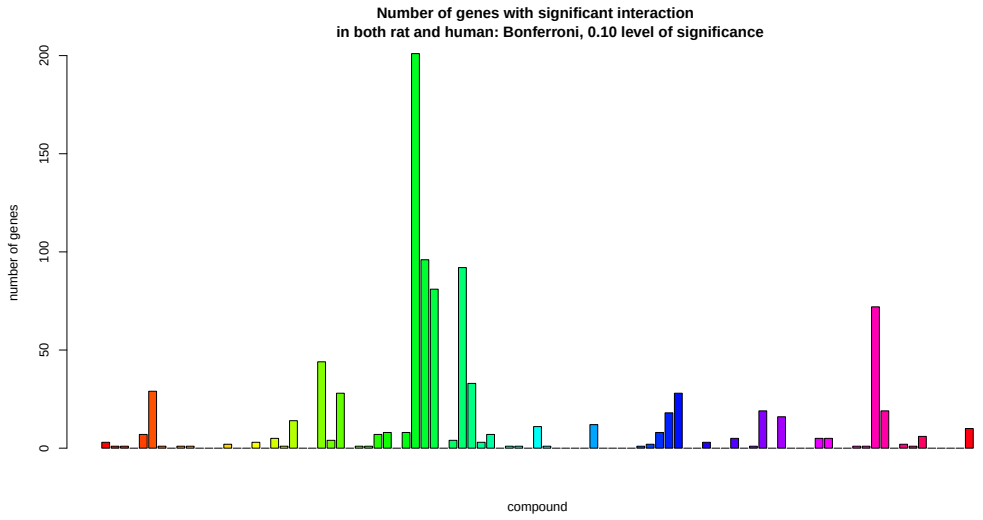


Figure 1: *Number of genes with significant interaction in two-way ANOVA for both rat and human. The p-values are adjusted using Bonferroni's method on significance level 0.10.*

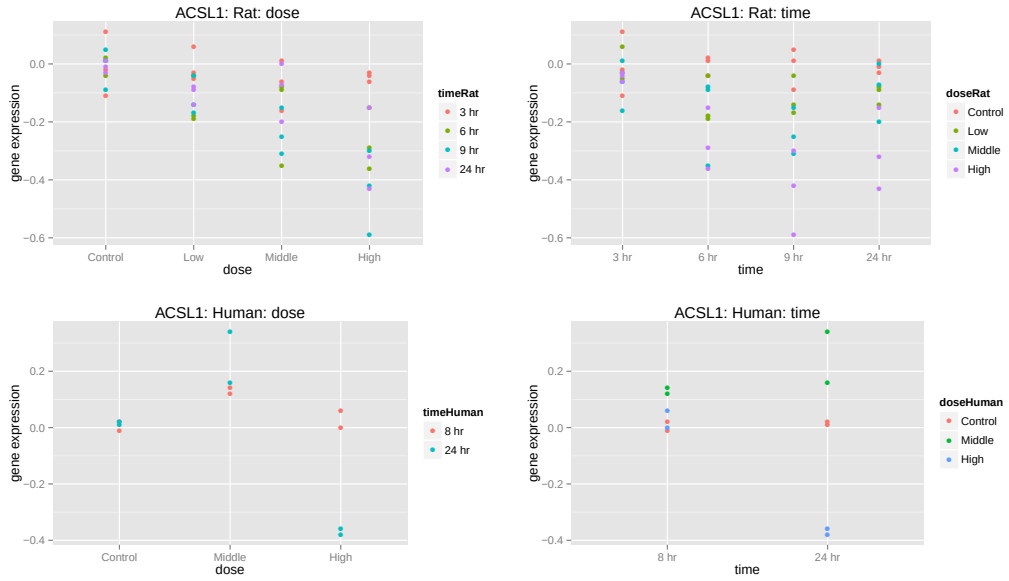


Figure 2: *Example of gene with significant interaction in both human and rat. Compound omeprazole and gene Acsl1 in rat, respectively ACSL1 in human.*

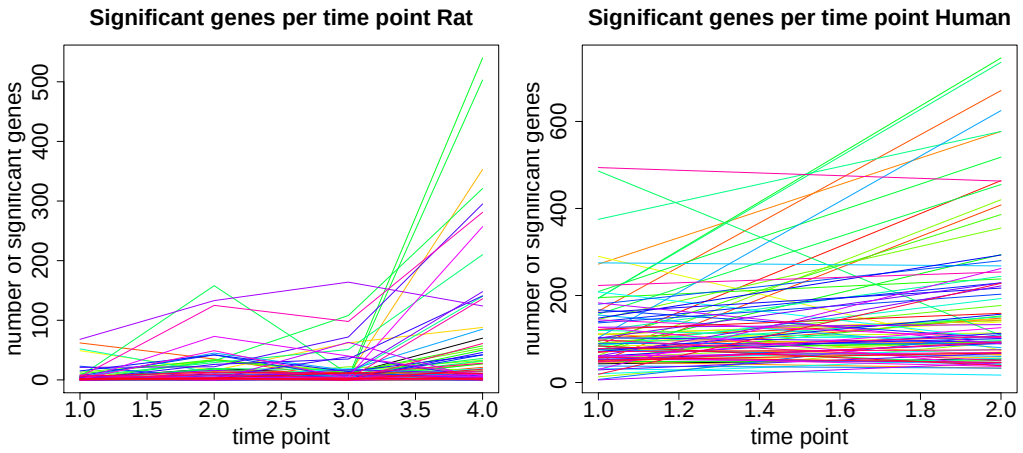
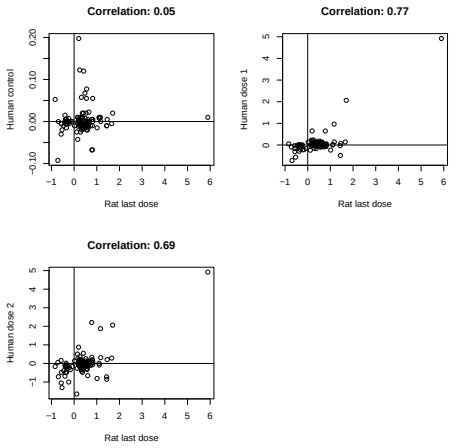
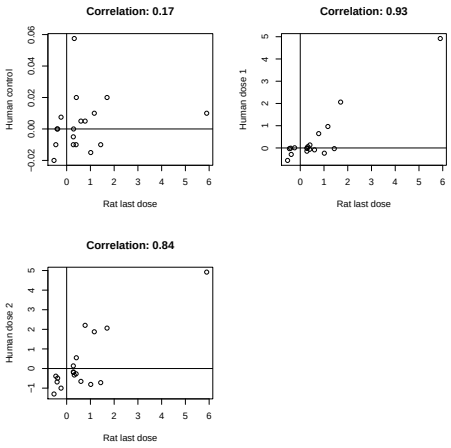


Figure 3: *Number of genes with significant dose-response profile per time point. Test is based on MCT and p-values are adjusted using Bonferroni's method on significance level 0.10. Rat data results are on left panel, human data on right panel.*



(a) *Genes with significant dose-response profile for rat (significance in human not considered).*



(b) *Genes with significant dose-response profile for both rat and human in last time point.*

Figure 4: *Dose effect for the compound omeprazole. Significance of genes is based on MCT adjusted by Bonferroni correction on level 0.10. On the x-axis is estimated isotonic mean in last dose in rat and on y-axis estimated isotonic mean in particular dose in human, both for last time point.*