Exploiting the Japanese Toxicogenomics Project for Predictive Modelling of Drug Toxicity

Djork-Arné Clevert[1†], Martin Heusel[1†], Andreas Mitterecker[1], Willem Talloen[2], Hinrich Göhlmann[2], Jörg Wegner[2], Andreas Mayr[1], Günter Klambauer[1], and Sepp Hochreiter[*1]

[1] Institute of Bioinformatics, Johannes Kepler University Linz, Linz, Austria; [2] Functional Genomics, Johnson & Johnson Pharmaceutical R&D, A Division of Janssen Pharmaceutica, Beerse, Belgium;
† Both authors contributed equally to this work.

Email: Djork-Arné Clevert - okko@clevert.de; Martin Heusel - mhe@gmail.com; Sepp Hochreiter[*]- hochreit@bioinf.jku.at;

[*]Corresponding author

Abstract

Motivation

In the last decade, surprisingly few drugs reached the market. Many promising drug candidates (approx. 80%) failed during or after Phase I, inter alia, due to issues with undetected toxicity [1]. The problem of undetected toxicity becomes even more apparent in the context of drug-induced illness which causes approximately 100,000 deaths per year solely in the USA [2]. Toxicogenomics tries to avoid such problems by prioritizing less toxic drugs over more toxic ones in early drug discovery. To this end, toxicogenomics employs high throughput molecular profiling technologies and predicts the toxicity of drug candidates. For this prediction, large-scale -omics studies of drug treated cell-lines and/or pharmacology model organisms are necessary. However, data exploitation of such large-scale studies requires a highly optimized analysis pipeline, that provides methods for correction of batch effects, noise reduction, dimensionality reduction, normalization, summarization, filtering and prediction.

In this work, we present a novel pipeline for the analysis of large-scale data sets in particular for transcriptomics data. Our pipeline was tested on the Japanese Toxicogenomics Project (TGP) [3], where we evaluated to what degree in vitro bioassays can be used to predict in vivo responses.

The evaluation tasks were to predict drug induced liver injury (DILI) concern [4] and the most prevalent *in vivo* pathological findings from the summarized *in vitro* gene expression values.

Methods and Material

The Japanese Toxicogenomics Project (TGP) is one of the most comprehensive efforts in toxicogenomics, including, among others, gene expression data, toxicological information and pathological data of 131 compounds *in vitro* and *in vivo* screened for toxicity in rat. In the course of the *in vitro* gene expression study, collagen cultured primary hepatocytes, isolated from Sprague-Dawley

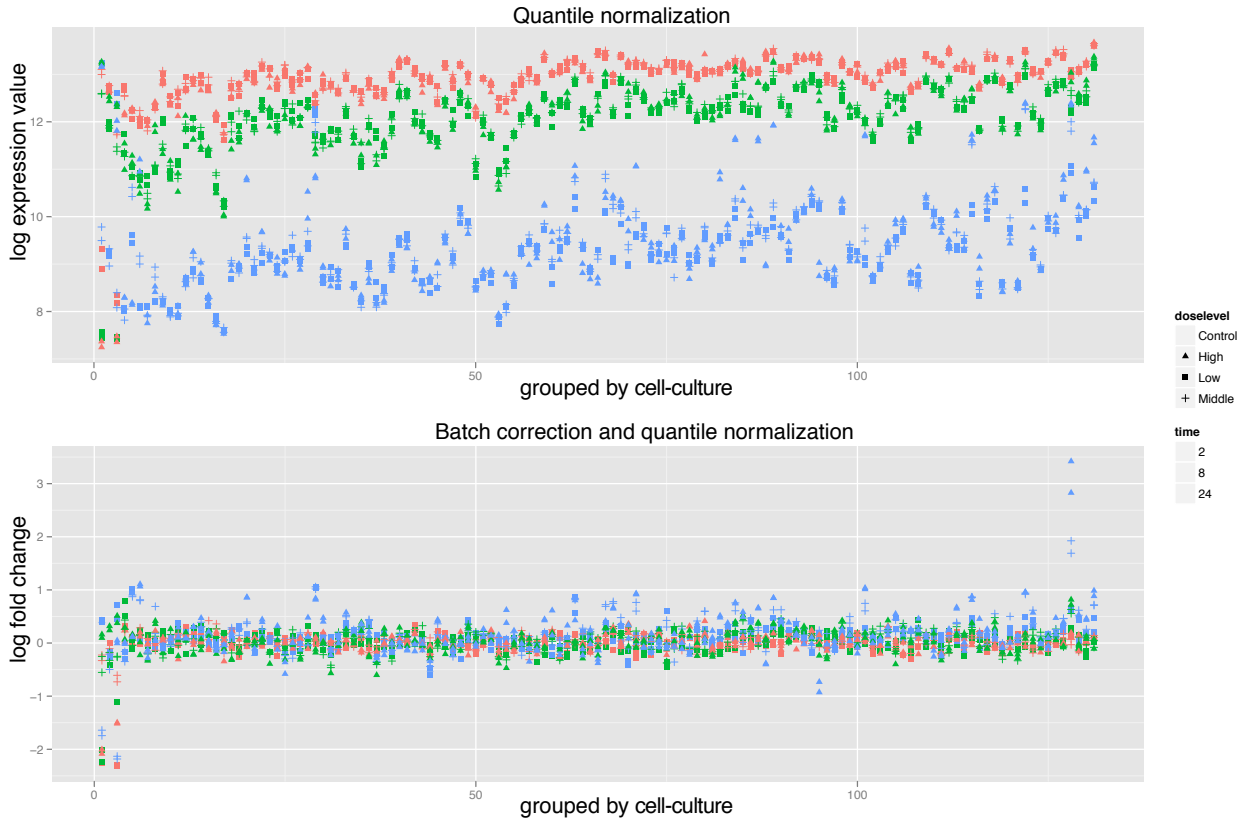

Figure 1: **(Upper panel)** The y-axis shows the log expression values of the fatty acid-binding protein 1 (Fabp1) estimated by FARMS after quantile normalization, while the grouped compounds are shown on the x-axis. The time points are encoded by orange, green and blue for 2h, 8h and 24h, respectively. The plot shows strong cell-culture effects, within the three time points and compounds, which could not be removed by the quantile normalization. **(Lower panel)** Same as upper panel but batch corrected before quantile normalization. The correction with the matched control within cell-culture clearly reduces the cell-culture effects, while compound induced expression changes are preserved.

rats, were treated with three compound concentrations (low, middle and high). To survey transcriptomic changes in response to the compound perturbation, mRNA was isolated at three time points (2h, 8h and 24h) and assayed in two replicates with Affymetrix RAE230_2.0 GeneChip microarrays.

The standard microarray preprocessing procedure consists of normalization, summarization and filtering. However, the standard preprocessing pipeline can not be applied to this data set, as the initial quality control of the microarray data revealed severe batch effects between the cell-cultures (see Figure 1). Therefore, we developed a three step normalization procedure which takes cell-culture batch effects into account. First, the probe-level data of the microarrays were baseline normalized to the same median. Secondly, a cell-culture batch correction was made by calculating probe intensity ratios using the corresponding control measurement for the cell-culture (only vehicle without compound) as reference. Finally, the probe intensity ratios were quantile-normalized [5] across all batches. For the next preprocessing step, summarization, we defined probe sets corresponding to genes using alternative CDFs (Version 15.1.0, ENTREZG) from Brainarray [6] and applied FARMS [7] for summarizing the intensity ratios at probe set level to obtain expression values per gene. Specially for this purpose, a new FARMS software package has been developed, that allows summarization of huge microarray data sets like those of the TGP. For the last preprocessing step, gene filtering, we applied the FARMS based informative/non-informative (I/NI) call [8–10] and excluded all non-informative probe sets.

After this data preprocessing, we predicted drug induced liver injury (DILI) concern and *in vivo* pathological findings for hypertrophy, vacuolization and ground glass appearance from the summarized *in vitro* gene expression values. We combined replicates by concatenating their features. Removing replicates is essential because otherwise the classification task leads to a trivial almost perfect solution in the LOO-CV task, by predicting one replicate by the other. For classification we used the Potential Support Vector Machine (P-SVM) [11] because it is well-suited for data sets with many samples. The PSVM is optimized for many samples since it isbased on a quadratic optimization problem in the number of features instead of the number of samples as standard SVMs. Therefore FARMS' gene filtering and P-SVM prediction are an ideal combination to process massive data sets.

Results

Leave-one-out cross-validation (LOO-CV) of the classifiers (see Table 1) for DILI concern, hypertrophy, vacuolization and ground glass appearance showed a sensitivity of 0.78, 0.81, 0.77 and 0.80 for a specificity of 0.63, 0.65, 0.58 and 0.56, respectively. These results are very promising, as due to the severe cell-culture effects we did not expect to obtain such a high classification performance.

Table 1: Summary of the LOO-CV for DILI concern and various pathological findings (first column "predicted finding"). The second column "# features" gives the number of features used for classification. The third column shows the "error rate" over all LOO runs, while the fourth and fifth columns report the "sensitivity" and "specificity", respectively.

predicted finding	# features	error rate	sensitivity	specificity
DILI concern	14	0.26	0.78	0.63
hypertrophy	47	0.32	0.81	0.65
vacuolization	39	0.38	0.77	0.58
ground glass appearance	56	0.42	0.80	0.56

References

1. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL: **How to improve R&D productivity: the pharmaceutical industry's grand challenge**. *Nature Rev. Drug. Discov.* 2010, **9**(3):203–214.
2. Lazarou J, Pomeranz BH, Corey PN: **Incidence of adverse drug reactions in hospitalized patients a meta-analysis of prospective studies**. *JAMA* 1998, **279**(15):1200–1205.
3. Uehara T, Ono A, Maruyama T, Kato I, Yamada H, Ohno Y, Urushidani T: **The Japanese toxicogenomics project: application of toxicogenomics**. *Mol. Nutr. Food. Res.* 2010, **54**(2):218–227.
4. Chen M, Vijay V, Shi Q, Liu Z, Fang H, Tong W: **FDA-approved drug labeling for the study of drug-induced liver injury.** *Drug Discov Today* 2011, **16**(15-16):697–703.
5. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185–193.
6. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, et al.: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data**. *Nucleic Acids Res.* 2005, **33**(20):e175.
7. Hochreiter S, Clevert DA, Obermayer K: **A new summarization method for Affymetrix probe level data**. *Bioinformatics* 2006, **22**(8):943–949.
8. Talloen W, Clevert DA, Hochreiter S, Amaratunga D, Bijnens L, Kass S, Göhlmann HWH: **I/NI-calls for the exclusion of non-informative genes: a highly effective feature filtering tool for microarray data**. *Bioinformatics* 2007, **23**(21):2897–2902.
9. Talloen W, Hochreiter S, Bijnens L, Kasim A, Shkedy Z, Amaratunga D, Göhlmann HWH: **Filtering data from high-throughput experiments based on measurement reliability**. *Proc. Natl. Acad. Sci. USA* 2010, **107**(46):173–174.
10. Kasim A, Lin D, Sanden SV, Clevert DA, Bijnens L, Göhlmann H, Amaratunga D, Hochreiter S, Shkedy Z, Talloen W: **Informative or noninformative calls for gene expression: a latent variable approach**. *Stat. Appl. Genet. Molec. Biol.* 2010, **9**:1–29.
11. Hochreiter S, Obermayer K: **Support Vector Machines for Dyadic Data**. *Neural Comput* 2006, **18**(6):1472–1510.