Detecting networks of gene expressions associated with human drug induced liver concern (DILI) using sparse principal components.

Ashley Bonner[1§], Joseph Beyene[1]

[1] Clinical Epidemiology and Biostatistics Dept., McMaster University, 1280 Main St. West, Hamilton, ON, L8S 4L8, Canada

§Corresponding author, bonnea@math.mcmaster.ca

Introduction

Accurately estimating a drug's potential to cause liver damage is especially important, as the liver is the organ most commonly interacting with consumed drugs. The Food and Drug Administration (FDA) developed a classification system [Chen et al, 2011] of drug-induced liver injury (DILI) potential (most, less, and no DILI concern). The drugs they applied their classification system to had been on the market for a minimum of 10 years, allowing sufficient public interaction to obtain updated and realistic DILI potential information. In contrast, new drugs will have only been tested in an experimental scene with much less data and typically with animal models. Toxicity effects from drugs might only become apparent after prolonged or human exposure, but it is not ethical to subject the public to unknown risks of this nature. Therefore, toxicogenomics, the study of drug-induced toxicity through biomarkers, is now a popular solution and the hunt is on for expression levels that predict high DILI potential.

The Japanese Toxicogenomics Project (TGP) has such motivations [Uehara et al. 2010]. With human in vitro, rat in vitro, and rat in vivo experiment models, they tested 131 drugs, many part of the FDA classification system, on liver samples for gene expression levels on thousands of probsets, using Affymetrix GeneChip® technology. This year's International Conference on Critical Assessment of Massive Data Analysis (CAMDA 2013) utilizes the TGP data to propose analysis challenges involving prediction of toxicity levels of drugs. Discovering novel biomarkers associated with DILI potential could aid in safely classifying the toxicity of new drugs. However, the breadth of genomic data makes analysis with simple statistical models challenging and dimension reduction techniques could be essential in some data scenarios.

Principal Component Analysis (PCA) is a multivariate dimension reduction and visualization technique that produces a new set of variables called principal components (PCs), constructed as linear combinations of the original variables, that efficiently organize the information in the original dataset and prepares for a computationally simpler analysis. The downfall to PCA is that PCs are comprised of *all* original variables, which is: a) unrealistic, if PCA is used for identifying group structure in the data, and b) confusing, since the interpretation of PCs near impossible. Sparse Principal Component Analysis (Sparse PCA) is a new extension to classical PCA that systematically forces variables with residual contribution to have 0-valued loadings, therefore attaining a) a more concise and realistic group structure of the data, and b) a more interpretable set of PCs for further analysis; absolutely critical for large genomic data.

In our contribution to the International Conference of CAMDA 2013, we utilize Sparse PCA to assemble organized gene expression profile variables (sparse PCs) for subsets of the human in vitro TGP data. We then use these sparse PCs to determine groups of gene expressions jointly associated with human DILI concern. By targeting linear combinations of gene expressions rather than individual probsets, we hope to uncover potentially interesting avenues for biological interpretation.

Methods

Data:

The TGP administered control, low, middle, and high doses of 119 to 131 drugs to human in vitro hepatocytes, rat in vitro hepatocytes, and rat in vivo. Gene expressions from the samples were measured at several time points after the drugs had been given.

Targeted Samples: We consider only samples from the human in vitro experiments. Of the 119 drugs applied to human samples, we consider only the 93 drugs that have human DILI concern classification as provided by Chen et al in 2007. Anticipating that higher doses result in more robust gene expression measurements (McMillian et al, 2005) and due to many drugs not being administered at low doses in the human in vitro samples, we consider only middle and high dose levels across all samples. Likewise, since measurements of gene expression in the human samples were not taken at 2hrs for many drugs, we restrict our analysis to only gene expressions measured at 8 and 24 hours. Of the two human samples found within each combination of drug, dose level, and gene expression sampling time, we used only the first, assuming the duplicates to be technical replicates.

Targeted Variables: Starting with the 54675 probsets retrieved from the Affymetrix GeneChip® Human Genome U133 Plus 2.0 Array with MAS5 summarization, we filtered out the bottom 75 percent in terms of Inter-Quartile Range (IQR), retaining only the 13669 most variable probsets. We transformed all gene expression values using log base 2 to dull the presence of outliers and achieve distributions closer to normal. Additionally, we measured gene expression changes between dose levels (High – Control, Middle – Control, High – Middle) at different sampling times and gene expression changes between sampling times (24 hours – 8 hours) at different doses; outcomes that better represent effects of dose levels and capture gene expression over time [Sukumaran et al, 2010]. Human DILI concern ('most', 'less', and 'no DILI concern' in humans) was used to detect if gene expression measurements differed across DILI classification. However, since only 8 drugs are classified as 'no DILI concern', we reclassified the human DILI concern variable to be binary: 'most DILI concern' vs. 'less or no DILI concern'. Of the 93 drugs we considered for humans, 40 are 'most DILI concern' and 53 are 'less or no DILI concern'; relatively balanced.

Analysis:

Investigate marginal associations: For each subset of data and gene expression outcome variable, we statistically tested marginal associations between each probset and human DILI concern by using moderated t statistics, tracking and counting those probsets with p-values < 0.05 and, more appropriately due to running many tests, those probsets with p-values < 0.05 after adjusting for false-discovery rate (FDR). Moderated t statistics, p-values, and FDR-adjusted p-values were calculated using the LIMMA package in R v3.0.0. We plan to examine if these probsets are found and grouped in our sparse PCA analysis.

Sparse PCA for finding joint associations: For each subset of data and gene expression outcome variable, we build 93 sparse PCs to summarize the gene expression data by using the sparse PCA method proposed by Witten, Tibshirani, and Hastie in 2009; the 'SPC' function in the authors R-package 'PMA'. Within such a high-dimensional data environment, their sparse PCA method is suggested as the best choice among competitors (Bonner and Beyene, 2012). We used several tuning parameters, 3, 5, 10, 20, 30, 40, to force different levels of sparseness to the PCs, looking for a balance between sparseness and percentage total variance retained among PCs. We statistically tested associations between sparse PCs and human DILI concern by using student t statistics, tracking and counting those sparse PCs with p-values < 0.05 and, more

appropriately, those sparse PCs with p-values < 0.05/93; a simple bonferroni adjustment for multiple testing, since sparse PCs are relatively independent. Within sparse PCs that were statistically significant, we identified the largest loadings and examined genetic structure for corresponding probsets We plan to report which genetic regions were most represented.

Results

Table 1 displays counts of probsets that were marginally associated with human DILI concern, for each combination of subgroup and gene expression outcome. We include results from a control dose subgroup analyses to highlight, through comparison, how many false-positives we expect to find in other analyses; control dose (0) across all drugs should generate the same gene expression levels, regardless of drug class, providing a baseline number of false-positives. Using FDR-adjusted p-values, we found only 3 probsets differentially expressed between most and less-or-no DILI concern: '1563061_at' for single value expression measurements at high doses and 8 hour sampling time, and '1567060_at' and '1557437_at' for single value expression measurements at high doses and 24 hour sampling time.

Moving to joint associations, we chose to examine only those sparse PCs obtained from using a tuning parameter of 30, since the immense sparseness induced by smaller tuning parameters reduced the percentage explained variance of the PCs too much. The sparse PCs we investigate had an average of 1727.11 non-zero loadings, down from an expected 13669 that classical PCA would produce. The total percentage of probset variance explained by all PCs ranged from 40.3% to 68.1%, depending on the subgroup and gene expression outcome. This is a substantial amount considering almost 90% of the loadings were forced to 0, validating the ability of Sparse PCA as a dimension reduction technique. As shown in Table 1, we found only 2 sparse PCs to be differentially expressed between most and less-or-no DILI concern after adjusting for multiple testing; Figure 1 presents their loading plots. Although there does not seem to be any trend in the probsets when ordering them as found in the database, we can investigate the top contributing probsets as they might provide insightful biological meaning regarding the PC. Loadings above the blue lines in Figure 1 correspond to probsets: 1557636_a_at, 215586_at, 243325_at, 1568751_at, and 1560349_at.

Discussion

Overall, analyzing human in vitro samples did not allow us to find any blaringly obvious gene expressions. Perhaps rat samples would boast more gene expression associations. It seems that high dose levels are slightly more able to detect probsets differentially expressed between most and less-or-no DILI concern in humans. Though, finding just three individual probsets differentially expressed after adjusting for multiple testing is a convincing argument towards investigating more complex relationships between human DILI concern and gene expressions. Coupling this motivation with high dimensional data issues, applying Sparse PCA seems to be an appropriate solution as not only does it automatically construct sparse linear combinations of the probsets, thus highlighting underlying structure among gene expression, but it also dramatically reduces the number of variables we need to analyze.

With just two sparse PCs considered differentially expressed after adjusting for multiple testing, it leads us to believe that just a few networks of human gene expressions are associated with DILI concern. However, since the sparse PCs host a collection of probsets, there exists more room for exploration, providing a more interesting avenue for biological investigation.

There were limitations with our analysis approach that we are looking forward to addressing. Due to each of our analyses having just 93 samples (1 per drug), using human DILI concern as a strict classification may have been presumptuous of drug homogeneity. Drugs are quite heterogeneous in their relations to gene expression (Afshari et al, 2011), so even if two drugs of most DILI concern were influential to a marker of gene expression, perhaps one up-regulates while the other down-regulates, leaving the resulting behavior deemed non-influential. Anticipating this, we had also investigated absolute changes in gene expression across doses and sampling times, but the findings were similar to those already presented. That said, perhaps we limited our results interpretation when restricting our view to only the probsets significant after adjusting for multiple testing. Figure 2, for example, shows that clustering samples by the top 100 differentially expressed probsets in the high dose, 8 hour subgroup is able to group DILI classes rather well. Sparse PCA can be regarded as a more statistically formal clustering method, so we have high hopes for extracting groups of gene expressions with our methods. Finally, sparse PCA is unsupervised, such that it does not build sparse PCs with the factor of interest, human DILI concern, in mind. Perhaps a supervised approach to selecting gene expressions such as Sparse Partial Least Squares (SPLS) would be more effective.

Future Directions

This is work in progress. In time for the CAMDA 2013 conference, we plan to integrate gene information to gain biological context and we will be including rat in vitro and rat in vivo samples to detect how sensitive gene expressions from rat samples are compared to human samples. Human DILI-associated gene structures obtained via sparse PCA on the rat samples can be compared to those found in humans, mapping common gene functions (Uehara et al., 2008). As well, the FARMS summarized data will be used.

References

- Afshari, C. A., Hamadeh, H. K., Bushel, P. R., The Evolution of Bioinformatics in Toxicology: Advancing Toxicogenomics, Toxicological Sciences (2011), doi:10.1093/toxsci/kfq373
- Bonner, A., Beyene, B., Sparse Principal Component Analysis for High-Dimensional Data: A Comparative Study, Open Access Dissertations and Theses - McMaster (2012), Paper 7146.
- Chen, M., et al., FDA-approved drug labeling for the study of drug-induced liver injury, Drug Discovery Today (2011), doi:10.1016/j.drudis.2011.05.007
- McMillian, M., et al, Drug-induced oxidative stress in rat liver from a toxicogenomics perspective, Toxicology and Applied Pharmacology (2005), doi: 10.1016/j.taap.2005.02.031
- Uehara, T., et al., Species-specific differences in coumarin-induced hepatotoxicity as an example toxicogenomics-based approach to assessing risk of toxicity to humans, Human & Experimental Toxicology (2008), doi: 10.1177/0960327107087910
- Uehara, T., et al., The Japanese toxicogenomics project: Application of toxicogenomics, Mol. Nutr. Food Res. (2010), DOI 10.1002/mnfr.200900169
- Sukumaran, S., Almod, R., DuBois, D., Jusko, W, Circadian rhythms in gene expression: Relationship to physiology, disease, drug disposition and drug action, Advanced Drug Delivery Reviews (2010), doi:10.1016/j.addr.2010.05.009
- Witten, D., Tibshirani, R., Hastie, T., A penalized matrix decomposition, with application to sparse principal components and canonical correlation analysis, Biostatistics (2009), 10:515-534.

Table 1: Counts of probsets that are differentially expressed between samples receiving drugs of most and less-or-no DILI concern; n = 93 samples (1 per drug) and p = 13669 gene expression measurements within each subgroup.

Subgroup	Gene Expression Measurement	# DEGs, p<0.05 (# passed FDR)	# DEsPCs, p<0.05 (# passed Bonferroni)
High dose, 8 hours	Single Value	827 (1)	6 (0)
Middle dose, 8 hours	Single Value	707 (0)	4 (0)
Control dose, 8 hours	Single Value	846 (0)	6 (0)
High dose, 24 hours	Single Value	700 (2)	2 (1)
Middle dose, 24 hours	Single Value	641 (0)	4(0)
Control dose, 24 hours	Single Value	680 (0)	3 (0)
8 hours	Change (H – C dose)	645 (0)	4 (0)
8 hours	Change (M – C dose)	717 (0)	6 (0)
8 hours	Change (H – M dose)	679 (0)	5 (0)
24 hours	Change (H – C dose)	676 (0)	4 (0)
24 hours	Change (M – C dose)	673 (0)	6 (0)
24 hours	Change (H – M dose)	702 (0)	4 (1)
High dose	Change (24 – 8 hours)	715 (0)	5 (0)
Middle dose	Change (24 – 8 hours)	690 (0)	3 (0)
Control dose	Change (24 – 8 hours)	700 (0)	7 (0)

Figure 1: Loading plots for the top significant sparse PCs. Probsets corresponding to the top loadings might be of significant interest.

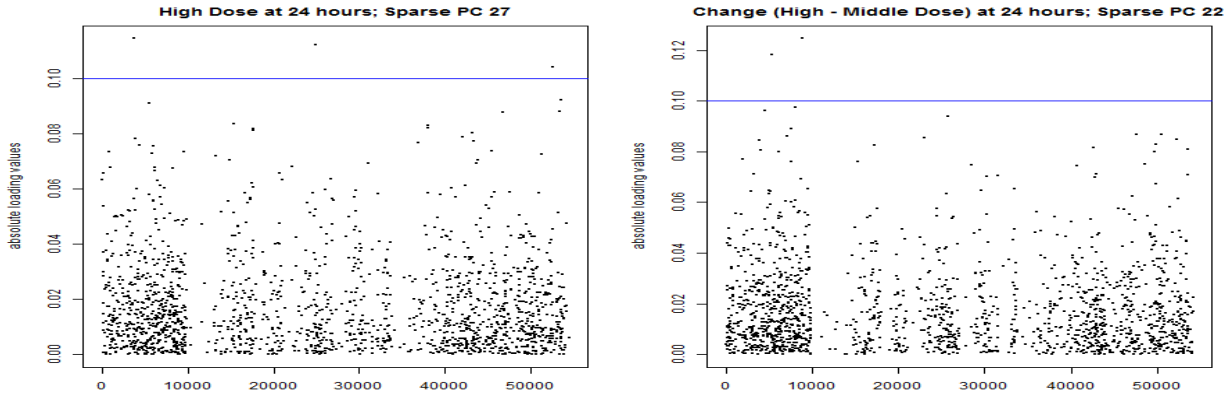

Figure 2: Heatmap showing clustering of most (red) vs. less-or-no (blue) DILI concern.