CAMDA2013

12th Annual International Conference on
Critical Assessment of Massive Data Analysis

Berlin, Germany | July 19-20, 2013

Conference Program and Abstracts

Organizers and chairs

Joaquin Dopazo
Bioinformatics Department
Centro de Investigación Principe Felipe
Valencia, Spain

Sepp Hochreiter
Institute of Bioinformatics
Johannes Kepler University Linz
Linz, Austria

Djork-Arné Clevert
Institute of Bioinformatics
Johannes Kepler University Linz
Linz, Austria

David Philip Kreil
Bioinformatics
Boku University Vienna
Austria

Simon Lin
Biomedical Informatics Research Center
Marshfield Clinic Research Foundation
Marshfield, WI USA

Scientific committee

Tim Beißbarth
Statistical Bioinformatics
University Medical Center Göttingen
Göttingen, Germany

Daniel Berrar
Tokyo Institute of Technology
Japan

Jong Bhak
Genome Research Foundation
South Korea

Mike Bowles
Biomatica, Founder
San Jose, CA, USA

Philippe Broët
University of Paris - XI
Paris, France

Susmita Datta
Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, KY USA

Joaquin Dopazo
Bioinformatics Department
Centro de Investigación Principe Felipe
Valencia, Spain

Wolfgang Huber
European Molecular Biology Laboratory
Heidelberg, Germany

David Philip Kreil
Bioinformatics
Boku University Vienna
Austria

Simon Lin
Biomedical Informatics Research Center
Marshfield Clinic Research Foundation
Marshfield, WI USA

Yves Moreau
K.U. Leuven
ESAT-SCD (Bioinformatics)
Leuven, Belgium

Ruchir Shah
Bioinformatics
SRA International, Inc.
Durham, NC USA

Ziv Shkedy
Center of Biostatistics and statistical
Bioinformatics
University Hasselt
Hasselt/Leuven, Belgium

Min He
Center for Human Genome Variation
Duke University School of Medicine
Durham, NC USA

Sepp Hochreiter
Institute of Bioinformatics
Johannes Kepler University Linz
Linz, Austria

Lan Hu
Center for Cancer Computational Biology
Dana-Farber Cancer Institute
Boston, MA, USA

Kun Huang
Comprehensive Cancer Center
Biomedical Informatics Shared Resource
The Ohio State University Medical Center
Columbus, OH USA

Willem Talloen
Functional Genomics Department
Johnson & Johnson, Pharmaceutical R&D
Beerse, Belgium

Weida Tong
Center for Bioinformatics, NCTR/FDA
Rockville, MD USA

Armand Valsesia
Genetics Group
Nestlé Institute of Health Sciences
Lausanne, Switzerland

Julie Zhu
University of Massachusetts Medical School
Worcester, MA, USA

Chunlei Wu
Department of Molecular & Experimental Medicine
The Scripps Research Institute
La Jolla, CA USA

Sponsors

CAMDA appreciates the generous support of the following sponsors:

Program Schedule for
Friday – July 19, 2013

07:30 - 09:00	Registration
09:00 - 09:15	Welcome
09:15 - 10:15	**Keynote: Atul Butte.** *Translating a trillion points of data into therapies, diagnostics, and new insights into disease*
10:15 - 10:45	Morning break
10:45 - 11:10	**Djork-Arné Clevert.** Setting the context – contest data set I
11:10 - 11:50	**Martin Otava.** *Prediction of Gene Expression in Human Using Rat in Vivo Gene Expression in Japanese Toxicogenomics Project*
11:50 - 12:30	**Hector A. Rueda-Zarate.** *A gene expression landscape of drug-induced liver hepatotoxicity*
12:30 - 13:30	Lunch break
13:30 - 14:10	**Ashley Bonner.** *Detecting networks of gene expressions associated with human drug induced liver (DILI) concern using sparse principal components*
14:10 - 14:50	**Tommi Suvitaival.** *Cross-organism prediction of drug hepatotoxicity by sparse group factor analysis*
14:50 - 15:30	**Marinka Zitnik.** *Matrix Factorization-Based Data Fusion for Drug-Induced Liver Injury Prediction*
15:30 - 16:00	Afternoon break
16:00 – 16:40	**Patricia Sebastián-Leon.** *Using probabilistic models of signaling pathways to predict in vivo drug activity*
16:40 - 17:00	**Ryan Gill.** *Similarity in Network Structures for in vivo and in vitro Data from the Japanese Toxicogenomics Project*
17:00 - 17:20	**Jose Luiz Rybarczyk-Filho.** *TRANSTAGING: Transcriptogram-based staging of cancer*
19:00	Conference dinner

Program Schedule for
Saturday – July 20, 2013

09:00 - 10:00	**Keynote: Nikolaus Rajewsky.** *Circular RNAs and other surprises from analysis of RNA:protein interactions*
10:00 - 10:30	Morning break
10:30 - 11:10	**Wei Zhang.** *Assessing Genomic Biomarkers of Toxicity in Drug Development*
11:10 - 11:50	**Jari Björne.** *Analyzing the Japanese Toxicogenomics Project Dataset with SVM and RLS Classifiers*
11:50 - 12:30	**Zhilong Jia.** *Reasonably integrating data for predicting drug toxicity by machine learning*
12:30 - 13:40	Lunch break
13:40 - 14:20	**Danyel Jennen.** *DILI classification model based on in vitro human transcriptomics and in vivo rat clinical chemistry data*
14:20 - 14:50	**Sepp Hochreiter.** Setting the context – contest data set II
14:50 - 15:30	**Huixiao Hong.** *Assessing single-nucleotide polymorphism and genotype calling using the KPGP-38 Human Genomes next-generation sequencing data from CAMDA*
15:30 - 16:00	Afternoon break
16:00 - 16:20	**Mathew Palakal.** *Characterization of the Korean Genome*
16:20 - 16:40	**Abhishek Kumar.** *Application of next-generation genome and transcriptome based methods for the exploration of secondary metabolites from marine fungi for the treatment of cancer*
16:40 - 17:10	Short break: vote for best presentations
17:10 - 17:30	Closing address and CAMDA contest awards

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.

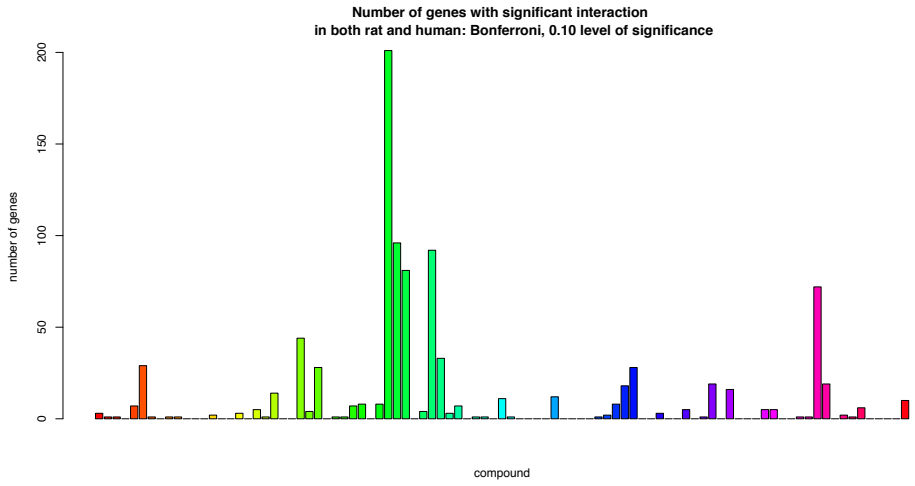


Figure 1: *Number of genes with significant interaction in two-way ANOVA for both rat and human. The p-values are adjusted using Bonferroni's method on significance level 0.10.*

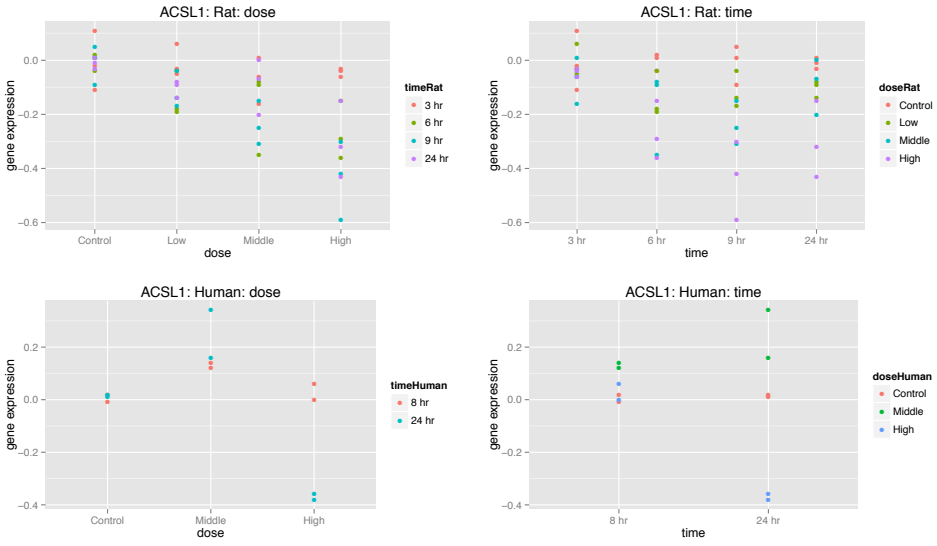


Figure 2: *Example of gene with significant interaction in both human and rat. Compound omeprazole and gene Acsl1 in rat, respectively ACSL1 in human.*

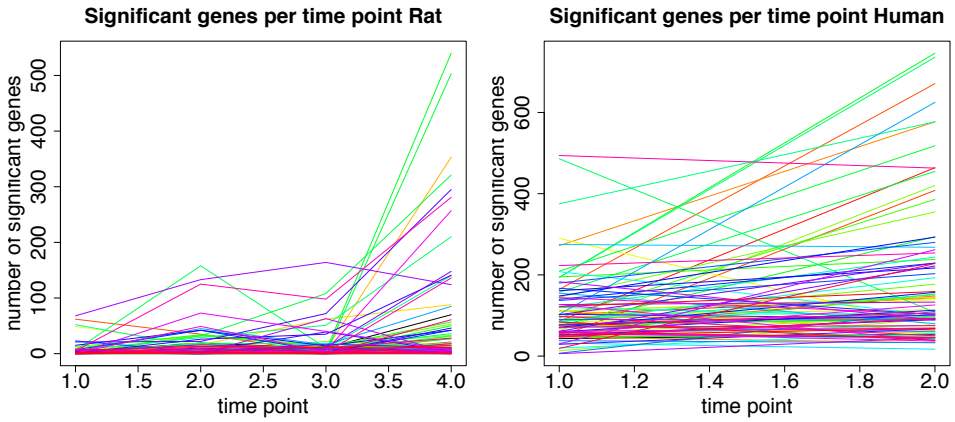
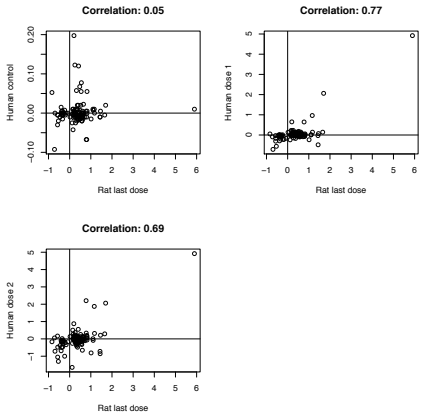
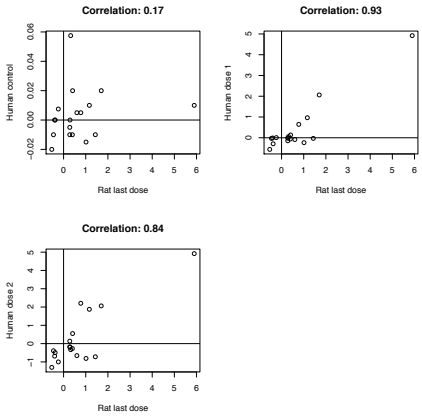


Figure 3: *Number of genes with significant dose-response profile per time point. Test is based on MCT and p-values are adjusted using Bonferroni's method on significance level 0.10. Rat data results are on left panel, human data on right panel.*



(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.*

A gene expression landscape of drug-induced liver hepatotoxicity

Héctor Rueda-Zárate[1,2], *Iván Imaz-Rosshandler* [1], *Julieta Noguez-Monroy*[2] *and Claudia Rangel-Escareño* [1*]

1 Computational Genomics, National Institute of Genomic Medicine Mexico

2 Computational Sciences Graduate Program, Instituto Tecnológico y de Estudios Superiores de Monterrey C.C.M.

*corresponding author crangel@inmegen.gob.mx

Introduction

The liver is one of the organs involved in biotransformation, chemical reactions that alter the structure, aqueous solubility and eventual disposition of non- nutritive compounds that enter into the organism. Xenobiotic biotransformation aims at controlling the toxication or detoxication of xenobiotic substances. However, during the biotransformation reactive intermediates may be produced, these could interact with critical cellular macromolecules and trigger the events that promote either tissue injury and cell death, permanent genomic changes, leading potentially to cancer.

Many currently and normally used drugs could affect the liver adversely in any combination of the reactions described. Liver injury can be classified as hepatocellular, cholestatic or mixed, based on criteria established by the Council for International Organizations of Medical Sciences (CIOMS) [14]. The drug-induced liver injury also known as DILI is classified as intrinsic and idiosyncratic hepatoxicity. The Intrinsic hepatotoxins cause hepatocellular damage and it is more related to other industrial agents more than it is to xenobiotics. However, xenobiotics are more closely related to idiosyncratic liver injury by its level of toxicity or to allergy reaction or other secondary effects. Toxic effects of drugs at all levels are extensively studied before these are administered to humans. The Toxicogenomics Project focuses on gene expression analyses in animals or in-vitro grown cells that have been exposed to the chemicals with the aim of understanding the molecular mechanisms of toxicity and eventualy be able to predict dangerous levels of toxicity.

Materials and Methods

We used gene expression data from the Japanese toxicogenomics project (TGP), a 5-year project that was completed in 2007. TGPs database comprises nearly 18,000 Affymetrix microarrays testing 131 compounds, mainly medical drugs and their effect in the liver. All microarrays targeted the liver in both in vitro and in vivo experiments. All .CEL files were downloaded into a 32GB server for the analyses. A primary test on processing capabilities and algorithm complexities showed that up to a maximum of 400 microarrays could be pre-processed using R and biocoductor `affy` library and its dependencies at once on this server. Hence, the strategy for this analysis would have to be design in such a way that it loads only those sets of microarrays involved in the actual biological question.

Strategy for integrative analysis

A map of how data are structured can be seen in Figure[1]. We need a strategy that will allow us to combine species(Hu, Rat), protocols (iVV, iVT), dosages (None, Low, Med, High) and time points. A mixture of differential expression analysis using `limma` and gene selection using ranking approach such as `timecourse` seems to be an appropriate beginning approach.

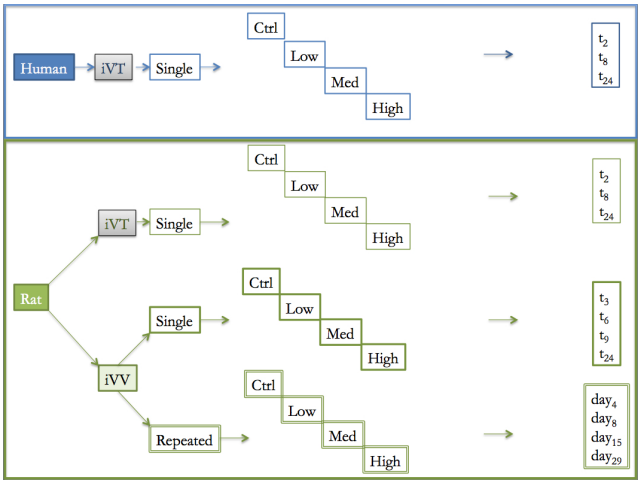


Figure 1. Group structure in TGP data

Just the human diagram (top portion in Figure [1]) would lead to up to 30 pairwise contrasts of interest per compound, roughly 117,900 comparisons if we were to use `limma` alone.

$$T_{contrasts} = N_{comp}\left[N_{timepts}\binom{N_{cond}}{2} + N_{cond}\binom{N_{timepts}}{2}\right]$$

That plus a similar number for RatiVT, also for RatiVV, plus repeated plus all cross-referenced contrasts. This, however, does not mean we cannot still do it. This approach should be taken on a more biological driven hypotheses rather than as massive computational analysis.

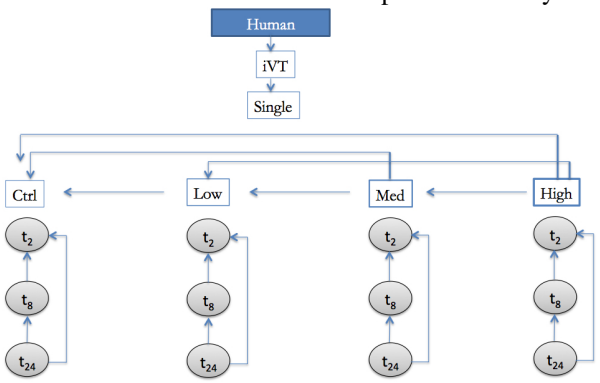


Figure 2. Diagram of contrasts within Human in-vitro samples per compound

Methods which rank genes (e.g. the MB statistic or the moderated Hotelling T2) perhaps provide easier access to genes whose absolute or relative expression varies over time. This approach is used for each of the four main paths in Figure [1] {HU.iVT, Rat.iVT.Single, Rat.iVV.Single,Rat.iVV.Repeated}. However, in order to see correlation of using animal model to infer potential toxicity in humans we would also need to take the differential expression approach. This will be applied to all groups with identical structure but different species. In our scenario that would only be {HU.iVT.Single, Rat.iVT.Single}.

The general structure of our strategy involves :
• Data storage and manipulation by using a relational database
• Raw-data analysis: quality metrics, normalization and background correction.
• Gene selection either by differential expression and/or by ranking method.
• Gene annotation and function. Conducting a gene set enrichment analysis on lists of ranked genes; a selection of orthologous genes using swissprotID and inParanoid database [11]. Simultaneously, research on drug toxicology to determine if compounds could be classified according to toxicity or type of liver damage.
• Machine learning approach: searching for possible patterns in data, clusters of compounds by unsupervised methods. Patterns in concentrations, patterns in the time-course results. • Tool development that will be available through R and Bioconductor.

Data

The TGP data contains a collection of 17,657 Affymetrix^{TM} microarrays from both in vitro and animal samples. Human samples were processed using Hgu133Plus2, animal samples were processed on the GeneChip Rat Genome 230 2.0 which is known to be a powerful tool for toxicology.

MySQL database

Due to data complexity in terms of number of groups, labels for all barcoded data were stored as a relational MySQL database. This allows faster, easy and optimum access to a specific set of .CEL files for further analysis. Even though it was one table at first, the database will grow as more information is developed. It will be constantly normalized and designed to be scalable. Access to it is through R scripts, an example is shown here:

```
findMicroarrays(species=c("Rat"),
      expType=c("in vitro", "in vivo"),
      dose=c("Control", "Middle"),
      singleRepeat=c("NA", "Single", "Repeat"),
      compound=c("AA", "ACA", "WY"),
      sacTime=c("2 hr", "8 hr"),
      experiment=c("CAMDA13"),
      path="CELS/",
      orderBy="DOSE_LEVEL", conn=conexion)
```

This function collects files according to the specified parameters and it modifies the file names to match the conditions making all more easy to follow. The resulting query is shown below.

```
SELECT BARCODE FROM MICROARRAY WHERE...
    SPECIES IN ('Rat') AND TEST_TYPE IN ...
    ('iVT','iVV') AND DOSE_LEVEL IN ...
    ('Control','Low') AND SINGLE_REPEAT_TYPE...
    IN ('NA','Single','Repeated') AND ...
    COMPOUND_ABBREVIATION IN ...
    ('AA','ACA','WY') AND SACRIFICE_PERIOD...
    IN ('2 hr','8 hr') AND EXPERIMENT IN...
    ('CAMDA13') ORDER BY DOSE_LEVEL;
```

And the resulting sample names are shown below.

```
 [1] "Rat.iVT.Control.NA.2 hr.WY-1"...
 [3] "Rat.iVT.Control.NA.8 hr.AA-3"...
[13] "Rat.iVT.Low.NA.2 hr.AA-13"...
[15] "Rat.iVT.Low.NA.8 hr.AA-15"...
[21] "Rat.iVT.Low.NA.8 hr.WY-21"...
```

Low-level analysis

Gene expression microarray raw data for subsets of samples collected through the database were pre-processed in the R statistical environment. A quality control tests

were run on randomly selected sets, showed a constant behavior of $MM-probes > PM-probes$ in a range from 22-30% causing serious concerns about using MAS5.0 algorithm for background correction. In fact, this analysis showed also that RMA [9] led to bimodal distribution indicating that background adjustment was unnecessary. Data were normalized using quantile normalization [4] , summarization was done using medianpolish. All methods from the Bioconductor `affy` library.

Timecourse analysis

Genes were ranked based on large absolute or relative amounts of change over time as a function of the drug concentration in relation to their replicate variances. For every selected subset, genes were classified according to a multivariate empirical Bayes statistic for replicated microarray time course data MB statistics implemented in the `timecourse` package [21].

Human-Rat orthologues

The human, mouse and rat genomes encode a very similar number of genes. Human-Rat share roughly 89 to 90% of genes [8] with a majority that have persisted without deletion or duplication since the last common ancestor. The most important aspect is perhaps that almost all human genes known to be associated with disease have orthologues in the rat genome. However, their rates of synonymous substitution are significantly different from the remaining genes. Hence, even though the high correlation we are also conducting an orthology analysis through related proteins using the InParanoid database [11]. This databases information is based on information about swissprotID. We are also exploring the ENSEMBLE database for this purpose. More tables are added to our database so gene-to-gene information could be generated.

Gene set enrichment analysis

After using timecourse approach, lists of genes of interest are generated. We may end up with way too many genes to examine in proper detail. Hence, a good way of comparing conditions is thorugh a gene set enrichment analysis that could tell us about cellular mechanisms behind different lists. The idea is to identify pathways affected by highly ranked genes in Human iVT and compare to those found in Rat iVT and Rat iVV. Tools used for this approach involve DAVID [5], GSEA [6] and BiNGO [3]

Machine Learning Approach

Even though we have access to a quite impressive sample size, this number is fastly diluted by the number of groups in the study. If we see Figure 1, we have four main groups {HuiVT, RatiVT, Rat iVV, RatiVV-Rep}, between 119 and 131 xenobiotics, and between 3 and 4 time points. So we basically have only either two or three replicates in each group for statistical assessment. The question we would like to address here is: What can we learn from data?
Hence, an unsupervised hierarchical clustering would allow to see sets of genes that follow a similar profile between the main groups. Properly validated these sets of genes could potentially be used as markers for in-vitro human models avoiding the need of performing animal model approaches. Class discovery and clustering validation can also be tested using Consensus Clustering method [18]

Results

Database implementation and data retrieval through R made all of the timecourse analyses time efficient. Only 48 compounds were selected since these are found in all four groups. Results from timecourse ranked all genes and only the top 50 from each group and each compund (2400 approx.) were selected for further analysis. Below there is a list of the 25 most common genes where column labeled as Count represents the number of times that gene was present across drugs, time and concentration in three groups: {HU iVT, Rat iVT, Rat iVV}.

Table 1. Top 50 genes in each compound and number of times they are present

Hu iVT	Count	Rat iVT	Count	Rat iVV	Count
CYP3A5	49	CXCL3	45	TXNRD1	23
CYP3A4	34	CYP1A1	32	ACOT2	17
NEAT1	31	SLC7A11	19	HSDL2	14
RRM2	29	SOX4	19	CCND1	13
RRAD	27	PDK4	17	SREBF1	13
CYP2C9	23	ANGPTL4	16	DUSP6	12
CYP1A1	20	HSDL2	16	SRXN1	12
MALAT1	20	ACAT3	14	PTPRF	11
ATF3	19	HMGCS2	13	STAC3	11
TSKU	18	NREP	13	ANKH	10
ANGPTL4	17	CD36	12	ATP1B1	10
ARL14	17	CPT1A	12	HAMP	10
CYP1A2	17	SERPINB9	12	HSPB8	10
PCK1	17	ACOT2	11	PPCS	10
C19orf80	16	DHRS3	11	SLC13A4	10
CYP3A7	14	TAGLN	11	TBC1D15	10
EGR1	14	FASN	10	TM2D3	10
GDF15	14	HSP90AB1	10	CAR14	9
PPP1R15A	14	LSS	10	GCLC	9
TRIB3	14	AKR1D1	9	PKLR	9
FAM13A	13	CYP26B1	9	CXCL12	8
IFRD1	13	FABP7	9	MGLL	8
MAFF	13	GDE1	9	PDK4	8
RPL38	13	PEX11A	9	ACACA	7

The level of correlation or intersection between these genes and assuming same symbol indicates orthologues

is shown in the Venn diagram in Figure 3.

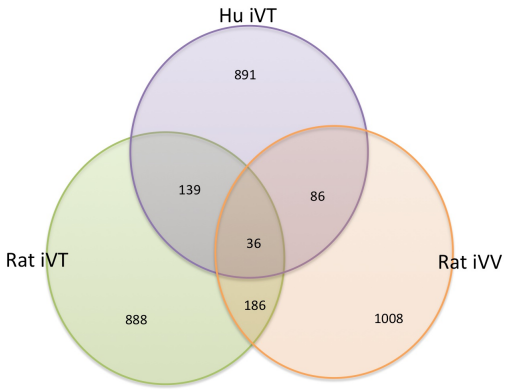


Figure 3. Genes found in all three groups and some of the 48 compounds were 36: {*PEX11A, PDK4, ANGPTL4, SERPINB9, CYP1A1, LSS, FASN, TRIB3, CREM, SWI5, PPP1R3B, PIR, NREP, HMGCR, ABCD3, RDX, TIPARP, SQLE, NQO1, HSPH1, YPEL5, EGR1, PTPRF, MDM2, JUN, BHLHE40, LDLR, TSKU, IFRD1, GCLM, SGK1, RRM2, EFNA1, IRF7, BCL6, INHBE* }

Since there is a vast amount of information for an abstract, we concentrated on one drug and pursued a more detailed analysis.

Case study intrinsic DILI: Acetaminophen

Acetaminophen toxicity is the leading drug-related cause. At low doses, the drug is conjugated to water-soluble metabolites in the liver and is excreted in the urine. At higher doses, glutathione depletion leads to saturation of the conjugation mechanism, leaving the parent compound to be metabolised to toxic intermediates. Moreover, toxicity risk increases if there exists chronic alcohol consumption, obesity, or drugs that induce the P-450 cytochrome system lowering the toxic threshold of acetaminophen [16], [17]. Timecourse results for this compound are shown in Figure 4.

We observe that patterns are different on each group even though is the same compound. We should consider however, that ranking is determine by genes with changes across time as a function of concentration also including that replicates do not vary much. A collective view of this including the top 100 genes can be seen in Figure 5 where we observe that overall gene profiles are different. This simply suggests that performing a timecourse analysis do not exclude the usual between groups differential expression approach. We would like to point out that this approach can also be done and has been done through the database.

One interesting remark is about the effect of high concentration effect at time 24 hrs. It is not clear whether gene is down-regulated as a response to high concentration or if we are facing a cell viability issue and the cell simply dies. We performed a gene set enrichment analysis and found out apoptosis pathways are significant for some drugs.

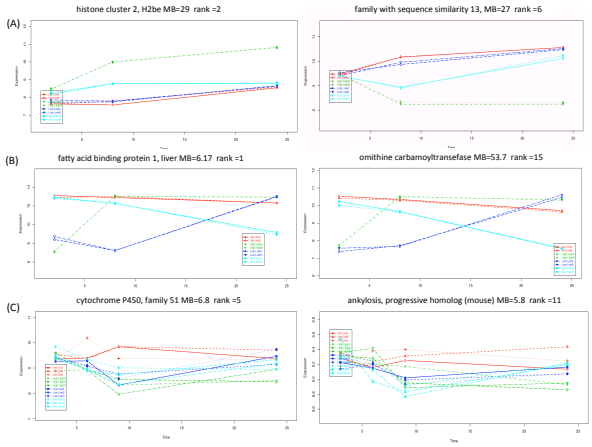


Figure 4. Top ranked genes in Acetaminophen (APAP): Top row (A) Human iVT, middle row (B) Rat iVT, bottom Row(C) Rat iVV Red = control, blue = Low, Cyan = Med and Green = High

Hierarchical clustering does not show interesting patterns in terms of gene profiles. However, among the three groups the Human iVT plots shows more interpretable results. As we can see that for most genes patterns of up-reglation occur at time 24 hrs.

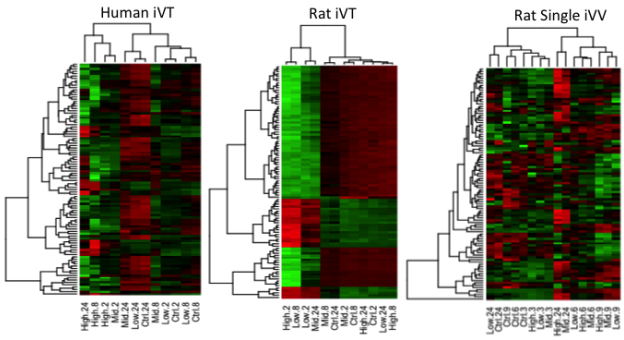


Figure 5. Hierarchical cluster plots for three groups.

Consensus clustering [18] on the selected 48 compounds are shown in Figure 6. Here we observe two interesting

patterns. It seems that after 2 hours most compounds tend to cluster into smaller number of groups (left panel top and bottom), but after 24 hours patterns are more heterogenous. This suggests that after 24 hours what we see is a putative different drug dependent metabolic process and biotransformation. Our hypothesis is that there may exist a molecular sub-classification of drugs based on gene expression profiles. We will further explore this by combining the Drug versus Disease data R- package and two databases DrugBank and ChemmineR

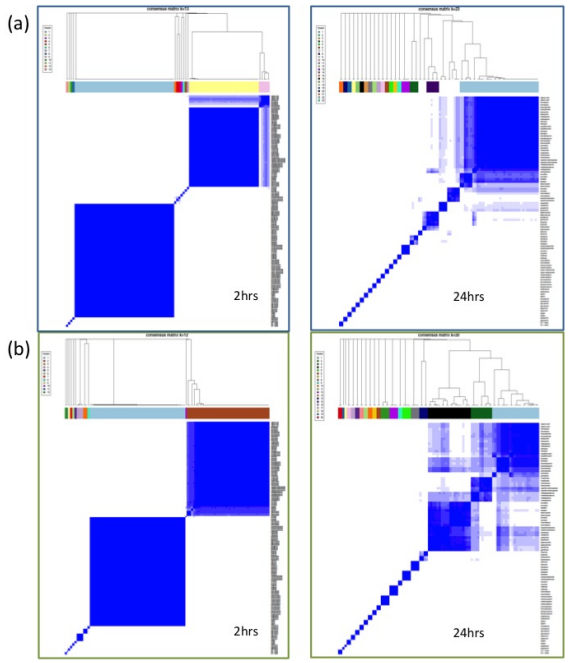


Figure 6. Consensus clustering for High concentration (a) Human iVT at 2 and 24hours.; (b) Rat iVT at 2 and 24 hours

Discussion

We have performed a broad analysis of this data set that has led us to pursue various hypotheses. Some of them are presented here and many others are currently under revision. It seems there is plenty of room for more discoveries and at this point we can only see the potential but not the end of the road. For instance, there is still work in progress for Rat in-vivo with repeated samples, a more specific gene set enrichment analysis, an extensive exploration of mechanisms for drug classification and feature selection using machine learning approaches among others.

Conclusions

The Japanese Toxicogenomics Project (TGPJ) is a combined efford between the National Institute of Health Sciences and 17 pharmaceutical companies. The purpose of the study and its results will impact drug development and toxicology research wordwide. A database fed by new gene information was created. In this work we propose an interactive model for analyses that uses a database that can be queried with specific biological questions. Then a collection of R functions will perform low-level analysis; classification providing a set of genes of interest either by timecourse, concentration or contrast specific approach; and data mining. We are currently working on an R package, as well as a manual for the scripts.

References

[1] 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. 54(2):218-27, 2010.

[2] Chen, M., et al., FDA-approved drug labeling for the study of drug-induced liver injury (DILI). Drug Discov Today, 2011. 16(15-16): p. 697-703.

[3] BiNGO: A Biological Network Gene Ontology tool. http://www.psb.ugent.be/cbd/papers/BiNGO/Home.html

[4] Bolstad, B.M., Irizarry R. A., Astrand M., and Speed, T.P. **(2003)**. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics* 19(2):185-193.

[5] DAVID: Bioinformatics Database. *Nature Protocols* 2009; 4(1):44 and *Nucleic Acids Res.* 2009;37(1):1 http://david.abcc.ncifcrf.gov/

[6] Tamayo, et al. (2005, PNAS 102, 15545-15550) and Mootha, Lindgren, et al. *Nat Genet* 2004 34, 267-273. http://www.broadinstitute.org/gsea/index.jsp

[7] Gentleman R., Carey V., HUber W., Irizarry R., Dudoit S. Bioinformatics and Computational Biology Solutions Using R and Bioconductor **2005** Statistics for Biology and Health - Springer.

[8] Richard A. Gibbs and George M. Weinstock et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution *Nature* 428, 493-521 (1 April 2004)

[9] Irizarry Rafael A. , Bolstad Benjamin M., Collin Francois , Cope Leslie M., Hobbs Bridget and Speed Terence P. **(2003)**. Summaries of Affymetrix GeneChip probe level data *Nucleic Acids Research* 31(4):e15.

[10] Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP **(2003)** Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics* Vol. 4, Number 2: 249-264.

[11] http://inparanoid.sbc.su.se

[12] Wit Ernest and McClure John. **(2004)**. Statistics for Microarrays Design, Analysis and Inference. *John Wiley & Sons Ltd.*

[13] R Ramachandran, S Kakar Histological patterns in drug-induced liver disease. *Journal of Clin Pathology(2008)*

[14] Watkins PB and Seeff LB. Drug-induced liver injury: summary of a single topic clinical research conference. *Hepatology* 43:61831 (2006)

[15] Williams R. Classification, etiology, and considerations of outcome in acute liver failure. Semin Liver Dis 1996;16:3438

[16] Lee WM. Acute liver failure. Clin Perspect Gastroenterol 2001;2:10110.

[17] Larson AM, Polson J, Fontana RJ, et al. Acetaminophen-induced acute liver failure: results of a United States multicenter, prospective study. Hepatology 2005;42:136472.

[18] Stefano Monti , Pablo Tamayo , Jill Mesirov , Todd Golub. Consensus clustering – A resampling-based method for class discovery and visualization of gene expression microarray data. *Journal Machine Learning*. Volume 52 Issue 1-2, July-August 2003, 91 - 118

[19] Zhijin Wu, Rafael A. Irizarry, Robert Gentleman, Francisco Martinez Murillo, and Forrest Spencer **(2004)** A Model Based Background Adjustment for Oligonucleotide Expression Arrays. *Johns Hopkins Univ, Dept. of Biostatistics Working Papers. Working Paper 1.*

[20] Vapnik, V. N. *The Nature of Statistical Learning Theory* (2nd Ed.), Springer Verlag, 2000

[21] Yu Chuan Tai and Terence P. Speed. A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann. Statist.* Volume 34, Number 5 (2006), 2387-2412.

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.

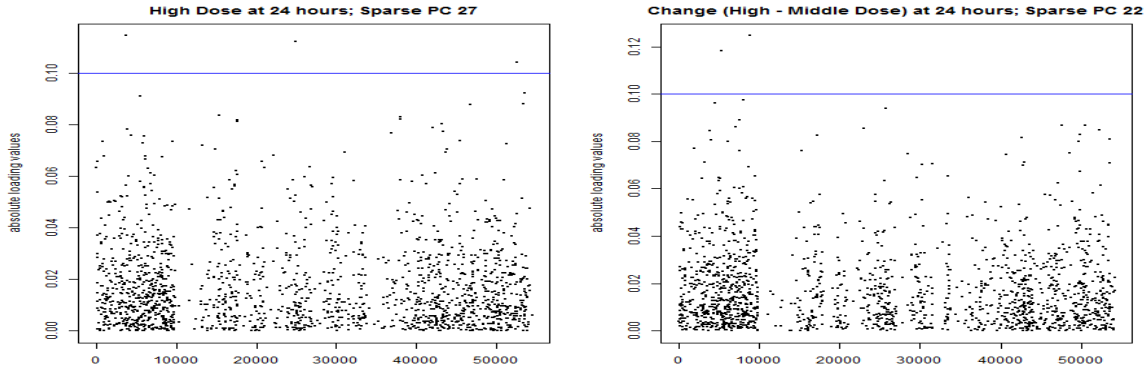
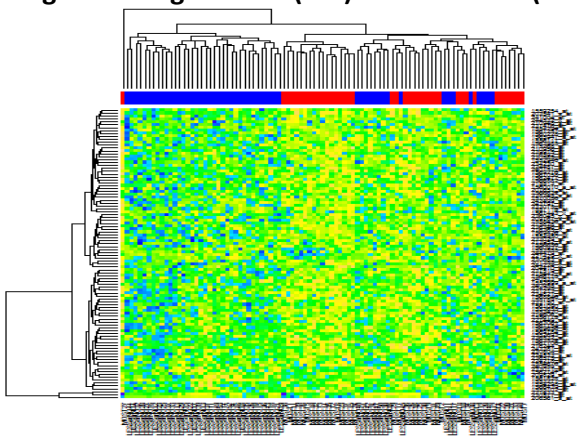


Figure 2: Heatmap showing clustering of most (red) vs. less-or-no (blue) DILI concern.



Cross-organism prediction of drug hepatotoxicity by sparse group factor analysis

Tommi Suvitaival*, Juuso A. Parkkinen, Seppo Virtanen
Helsinki Institute for Information Technology HIIT,
Department of Information and Computer Science, Aalto University
{tommi.suvitaival,juuso.parkkinen,seppo.j.virtanen}@aalto.fi

Samuel Kaski
Helsinki Institute for Information Technology HIIT,
Department of Information and Computer Science, Aalto University
Helsinki Institute for Information Technology HIIT,
Department of Computer Science, University of Helsinki
samuel.kaski@aalto.fi

Abstract

We investigate the problem of how to computationally generalize cell-level and clinical-level responses from model organisms to humans. We use a multi-view machine learning approach to detect associations between drug-induced transcriptional changes and organ-level damage. We show that the model learns associations that enable us to predict liver injury across organisms based on transcriptional responses. Moreover, the learned structure in the transcriptional data of the model organisms can separate drug compounds by both their therapeutic and toxicological effects on humans.

1 Introduction

We study the problem of how to computationally generalize associations between omics data and clinical-level data from model organisms to humans. The task is highly non-trivial because the organisms are different by their biological systems regardless of their distant relatedness. Additionally, ground-truth data for learning the effects of harmful interventions on humans are hard or impossible to obtain.

There is existing work on modeling conserved responses across organisms and for separating these responses from organism-specific signals in high-dimensional omics data [4, 5]. In these studies, the focus has been on detecting similarities between the biological systems in the two species. The next step to that is to translate the expected response to a condition from a model organism to the organism of interest.

In this paper, our goal is to find associations between high-throughput data views and generalize findings across organisms. Specifically, we formulate two modeling tasks: prediction of drug hepatotoxicity by gene expression across organisms (Task 1) and translation of drug effects from model organisms to humans (Task 2).

To solve the two tasks, we introduce a probabilistic multi-view model, sparse group factor analysis (GFA), and demonstrate its performance on the data collected by The Japanese Toxicogenomics Project [7]. The TGP data set includes clinical and gene expression data from three organisms after over 100 different medical treatments at multiple experimental conditions.

*To whom correspondence should be adressed.

2 Methods

2.1 Data set

The JTG data set includes gene expression data from three model organisms (primary hepatocyte cells from rat *in vivo*, rat *in vitro* and human *in vitro*) under conditions that can be summarized as three experimental factors (administered drug compound, dosage and time from the administration). For this analysis, we select the subset of experimental factor levels that are observed in all three organisms. This set includes 119 drug compounds administered at two dosage levels (middle and high) and measurements made at two time points after the treatment (8/9 h and 24 h). Histopathology of the liver has been examined from the rat *in vivo* experiments at the same time points and dosage levels, providing a pathological finding class and severity grading for each sample.

For the modeling task, we consider each combination of compound, dose and time as a single sample in the model. The gene expression observations were provided in the FARMS-summarized [2] format, which we use to compute the differential expression of the treated samples against the controls. We represent the pathological finding classes for each sample as a grade-weighted count. As the four data matrices (differential gene expression $\mathbf{X}^{\text{rat}}_{\text{in vivo}}$, $\mathbf{X}^{\text{rat}}_{\text{in vitro}}$ and $\mathbf{X}^{\text{human}}_{\text{in vitro}}$, and pathological findings $\mathbf{Y}$) are now matched by their samples, we call the matrices different *views* of the data.

2.2 Model

We use group factor analysis (GFA [8]) to learn associations between the gene expression measurements and pathological findings. GFA is an unsupervised Bayesian latent variable model designed to learn associations between multiple observed views of data [3] – i.e. associations between data matrices with matched samples.

GFA allows us to explore the data in a low-dimensional latent representation, where the data is decomposed into shared and view-specific components. Additionally, GFA can be used for prediction from a set of views to another set of views. In Task 1, we utilize the gene expression views to predict the pathological findings. We can also study the similarity of the samples, based on correlations between their latent space representations. We use this in Task 2 to evaluate whether the compounds deemed similar in the latent space are similar by their known therapeutic or toxic effects in humans.

To avoid overfitting to the high-dimensional gene expression data and to increase the interpretability of the model, we introduce sparsity to the projections between the latent space and the observed data views. Sparsity leads to a smaller subset of variables of the data being active in the model – also in the target view of the cross-view prediction task. Effectively, the model selects the variables that have the strongest associations within and between the views.

Many of the pathological finding classes, which we attempt to predict in Task 1, appear only few times in the entire TGP data set. A model predicting such targets is prone to overfitting. Sparse GFA overcomes this risk by automatically selecting the target classes that are feasible to predict.

3 Results

3.1 Task 1: Prediction of drug hepatotoxicity by gene expression across organisms

To investigate the strength of associations between the clinical-level responses and the changes in gene expression, we quantify the success at predicting pathological findings of the *in vivo* rats ($\mathbf{Y}$) based on gene expression data from the three organisms ($\mathbf{X}^{\text{rat}}_{\text{in vivo}}$, $\mathbf{X}^{\text{rat}}_{\text{in vitro}}$ and $\mathbf{X}^{\text{human}}_{\text{in vitro}}$). In a cross-validation setting, we learn GFA jointly using training data of the three gene expression views and the pathology view, and compare predictions from each of the gene expression views to the pathology view on test data.

We discover that predictors based on gene expression of the human and rat cell lines yield a mutually comparable prediction accuracy, while the predictor based on gene expression of the live rats yields a clearly superior performance (Fig. 1a). This is expected, as the pathological findings are also made on the live rats.

In comparison to a standard multi-output ℓ_1-regularized regression model [6], sparse GFA yields comparable or better predictions (Fig. 1b).

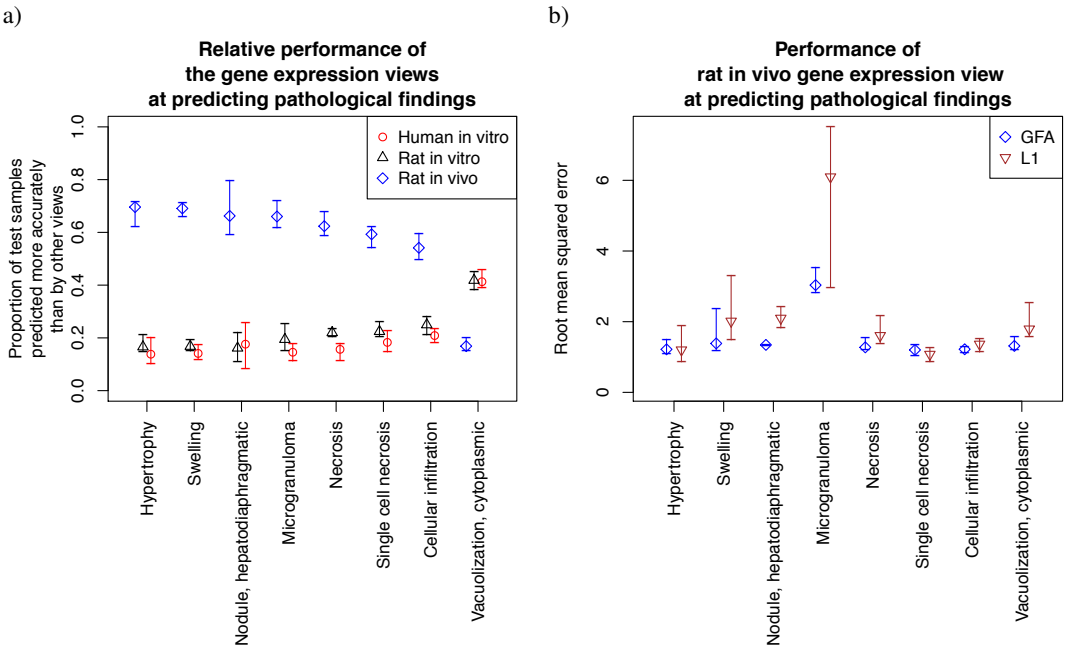


Figure 1: GFA-predictor based on gene expression of rat *in vivo* samples yields a superior prediction on pathological findings in the test data compared to gene expression of the *in vitro* samples (left). The absolute performance of GFA is comparable to or better than the performance of the ℓ_1-regularized multi-output regression model at predicting pathological findings in the test data based on gene expression of rat *in vivo* samples (right). The pathological finding classes (x-axis) are sorted by the performance of the predictor based on gene expression of rat *in vivo* samples. The confidence intervals are the maximum and minimum from 10 randomizations of cross-validation.

3.2 Task 2: Translation of drug effects from model organisms to humans

GFA allows us to explore the data in an unsupervised way in the low-dimensional latent space. Specifically, we want to investigate the model's ability to learn drug-induced changes in gene expression of the model organisms that can be generalized to system-level responses in humans.

To evaluate the generalizability, we use two types of ground-truth labels representing drug-induced effects in humans that have not been utilized by GFA: anatomical therapeutic chemical classification (ATC [9]) codes and drug-induced liver injury (DILI) labels [1]. We learn GFA for the three differential gene expression views of the model organisms ($\mathbf{X}^{\text{rat}}_{\text{in vivo}}$, $\mathbf{X}^{\text{rat}}_{\text{in vitro}}$ and $\mathbf{X}^{\text{human}}_{\text{in vitro}}$) and study the aggregation of similar drug compounds in the latent space of this joint model. We quantify the aggregation by computing the mean average precision score of the retrieval of similar compounds in the latent space. We also compute the randomized retrieval performance, providing a baseline for the study.

We discover that compounds with same ATC code (level 4) are strongly aggregated in the latent representation (Fig. 2a). Also the DILI labels are aggregated more than what would be expected (Fig. 2b). Aggregation by the DILI labels is not as strong as by the ATC codes. This may be due to the more heterogeneous nature of the responses to toxic compounds in comparison to the more coherent responses to normal therapeutic drugs.

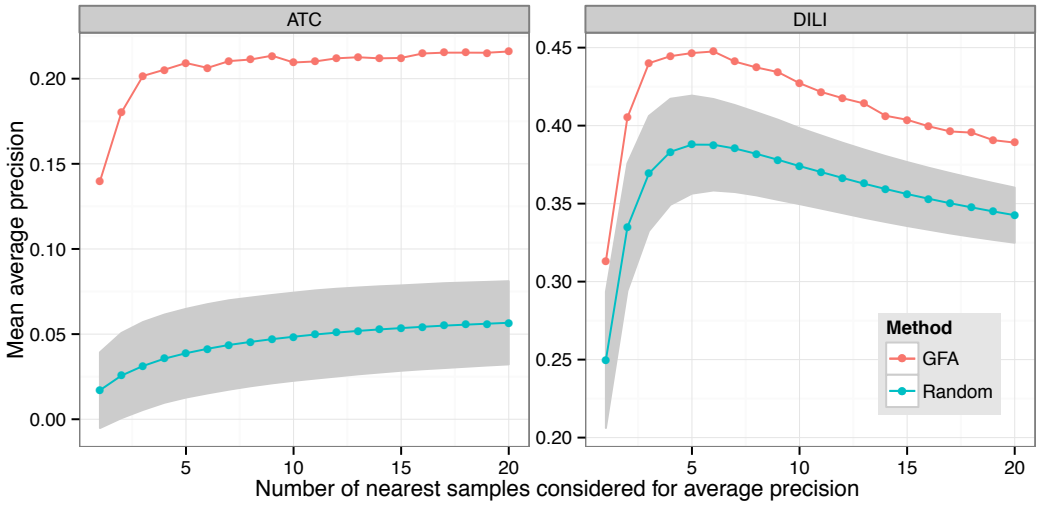


Figure 2: Similar drug compounds are significantly aggregated in the latent space in terms of both the ATC codes and DILI labels (left and right, respectively). The aggregation is quantified as mean average precision score of the retrieval of similar compounds in the latent space of GFA. The retrieval performance is shown as a function of the number of nearest neighbor compounds and compared to the performance in the same retrieval task after the random permutation of the compound labels.

4 Discussion

We have demonstrated that the proposed model – sparse group factor analysis – detects associations between transcriptional and clinical views across organisms in a way that generalizes beyond the immediate prediction task. The model allows us to explore the data in a low-dimensional latent space, revealing structure that can describe biological responses to drug compounds. In addition, we have shown that the cross-view predictive power of the model is comparable to a standard regularized regression model designed for the task.

Acknowledgments

Funding: The Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, 251170; Computational Modeling of the Biological Effects of Chemicals, 140057), Finnish Doctoral Programme in Computational Sciences FICS and Helsinki Doctoral Programme in Computer Science.

References

[1] Minjun Chen, Vikrant Vijay, Qiang Shi, Zhichao Liu, Hong Fang, and Weida Tong. FDA-approved drug labeling for the study of drug-induced liver injury. *Drug Discovery Today*, 16(15):697–703, 2011.

[2] Sepp Hochreiter, Djork-Arne Clevert, and Klaus Obermayer. A new summarization method for Affymetrix probe level data. *Bioinformatics*, 22(8):943–949, 2006.

[3] Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14:965–1003, 2013. Implementation in R available at http://research.ics.aalto.fi/mi/software/CCAGFA/.

[4] Hai-Son Le and Ziv Bar-Joseph. Cross species expression analysis using a Dirichlet process mixture model with latent matchings. *Advances in Neural Information Processing Systems*, 23:1270–1278, 2010.

[5] Tommi Suvitaival, Ilkka Huopaniemi, Matej Orešič, and Samuel Kaski. Cross-species translation of multi-way biomarkers. In Timo Honkela, Wlodzislaw Duch, Mark Girolami, and Samuel Kaski, editors,

Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN), Part I, volume 6791 of *Lecture Notes in Computer Science*, pages 209–216. Springer, 2011.

[6] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[7] Takeki Uehara, Atsushi Ono, Toshiyuki Maruyama, Ikuo Kato, Hiroshi Yamada, Yasuo Ohno, and Tetsuro Urushidani. The Japanese toxicogenomics project: application of toxicogenomics. *Molecular Nutrition & Food Research*, 54(2):218–227, 2010.

[8] Seppo Virtanen, Arto Klami, Suleiman A. Khan, and Samuel Kaski. Bayesian group factor analysis. In Neil Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *JMLR W&CP*, pages 1269–1277. JMLR, 2012. Implementation in R available at http://research.ics.aalto.fi/mi/software/CCAGFA/.

[9] WHO Collaborating Centre for Drug Statistics Methodology. ATC classification index with DDDs, 2013, Oslo 2012.

Matrix Factorization-Based Data Fusion for Drug-Induced Liver Injury Prediction

Marinka Žitnik[1] and Blaž Zupan[1,2]

[1] Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia
[2] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX-77030, USA

marinka.zitnik@fri.uni-lj.si, blaz.zupan@fri.uni-lj.si

Abstract. We report on a data fusion approach for prediction of outcome of drug-induced liver injury (DILI) in humans from gene expression studies as provided by the CAMDA 2013 Challenge. Our aim was to investigate if the data from all four toxicogenomics studies can be fused together to boost prediction accuracy. We show that recently proposed matrix factorization-based fusion provides an elegant framework for integration of CAMDA and related data sets. Our data fusion approach yields a high cross-validated AUC of 0.819 (in vivo assays), which is above the accuracy of standard machine learning procedures (stacked classification with feature selection). Achieved accuracy is also a substantial improvement of the highest scores on the same data sets reported in CAMDA 2012. Our data analysis shows that animal studies can be replaced with in vitro assays (AUC = 0.799) and that we can predict liver injury in humans from animal data (AUC = 0.811).

1 Introduction

Molecular biology abounds with data from sequencing, expression studies, function annotations, studies of interactions and other. These data sources are related, and analysis of one data set could benefit from inclusion of others. We have recently proposed a data fusion approach [1] that can elegantly integrate heterogeneous data sources, representing each data set in a matrix and fusing the data sets by simultaneous matrix factorization. We here report on the fusion of 29 data sets from CAMDA Challenge and related data repositories to predict DILI potential. We compare the accuracy of data fusion to that of a standard multi-classifier approach where we stack four state-of-the-art classification algorithms. We additionally investigate feature subset selection by CUR matrix decomposition [2] applied before stacking [3]. Our principal contribution is a demonstration that toxicogenomics studies can substantially benefit from data fusion.

2 Data fusion by Matrix Factorization

We use data fusion by matrix factorization [1], an intermediate data integration approach that is able to fuse heterogeneous data sources. Intermediate integration is often the preferred integration strategy [4,5,6] as it embeds the structure of the data into a predictive model and for this reason often achieves higher accuracy.

Data fusion considered 14 object types (nodes in Fig. 1, *e.g.,* drug, GO term, or drug type) and a collection of 29 data sources, each relating a pair of object types (arcs in Fig. 1, *e.g.,* gene annotations that relate genes and GO terms). In addition to FARMS-summarized expression data sets we include data on drugs available from DrugBank[3], gene annotations from Gene Ontology[4], protein-protein interactions from STRING[5], and

[3] http://www.drugbank.ca
[4] http://www.geneontology.org
[5] http://string-db.org

hematological and clinical chemistry data for each animal and array metadata information, the latter being provided by the challenge organizers. We did not use in vivo pathological findings in the fused model.

We represent the observations from a data source that relates two distinct objects types i and j in a sparse relation matrix $\mathbf{R}_{ij}$ (*e.g.,* $\mathbf{R}_{1,13}$ for annotations of genes in rat in vivo single study). A data source that provides relations between objects of the same type i is represented by a constraint matrix $\mathbf{\Theta}_i$ (*e.g.,* $\mathbf{\Theta}_{10,10}$ for DrugBank's drug interactions). Relation matrices $\mathbf{R}_{ij}$ are simultaneously factorized under constraints by $\mathbf{\Theta}_i$ [1]. The resulting system contains factors $\mathbf{S}_{ij}$ that are specific to each data source and factors $\mathbf{G}_i$ that are specific to each object type, such that each relation matrix $\mathbf{R}_{ij}$ is approximated as $\widehat{\mathbf{R}}_{ij} = \mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j^T$. Fusion takes place due to matrix factor sharing during decomposition of relation matrices.

We apply data fusion to infer relations between drugs and DILI potential, respectively. This relation, encoded in a target matrix $\mathbf{R}_{10,14}$, is observed in the context of all other data sources. Matrix $\mathbf{R}_{10,14} \in \mathbb{R}^{131\times 3}$ is a $[0,1]$-matrix that is only partially observed. Its entries indicate drugs' degree of membership to the three DILI severity classes, which are "No concern DILI", "Less concern DILI" and "Most concern DILI", respectively. We aim to predict the unobserved entries in $\mathbf{R}_{10,14}$ by reconstructing them through matrix factorization. The DILI severity of p-th drug is determined as $\arg\max_i \widehat{\mathbf{R}}_{10,14}(p,i)$.

3 Multi-Classifier Approach and Feature Subset Selection by CUR Matrix Decomposition

We use FARMS-summarized gene expression data for the four toxicogenomics studies that were provided by the organizers of the challenge [7]. We employ CUR matrix decomposition [2] to identify a small set of information carrying genes. CUR matrix decomposition in an unsupervised manner approximates target matrix $\mathbf{A}$ as $\mathbf{A} \approx \mathbf{CUR}$, where $\mathbf{C}$ and $\mathbf{R}$ are low-dimensional matrix factors that contain a subset of columns and rows from $\mathbf{A}$, respectively. The advantage of CUR decomposition over some well known low-rank matrix decompositions such as principal component analysis (PCA) or singular value decomposition (SVD) is its explicit representation in terms of a small number of actual columns and rows of target data matrix. The CUR decomposition-selected features correspond to original gene expression profiles instead of their linear combinations as with PCA and SVD. We then apply several state-of-the-art classifiers to predict the DILI concern in human from the matrix factor $\mathbf{C}$ obtained for each toxicogenomics study separately. We use gradient tree boosting with multinomial deviance as a loss function to model the three classes of DILI severity, random forests, support vector machine with polynomial kernel. Individual predictions are ensembled through stacking with logistic regression [3].

4 Results and Discussion

The performance of proposed inference approaches was estimated through 10-fold cross-validation. Feature subset selection for multi-classifier approach was performed on training data sets. Parameters of the classification and matrix decomposition algorithms, such as the number of iterations and the sizes of the constituent trees in gradient tree boosting, were estimated through internal cross-validation on the training data.

In our first experiment we considered the DILI prediction problem for each study separately and pursued a multi-classifier approach (Table 1). Feature subset selection by CUR

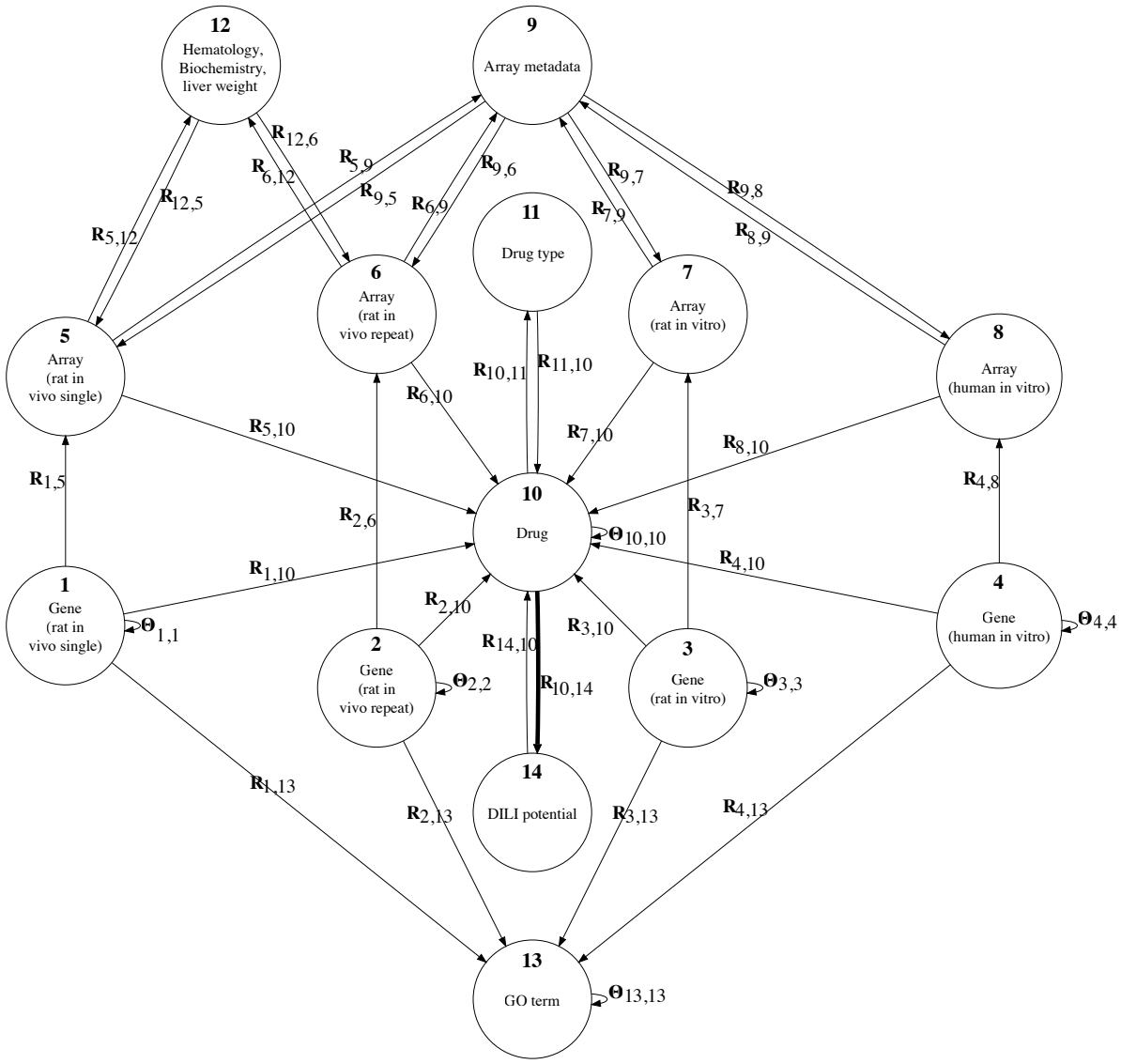


Fig. 1: Fused data sources. Nodes represent 14 object types. Arcs denote data sources that relate objects of different type (relation matrices, $\mathbf{R}_{ij}$) or objects of the same type (constraints, $\mathbf{\Theta}_i$) for a total of 29 matrices-data sources. Bold arc ($\mathbf{R}_{10,14}, \mathbf{R}_{14,10} = \mathbf{R}_{10,14}^T$) represents relation between drugs and DILI potential that we try to augment. Fused data sources include gene annotations that are encoded in $\{0,1\}$-matrices $\mathbf{R}_{1,13}$, $\mathbf{R}_{2,13}$, $\mathbf{R}_{3,13}$ and $\mathbf{R}_{4,13}$, expression profiles ($\mathbf{R}_{1,5}$, $\mathbf{R}_{2,6}$, $\mathbf{R}_{3,7}$, $\mathbf{R}_{4,8}$), hematology, body weight and clinical chemistry data for each rat ($\mathbf{R}_{5,12}$, $\mathbf{R}_{6,12}$, $\mathbf{R}_{12,5} = \mathbf{R}_{5,12}^T$, $\mathbf{R}_{12,6} = \mathbf{R}_{6,12}^T$), array metadata information such as dose level, dosage time and sacrifice time ($\mathbf{R}_{5,9}$, $\mathbf{R}_{6,9}$, $\mathbf{R}_{7,9}$, $\mathbf{R}_{8,9}$, $\mathbf{R}_{9,5} = \mathbf{R}_{5,9}^T$, $\mathbf{R}_{9,6} = \mathbf{R}_{6,9}^T$, $\mathbf{R}_{9,7} = \mathbf{R}_{7,9}^T$, $\mathbf{R}_{9,8} = \mathbf{R}_{8,9}^T$), drug targets ($\mathbf{R}_{1,10}$, $\mathbf{R}_{2,10}$, $\mathbf{R}_{3,10}$, $\mathbf{R}_{4,10}$), indication of medical drugs tested with arrays ($\mathbf{R}_{5,10}$, $\mathbf{R}_{6,10}$, $\mathbf{R}_{7,10}$, $\mathbf{R}_{8,10}$), structure and categorization of drugs ($\mathbf{R}_{10,11}$, $\mathbf{R}_{11,10} = \mathbf{R}_{10,11}^T$). Constraint matrices encode protein-protein interactions ($\mathbf{\Theta}_{1,1}$, $\mathbf{\Theta}_{2,2}$, $\mathbf{\Theta}_{3,3}$, $\mathbf{\Theta}_{4,4}$), drug interactions ($\mathbf{\Theta}_{10,10}$) and semantic structure of Gene Ontology graph ($\mathbf{\Theta}_{13,13}$).

matrix decomposition substantially reduced the number of features. For instance and as averaged across cross-validation folds, only about 300 features were used for training the prediction models in human in vitro study instead of original 18,988 features included by FARMS summarization. Solid performance of multi-classifier approach was not surprising [8,9], yet the substantial improvement of the AUC scores from CAMDA 2012 was.

Notice that we did not reimplement the procedures from [10], so the comparison of AUC scores is only indicative as they were obtained on different data samples chosen by cross validation. Yet the relatively large gains in AUC by our methods do provide evidence for improvements in prediction performance.

Notice also comparable performance of data preprocessing by CUR factorization and PCA. As CUR performs feature selection rather than feature transformation, it could be a preferable procedure to identify gene biomarkers.

Table 2 reports on 10-fold cross-validated accuracy for seven data fusion configurations that considered various subsets of the complete fusion model in Figure 1. The model inferred from all assays used an entire collection of data sources from Figure 1. Other models considered only selected toxicogenomics studies and associated non-expression data. For instance, fusion of in vivo assays omitted all data sets from in vitro studies (object types 3, 4, 7, and 8).

Data fusion surpassed the accuracy of multi-classifier approach to predict DILI potential in humans (Table 2). The most accurate model was inferred by fusing in vivo assays, which scored AUC of 0.819. It is surprising that in vivo assays, which relied on animal model, performed better than human assays, as we aim at predicting DILI potential in humans. However, last year's participants Pessiot *et al.*, 2012 [10] similarly observed that using in vivo animal data was more informative than using in vitro data from humans. Their AUC scores obtained by linear support vector machine classifier and inferred from separate toxicogenomics studies were substantially lower than those reported by our fusion-based approach. Also, fusion-based model inferred from animal assays (these are three studies, two in vivo and one in vitro study) outperformed model obtained by fusing human assays only (one human in vitro study), where the first achieved AUC of 0.811 and the latter AUC of 0.792. One might expect that administration of drugs to animal models would fail to identify the risk of liver injury for drugs prescribed to human due to differences in metabolic pathways and the current lack of suitable animal models that reproduce the human risk factors [11]. Our results do not confirm this hypothesis, although differences in performance are small and further investigations seem worthwhile pursuing.

Machine learning method	human in vitro	rat in vitro	rat in vivo single	rat in vivo repeated
Log. reg. stack. (RF, MD GBT, LR, SVM) w. PCA	0.741	0.765	0.748	0.761
Log. reg. stack. (RF, MD GBT, LR, SVM) w. CUR	0.758	0.755	0.764	0.778
Pessiot *et al.*, 2012 [10]	0.59	0.58	0.67	0.66
Clevert *et al.*, 2012 [12]		0.26*		

Table 1: Predictive performance of multi-classifier approach for DILI potential prediction with and without CUR dimensionality reduction. Reported are 10-fold cross-validated AUC scores. Acronyms: RF - random forests [13], MD GBT - multinomial deviance gradient boosting trees [14], LR - logistic regression, SVM - support vector machine (polynomial third degree kernel). CAMDA 2012 scores are from Pessiot *et al.* [10] and Clevert *et al.* [12] who used different cross-validation indices and data preprocessing. *Clevert *et al.* [12] reported the error rate and not AUC score.

Fused data	AUC
In vivo assays	0.819
All in vitro assays	0.790
Human in vitro assays	0.793
Animal in vitro assays	0.799
Animal assays	0.811
Human assays	0.792
All assays	0.810

Table 2: Predictive performance of fusing various subsets of assays for DILI potential prediction. Reported are 10-fold cross-validated AUC scores.

5 Conclusion

Data fusion allows us to simultaneously consider the available data for outcome prediction of drug-induced liver injury. Its models can surpass accuracy of standard machine learning approaches. Our results also indicate that future prediction models should exploit circumstantial evidence from related data sources in addition to standard toxicogenomics data sets. We anticipate that efforts in data analysis have the promise to replace animal studies with in vitro assays and predict the outcome of liver injuries in humans using toxicogenomics data from animals.

References

1. M. Žitnik and B. Zupan, "Data fusion by matrix factorization," (submitted), 2013.
2. M. W. Mahoney and P. Drineas, "CUR matrix decompositions for improved data analysis," *Proceedings of the National Academy of Sciences*, vol. 106, no. 3, pp. 697–702, 2009.
3. D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.
4. M. H. van Vliet, H. M. Horlings, M. J. van de Vijver, M. J. T. Reinders, and L. F. A. Wessels, "Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome," *PLoS One*, vol. 7, no. 7, p. e40358, 2012.
5. O. Gevaert, F. De Smet, D. Timmerman, Y. Moreau, and B. De Moor, "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks," *Bioinformatics*, vol. 22, no. 14, pp. e184–90, 2006.
6. G. R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble, "A statistical framework for genomic data fusion," *Bioinformatics*, vol. 20, no. 16, pp. 2626–2635, 2004.
7. S. Hochreiter, D.-A. Clevert, and K. Obermayer, "A new summarization method for affymetrix probe level data," *Bioinformatics*, vol. 22, no. 8, pp. 943–949, 2006.
8. S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?" *Machine learning*, vol. 54, no. 3, pp. 255–273, 2004.
9. G. Pandey, B. Zhang, A. N. Chang, C. L. Myers, J. Zhu, V. Kumar, and E. E. Schadt, "An integrative multi-network and multi-classifier approach to predict genetic interactions," *PLoS Computational Biology*, vol. 6, no. 9, 2010.
10. J.-F. Pessiot, P. S. Wong, T. Maruyama, R. Morioka, S. Aburatani, M. Tanaka, and W. Fujibuchi, "The impact of collapsing data on microarray analysis and DILI prediction," *CAMDA 2012 Challenge*, pp. 21–25, 2012. [Online]. Available: http://camda.bioinfo.cipf.es/camda2012/_media/camda2012abstracts_updated.pdf
11. N. Kaplowitz, "Avoiding idiosyncratic DILI: Two is better than one," *Hepatology*, 2013.
12. D.-A. Clevert, M. Heusel, A. Mitterecker, W. Talloen, H. Göhlmann, J. Wegner, A. Mayr, G. Klambauer, and S. Hochreiter, "Exploiting the Japanese toxicogenomics project for predictive modelling of drug toxicity," *CAMDA 2012 Challenge*, pp. 26–29, 2012. [Online]. Available: http://camda.bioinfo.cipf.es/camda2012/_media/camda2012abstracts_updated.pdf
13. L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
14. J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

Using probabilistic models of signaling pathways to predict *in vivo* drug activity.

Patricia Sebastián-León[1] and Joaquín Dopazo[1,2,3,*]

1 Department of Computational Genomics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, 46012, Spain.
2 CIBER de Enfermedades Raras (CIBERER), Valencia, 46012, Spain,
3 Functional Genomics Node (INB) at CIPF, Valencia, 46012, Spain;
* To whom correspondence should be addressed. Tel: +34 963289680; Email: jdopazo@cipf.es

Abstract

Signaling pathways constitute a valuable source of information that allows interpreting the way in which the cell respond to external stimulus and the aspects of the cell functionality affected by these. Here we explore the effect of drugs in cell signaling and the feasibility of using signaling to predict drug effect. A simple probabilistic model of 23 rat KEGG signaling pathways is used to compare the impact of drugs *in vitro* and *in vivo*. Our results document that almost all the pathways (20 out of the 23 pathways modeled) were affected in one or more stimulus-response circuits in the same way both in cell lines and in the *in vivo* experiment. This effect was observed for half of the drugs tried. Therefore models of cell signaling can be used as predictor of *in vivo* activity from *in vitro* activity in a reasonable number of cases. The advantage of using such models is that they permit an unprecedented insight into the mechanisms of drug effect and also understanding the differences between the *in vitro* and the *in vivo* systems.

Introduction

Signaling pathways represent the way in which the combined effect of gene activity elicits cell-level responses by activating/deactivating specific functionalities in response to particular stimulus through a chain of intermediate molecules. Drugs can either act as external stimulus or directly interfere with the genes of the pathway, causing changes in the "normal" responses. Such changes can be used to understand the biological consequences of the effect of drug in the cell, as well as to give clues on the drug mechanism of action. Despite a different behavior of signaling pathways is expected when cell lines are compared to organs or tissues, some affected signaling mechanisms could be common to certain drugs and could be used to predict *in vivo* activity. We have used the KEGG (Kanehisa, et al., 2012) repository, which contain detailed information about pathways, to obtain the templates for the derivation of the probabilistic models.

Methods

If the individual probabilities of protein presence/absence of all the proteins in the pathway are known, a simple probabilistic model of the pathway can be used to calculate the probabilities for signal transmission from any receptor protein to any final effector protein (taking into account all the intermediate activator and/or repressor proteins in between). Here, we take gene expression values as proxies of gene activity

and, consequently, presence/absence of the corresponding protein (Efroni, et al., 2007). We have used more than 10,000 Affymetrix microarrays downloaded from the GEO database (Barrett, et al., 2013) to derive the empirical distributions of presence/absence for each probe, that are further used to calculate the probability of presence/absence for the to the genes involved in the studied pathways (Efroni, et al., 2007; Sebastian-Leon, et al., 2013). Nodes have been treated in different ways depending on whether they were composed by alternative proteins (redundancy: only one of them keeps the node working) or complexes (all proteins are indispensable to keep the node working). This simplification has proven to be useful in practical terms (Sales, et al., 2012). Therefore, given the measurements of gene expressions in a particular experiment, the reference distributions can be sued to estimate the probabilities of presence/absence of each protein (and each node) of the pathway.

Once such probabilities have been estimated, the probability of signal transmission along a stimulus-response circuit can easily be inferred from the probabilities of activation of all the connecting nodes that constitute the circuit (providing that inhibitor nodes allow signal transmission when they are deactivated). The circuits are defined by the 23 KEGG pathways of rat used here (see Table 1). Therefore, the stimulus-response circuits of any of the pathways can easily be modeled by means of a simple product of probabilities (using the principle of inclusion/exclusion when bi- or multi-furcating stretches are present) (Sebastian-Leon, et al., 2013). This provides a straightforward approach to estimate the probability of signal transmission from gene expression values. However, such probabilities of signal transmission when out of context are not informative. What is interesting is the comparison of such probabilities in two different conditions (typically cases versus controls). We apply a Wilcoxon test (Wilcoxon, 1945) that allows detecting which stimulus-response circuits significantly change their probabilities of signal transmission between the compared conditions.

Here, we compare the changes induced by a collection of 132 drugs from the TGP dataset from the Japanese Toxicogenomics Project (Uehara, et al., 2010) in the different circuits of different pathways both, *in vitro* and *in vivo*.

The models of the pathways have recently been published (Sebastian-Leon, et al., 2013) and are available at: http://pathiways.babelomics.org/

Results

For each drug, we carried out all the comparisons between the doses tried *in vitro* and *in vivo*, independently. For any of these comparisons, we studied which circuits in which pathways displayed a significant change in the activity induced by the drug, as well as the type of change experimented (activation or inhibition). Table 1 shows the pathways in which the drugs caused the same type of alterations in one or several stimulus-response circuits. A total of 931 different circuits from all the pathways were affected by one or more drugs. Cell lines are more affected by drugs than the corresponding *in vivo* counterparts (by more than a 25% in average). However, only 207 stimulus-response circuits, corresponding to almost all the pathways (20 out of a total of 23

modeled) represented in Table 1 display coincident patterns of activation in response to several of the drugs tried. Almost half of the drugs tried (58 out of 132) caused an identical effect both *in vitro* and *in vivo* in at least one circuit of at least one pathway.

KEGG ID	Name	Drugs
rno03320	PPAR SIGNALING PATHWAY	bendazac, benzbromarone, benziodarone, clofibrate, fenofibrate, furosemide, gemfibrozil, simvastatin, sulfasalazine, WY-14643
rno04115	p53 SIGNALING PATHWAY	colchicine, disopyramide, ethionine, moxisylyte, nitrosodiethylamine, propylthiouracil, puromycin_aminonucleoside, quinidine
rno04060	CYTOKINE-CYTOKINE RECEPTOR INTERACTION	diazepam
rno04210	APOPTOSIS	hydroxyzine, nitrofurantoin
rno04340	HEDGEHOG SIGNALING PATHWAY	
rno04514	CELL ADHESION MOLECULES	caffeine, aproxen, nitrofurazone, tacrine, colchicine, gentamicin
rno04612	ANTIGEN PROCESING AND PRESENTATION	flutamide, puromycin_aminonucleoside
rno04662	B CELL RECEPTOR SIGNALING PATHWAY	nimesulide, nitrofurazone, chloramphenicol, colchicine, mexiletine, gentamicin, hydroxyzine, sulpiride
rno04916	MELANOGENESIS	Doxorubicin, isoniazid
rno04012	ERBB SIGNALING PATHWAY	hydroxyzine, nitrofurantoin, colchicine, ethionine, colchicine, caffeine
rno04310	WNT SIGNALING PATHWAY	Caffeine, ibuprofen
rno04370	VEGF SIGNALING PATHWAY	Acetamidofluorene, cyclophosphamide, danazol, diazepam, ethambutol, ethinylestradiol, ibuprofen, cyclosporine_A, diazepam, ajmaline, ethinylestradiol, ethambutol, nitrofurantoin, nitrofurantoin
rno04530	TIGHT JUNCTION	caffeine, cisplatin, naproxen, sulindac, ethionine, gentamicin, monocrotaline, puromycin_aminonucleoside
rno04630	JAK-STAT SIGNALING PATHWAY	diclofenac, disopyramide, furosemide, ibuprofen, sulindac
rno04664	Fc EPSILON RI SIGNALING PATHWAY	colchicine, ethionine, gentamicin, penicillamine, valproic_acid
rno04920	ADIPOCYTOKINE SIGNALING PATHWAY	diclofenac, naphthyl_isothiocyanate, naproxen, colchicine,
rno04020	CALCIUM SIGNALING PATHWAY	ethionine, hydroxyzine, caffeine
rno04330	NOTCH SIGNALING PATHWAY	Methimazole, naproxen
rno04512	ECM-RECEPTOR INTERACTION	nifedipine
rno04540	GAP JUNCTION	carbon_tetrachloride
rno04660	T CELL RECEPTOR SIGNALING PATHWAY	colchicine, ethionine, gentamicin, penicillamine, valproic_acid, caffeine, disopyramide, naproxen, sulindac, naphthyl_isothiocyanate, hydroxyzine, coumarin
rno04912	GnRH SIGNALING PATHWAY	disopyramide, naproxen, iproniazid

Figure 1 shows in detail the activity of several drugs in the PPAR signaling pathway. A total of twelve drugs significantly trigger the activation of the lipid metabolism and the adipocyte differentiation both *in vitro* and *in vivo*. This functional activation is attained through the activation three main stimulus-response circuits (in red in the figure). The detail provided by the model allows understanding the ways through the drugs are acting in the cell, as well as detecting other valuable collateral drug effects, as side effects, drug resistances, etc., providing these have a significant impact in any of the modeled pathways.

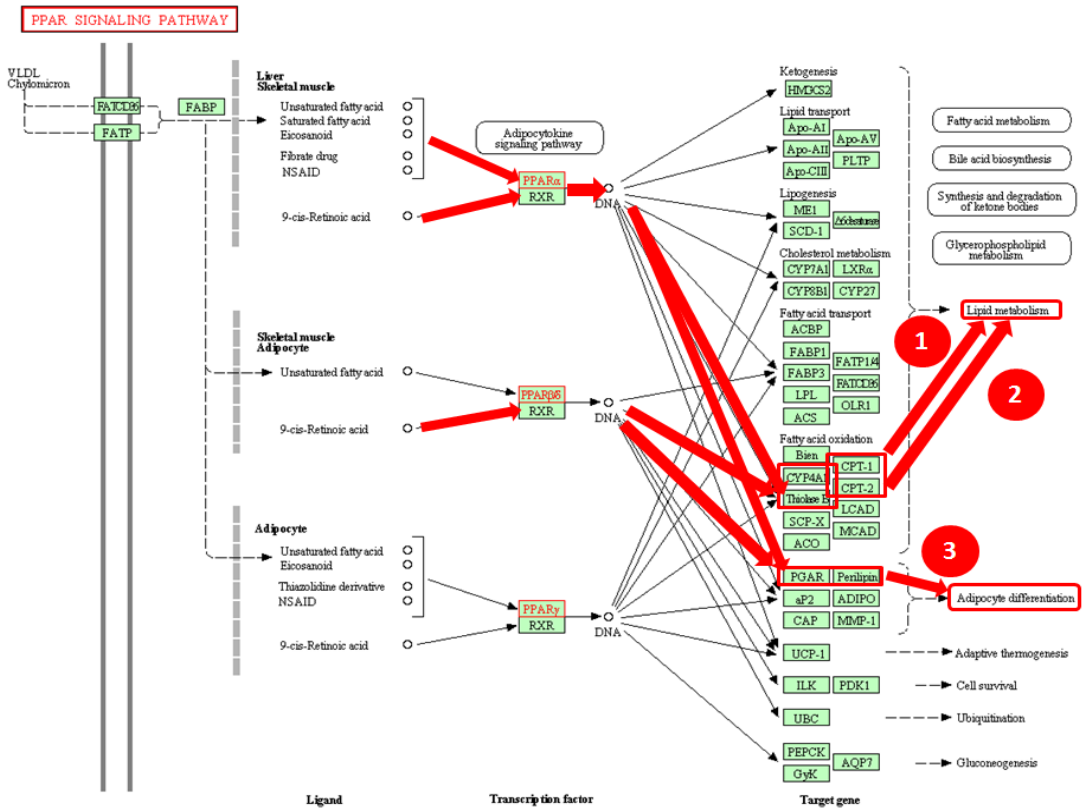


Figure 1. Rat pathway PPAR (rno03320) with red arrows indicating activation of signaling circuits by different drugs. 1) Circuit activated by: benzbromarone, clofibrate, fenofibrate, naproxen, WY-14643, omeprazole; 2) circuit activated by: benziodarone, sulfasalazine; 3) circuit activated by: bendazac, benzbromarone, fenofibrate, gemfibrozil, simvastatin, WY-14643. The effect of the drugs is an activation of the lipid metabolism and the adipocyte differentiation.

Discussion

Cell lines have extensively been used for initial *in vitro* testing of drugs. However, its validity as models of *in vivo* systems is questionable. Recent studies demonstrate that the global pattern of gene expression of cell lines is completely different to any other cell type, either healthy or diseased (Lukk, et al., 2010). However, this quantitative observation does not provide any information about the extent at which cell lines still retain similar functionalities of the cell type from which they have been derived from. Here we have used a simple probabilistic model that transforms gene expression levels into probabilities of signal transmission across signaling pathways, from receptor nodes, which receive the stimulus, to the effector nodes that trigger the corresponding response. In this way, gene expression data, of often difficult interpretation, are transformed into meaningful functional information regarding changes in the different pathway responses triggered by particular stimulus.

Our observations document a different behavior of cell lines with respect to their *in vivo* counterparts. However, such differences are not as radical as the behaviors described for the global gene expression (Lukk, et al., 2010) and only affects to about a 25% of the

signaling circuits in the average. This indicates that, despite the disparity in global gene expression, the global behaviors are, probably, not so dissimilar.

Using pathways to assess drug responses have a number of limitations. Firstly, there are drugs (half of the drugs tested here) that will not affect to the set signaling pathways modeled and therefore their effects will remain undetectable. In other cases, extensive responses, mainly observed *in vitro*, mask the induction or repression of common circuits that might be useful to predict drug activity.

Despite the described limitations, our results suggest that the use of models of pathways can offer an interesting alternative to other "black box" methods for drug activity prediction. More detailed modeling of cell activity, including metabolic pathways and other aspects such as regulation, protein interaction, etc., will probably increase the predictive accuracy offering, at the same time, valuable information on the drug action mechanisms.

Funding

This work is supported by grants BIO2011-27069 from the Spanish Ministry of Economy and Competitiveness (MINECO), PROMETEO/2010/001 from the Conselleria de Educacio of the Valencia Community. We also thank the support of the National Institute of Bioinformatics (www.inab.org), the CIBER de Enfermedades Raras (CIBERER), both initiatives of the ISCIII, MINECO and the Bull Chair in Computational Genomics (http://bioinfo.cipf.es/chair_compgenom).

References

Barrett, T., *et al*. (2013) NCBI GEO: archive for functional genomics data sets--update, *Nucleic Acids Res*, **41**, D991-995.
Efroni, S., Schaefer, C.F. and Buetow, K.H. (2007) Identification of key processes underlying cancer phenotypes using biologic pathway analysis, *PLoS ONE*, **2**, e425.
Kanehisa, M., *et al*. (2012) KEGG for integration and interpretation of large-scale molecular data sets, *Nucleic Acids Res*, **40**, D109-114.
Lukk, M., *et al*. (2010) A global map of human gene expression, *Nat Biotechnol*, **28**, 322-324.
Sales, G., *et al*. (2012) graphite - a Bioconductor package to convert pathway topology to gene network, *BMC Bioinformatics*, **13**, 20.
Sebastian-Leon, P., *et al*. (2013) Inferring the functional effect of gene expression changes in signaling pathways, *Nucleic Acids Res*, **In press**.
Uehara, T., *et al*. (2010) The Japanese toxicogenomics project: application of toxicogenomics, *Mol Nutr Food Res*, **54**, 218-227.
Wilcoxon, F. (1945) Individual comparisons by ranking methods, *Biometrics Bulletin*, **1**, 80-83.

Similarity in Network Structures for in vivo and in vitro Data from the Japanese Toxicogenomics Project

Ryan Gill[1], Somnath Datta[2], Susmita Datta[2]
[1]Department of Mathematics, University of Louisville, Louisville, KY 40292, USA
[2]Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, USA

1. Introduction We provide a partial answer to the important question in Toxicogenomics whether in-vivo microarray expression data based on animal studies can be replaced by in-vitro data. We consider the TGP dataset which contains over 21,000 arrays for rats treated with mainly human drugs and profiled using the Affymetrix RAE230_2.0 GeneChip®. The main target organ profiled is liver. In a previous study, Uehara et al. (2010) identified the genes commonly up-regulated both *in vivo* and *in vitro* after treatment with three different drugs clofibrate, WY-14643 and gemfibrozil. This study was one of the first to create an *in vivo–in vitro* bridge for the validation of a genomic biomarker with those three compounds. In this analysis, we try to provide a comprehensive view of the *in vivo–in vitro* bridging across all the genes (probe sets) for all the 131 drugs provided in the challenge data. Moreover, our approach is not only to observe the similarities in gene expressions of individual genes but to identify the similarities of the network connectivity of all the similar genes across all the chemicals. Methodologically, we consider this question from a statistical perspective and apply a significance test to examine if there is a difference between the genomic networks for the two different types (*in vivo*/*in vitro*) after accounting for different dosages of the drugs, and sacrifice times of the rats. In order to construct the networks of genes and then finding the differences/similarities of the networks for the two types we use the approach similar to the framework for differential network analysis described in our earlier work in Gill et al. (2010). Construction of the networks for each type of data is based on a connectivity score measuring the association between each pair of genes. We apply a connectivity score constructed using a partial least squares (PLS) method that captures the predictability of each gene's expression from a pairing gene after adjusting for other genes and additional covariables (such as dosage) and thus extending our earlier approach to network and differential network analysis (Pihur et al., 2008; Gill et al., 2010; Gill et al., 2012).

In order to study the expression pattern and the network structures, important data preprocessing is required to account for type, dose, and sacrifice time effects. There are substantial differences between the expression values of the MAS5 preprocessed data from the *in vivo* and *in vitro* samples and any naive attempt (such as a gene by gene t-test) might find that all genes are significantly differentially expressed in the two types. We build in the additional preprocessing in our linear model (ANOVA) for log-gene expressions. Similarly, these effects are included in our model for the computation of the PLS scores for the network analysis. These are detailed in the next section.

2. Data We analyze part of the challenge dataset from the Japanese Toxicogenomics Project and compare the MAS5 preprocessed data from the "single dose study *in vivo* experiment using Sparague-Dawley rats" with the "*in vitro* study using hepatocytes from Sparague-Dawley rats" for 131 drugs. The *in vivo* dataset for each drug has microarray expression

values of 31099 genes for 48 rats at four different dose concentrations (control, low, middle, and high) and four different sampling times (3, 6, 9, and 24 hours) with three observations at each combination of the levels for these factors. The in-vitro dataset for each drug has microarray expression values of the same genes for 24 rats at four dose concentrations with the same labels and three different sampling times (2, 8, and 24 hours) with two observations at each combination of the levels. The possibility of using the FARM preprocessed data was also considered, but many of the drugs have many genes with expression value 0 for all observations which precludes the use of regression or even correlation methods since there is no variation in the value of these variables.

3. Methods First, we used a nested ANOVA model to assess the effects of *TYPE* (*in vivo*/*in vitro*), drug dose (*DOSE*) , and sacrifice time (*SAC*) on the expression levels of 31099 genes for each drug. Specifically, for each drug the mean expression value for the ith observation for the gth gene is modeled as

$$\mu_{ig} = TYPE_{ig} + (TYPE{*}SAC{*}DOSE)_{ig}.$$

Before fitting the ANOVA model we take the logarithm of the centered expression levels; the logarithm of the expression values are centered with respect to all genes of the given type. For each drug, the p-values for *TYPE* are computed for each gene under the assumption that the expression values follow a normal distribution with homogeneous error variance. We use these preliminary ANOVA analysis to determine the genes for which the expression are not significantly different for two different types (*in vivo* vs. *in vitro*) at a pairwise type 1 error rate of 0.05. Summarizing the results for all the drugs we find there are 473 genes for which the *TYPE* effect is not significant for at least 80% of the drugs. In other words, the expression profiles of this common set of genes appear to be similar for many of the drugs. Thus, these 473 genes can be taken as common bridging genes between *in vivo* and *in vitro* studies across a great majority of the drugs. However, as the genes do not work independently we want to construct the network of those genes and check their differential behavior across two types.

The tests described in this section are based on connectivity scores s_{ik} which measures the association between the ith and kth genes in a network. Our earlier methods (Gill et al., 2010) for differential network connectivity are modified to allow for additional covariates. We estimate the coefficients for these additional covariates at the same time that the coefficients used to compute the connectivity scores are obtained. Let x_i be the centered and scaled n-dimensional expression vector for the ith gene. The method of computing the PLS scores that is described in Pihur et al. (2008) uses separate PLS models $x_i = \sum_{j \neq i} b_{ij} x_j +$ error, for each gene i. However, in the present context, adjustments for additional effects such as the dose levels are needed; thus we create additional covariate vectors z_1, ..., z_m and fit a set of linear models of the form $x_i = \sum_{k=1}^{k} a_{ik} z_k + \sum_{j \neq i} b_{ij} x_j +$ error. PLS regression is used to estimate the coefficients $a_{i1}, ..., a_{im}, b_{i1}, ..., b_{i,i-1}, b_{i,i+1}, ..., b_{ip}$ based on the design matrix formed by the covariates in the PLS model. The PLS scores are computed based on the estimates $b_{i1}, ..., b_{i,i-1}, b_{i,i+1}, ..., b_{ip}$. The details of the method for computing the PLS regression estimates of the regression coefficients and their conversion to PLS scores are omitted in this extended abstract; these were along the same lines as Pihur et al. (2008). A symmetrized estimate of regression coefficient b_{ij} is taken as the PLS association score $s_{ik} = (\widehat{b}_{ij} + \widehat{b}_{ji})/2$.

Once the connectivity scores are computed for each network, a permutation test is performed to test for differential connectivity of the class of all genes or the test for a single gene. Let $s_{ik}^{(1)}$ and $s_{ik}^{(2)}$ denotes the connectivity scores between genes i and k for networks 1 and 2, respectively. The test statistic for the class of all genes $\mathcal{F}$ with cardinality f is

$$\Delta = \frac{1}{f(f-1)} \sum_{i \neq j \in \mathcal{F}} D(s_{ik}^{(1)}, s_{ik}^{(2)}) \tag{1}$$

and the test statistic for a single gene g is

$$d(g) = \frac{1}{p-1} \sum_{i \neq g} D(s_{ig}^{(1)}, s_{ig}^{(2)}), \tag{2}$$

where D computes the distance between the connectivity scores. We have worked with the L_1-distance $D(s^{(1)}, s^{(2)}) = |s^{(1)} - s^{(2)}|$ rather than the more commonly used L_2-distance leading to a more robust analysis. The permutation test is performed by randomly assigning the labels to each observation in the data set formed by combining the observations from both networks.

4. Results For each of the 131 drugs, tests for differential connectivity of the networks on the set of all 473 non-differentially expressed genes (1) were performed using 1000 permutations based on the L_1 distance function and the PLS connectivity scores. No significant differences in the overall connectivity scores of the networks of this set of 473 genes were found for 77 of the 131 drugs at a 5% significance level. These drugs are listed in Table 1.

acarbose	disopyramide	nimesulide
acetamidofluorene	disulfiram	nitrosodiethylamine
acetaminophen	doxorubicin	papaverine
acetazolamide	enalapril	penicillamine
adapin	erythromycin ethylsuccinate	phenacetin
amitriptyline	ethambutol	phenobarbital
bendazac	ethinylestradiol	phenylanthranilic acid
benziodarone	ethionamide	propylthiouracil
bromoethylamine	etoposide	puromycin aminonucleoside
bucetin	famotidine	quinidine
captopril	fenofibrate	simvastatin
carboplatin	fluphenazine	sulindac
cephalothin	flutamide	sulpiride
chloramphenicol	gentamicin	tamoxifen
chlormadinone	griseofulvin	tannic acid
chlormezanone	hydroxyzine	terbinafine
chlorpheniramine	imipramine	tetracycline
chlorpromazine	labetalol	theophylline
chlorpropamide	lomustine	thioridazine
ciprofloxacin	lornoxicam	ticlopidine
clomipramine	mefenamic acid	tiopronin
colchicine	meloxicam	tolbutamide
cyclosporine A	metformin	triamterene
danazol	methyltestosterone	triazolam
dantrolene	mexiletine	trimethadione
diltiazem	nifedipine	

Table 1: Drugs with similar connectivity scores in the two networks.

Even among the 54 drugs for which the set of all genes are significantly different in terms of overall network connectivities, there are many genes that are not significantly different in terms of individual connectivity scores in the two networks at a 5% level. Tests for the significance difference of the connectivity score of each individual gene within the network (2) were performed for the 54 drugs, and there were 35 genes that were not differentially connected for at least 70% of the drugs. These genes are shown in Table 2.

GENE	prop.	GENE	prop.	GENE	prop.	GENE	prop.	GENE	prop.
1385656_at	0.833	1397371_at	0.759	1395446_at	0.741	1381550_at	0.722	1392859_at	0.704
1395874_at	0.815	1396604_at	0.759	1375063_at	0.741	1370626_at	0.722	1388033_at	0.704
1378788_at	0.796	1392389_at	0.759	1396731_at	0.722	1398741_at	0.704	1385031_at	0.704
1396340_at	0.778	1391493_at	0.759	1385655_at	0.722	1398675_at	0.704	1383272_at	0.704
1393711_at	0.778	1368887_at	0.759	1385589_at	0.722	1397850_at	0.704	1383195_at	0.704
1391313_at	0.778	1368854_at	0.759	1384683_at	0.722	1397720_at	0.704	1381502_at	0.704
1398707_at	0.759	1397339_at	0.741	1384061_at	0.722	1395490_at	0.704	1377391_at	0.704

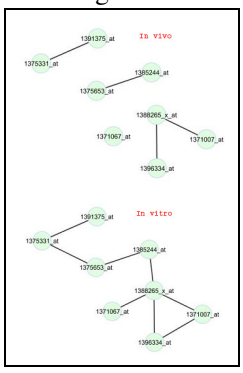
Table 2: Genes not differentially expressed for at least 70% of the remaining 54 drugs. The respective gene names (probe set IDs) and proportion of drugs with similar connectivity scores for that gene in the *in vivo* and *in vitro* networks.

In order to characterize the 473 genes which have shown no significant difference between the *in vivo* and *in vitro* types with more than 80% of the drugs we used functional annotation tool DAVID (Huang et al., 2009a; 2009b). Results of that analysis for the top five functional clusters out of the 473 genes are given in Table 3. Most of the genes in the first functional cluster are involved in neuron development, neuron differentiation, neuron projection morphogenesis and cell morphogenesis activities. The genes in the second most important cluster are involved with proteins in cell-cell junctions of multi-cellular species and also most of them are associated with some synaptic activities. The third most important functional cluster of the genes are associated with epidermal growth factor (EGF) proteins.

Cluster	Enrichment Score	% of drugs
1	4.05	87
2	3.08	92
3	2.20	90
4	1.49	92
5	1.48	95

Table 3: Tests of differential connectivity for the top 5 clusters obtained from the DAVID Functional Annotation Tool. The last column shows the percentages of drugs for which the corresponding sub-networks were not significantly different.

Figure 1: *In vivo* and *in vitro* networks for cluster 4 and the drug phenylbutazone. Edges are displayed for gene pairs with connectivity scores (rescaled so that the largest score for the network is 1 in magnitude) greater than 0.5 in magnitude.



Next, we reconstructed the networks separately for each functional cluster. These networks had fewer significant differences between the *in vivo* and *in vitro* types than the overall

networks. As seen in the Table 3, the difference between the *in vivo* and *in vitro* networks are not statistically significant for at least 87% of the drugs among these top five clusters.

We also annotated 35 genes for each of which the individual network connectivity score between the *in vivo* and *in vitro* types remained unchanged in spite of having significantly different total gene set network connectivity scores under the treatment of 54 drugs. With DAVID annotation tool we figured that all these 35 genes are in one functional cluster and they are associated with cellular macromolecular complex assembly.

Lastly, we wanted to illustrate how these sub-networks behave for a given drug. Figure 1 illustrates the constructed *in vivo* and *in vitro* networks for the genes in cluster 4 for phenylbutazone, a non-steroidal anti-inflammatory drug (NSAID). For these networks, the test for differential connectivity is not significant (p-value is 0.42). All edges in the *in vivo* network also appear in the *in vitro* network, and only 4 edges in the *in vitro* network do not appear in the *in vivo* network.

5. Conclusion A comprehensive view of the *in vivo - in vitro* bridge of the genes using the rat microarray TGP study under all the drugs is undertaken. We not only provide the similarity of individual gene expression pattern but also that of the association networks under *in vivo* and *in vitro* experiments. The systems are scrutinized in terms of overall network connectivity and also in terms of individual gene connectivity. We use PLS based association scores adjusted for sacrifice time and dosage followed by a permutation based statistical test with those scores. Since we are trying to identify genes that are not different, a conservative approach in this context will be not to apply a multiple testing p-value correction unlike typical gene expression studies where the goal is to identify genes that are differentially expressed and/or connected under two biological conditions. It is interesting to observe that, similar to Uehara et al. (2010) who studied three of the drugs, none of the bridging genes that we found are involved with cell proliferation and apoptosis.

A potential limitation of our study is that our findings are based on a specific type of statistical model. In the future we plan to undertake additional investigation where networks are constructed by fitting other types of predictive models such as lasso (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005) and the results are compared.

The findings must be interpreted carefully. First of all, we have highlighted the genes which were not significantly different. However it does not quite imply that *in vivo* and *in vitro* studies are completely interchangeable since there are genes that show differential expression and network profiles in the two networks. Furthermore, lack of statistical significance does not necessarily imply that the objects under comparison are indeed equal.

References

Gill, R., Datta, S., and Datta, S. (2010). *BMC Bioinformatics*, **11**, 95.

Gill, R., Datta, S., and Datta, S. (2012). http://CRAN.R-project.org/package=dna

Pihur, V., Datta, S., and Datta, S. (2008). *Bioinformatics*, **24**, 561-568.

Uehara T., Ono A., Maruyama T., Kato I., Yamada H., Ohno Y., Urushidani T. (2010). *Mol. Nutr. Food Res.*, **54**, 218-227.

Huang, D.W., Sherman, B. T., Lempicki, R. A. (2009a). *Nature Protoc.*, **4**, 44-57.

Huang, D. W., Sherman, B. T., Lempicki, R. A. (2009b). *Nucleic Acids Res.*, **37**, 1-13.

Zou, H., Hastie, T. (2005). *J. Royal. Statist. Soc. B.*, **67**, 301–320.

Tibshirani, R. (1996). *J. Royal. Statist. Soc. B.*, **58**, 267-288.

TRANSTAGING: Transcriptogram-based staging of cancer

Jose Luiz Rybarczyk-Filho, Marcio Luis Acencio, and Ney Lemke

Departamento de Física e Biofísica, Instituto de Biociências de Botucatu, Unesp - Univ Estadual Paulista

The classification of different tumor types is most important in cancer diagnosis. The cancer classification studies are clinical based and have restricted diagnostic ability. Cancer classification using gene expression data is known to contain the keys for addressing the central problems relating to cancer diagnosis and drug discovery. Analysis of genome-wide expression data poses a challenge to extract relevant evidence. We use computational method that order genes on a line and clusters genes by the probability that their products interact. Protein–protein association information can be obtained from large data bases as STRING. The genome organization obtained this way is independent from specific experiments, and defines functional modules that are associated with gene ontology terms. The starting point is a gene list and a matrix specifying interactions. Considering the Homo sapiens genome, we projected on the ordering gene expression, producing plots of transcription levels for three different tumor types (lung, neuroblastome, breast), whose data are available at Gene Expression Omnibus database. This analysis differentiated normal and tumor tissues. Moreover, the subdivision of the tumor tissues in many classes that were previously inspected with biological process ontologies (Gene Ontology) shown that each class has a set of modified process. This result is the first evidence to find biomarkers for tumor staging by a computational method.

Assess Genomic Biomarkers of Toxicity in Drug Development

Wei Zhang, Scott Emrich and Erliang Zeng*
Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA
* **E-mail: ezeng@nd.edu**

1 Introduction

Typical risk assessment strategies use animal models and *in vitro* experiments as surrogates for human studies in the early stages of drug development. Toxicity assessment is then conducted using conventional indicators such as pathology and clinical chemistry data. Although these methods are widely used, around 40% of drug-induced liver injury (DILI) cases are not detected in the preclinical studies using these conventional indicators, and agreement between studies on animal models and human clinical trials is often poor. To overcome these issues, advances in modern "-omics" including high-throughput microarrays and next-generation sequencing technologies have allowed using genomic biomarkers in risk assessment. The underlying hypothesis is that genomic biomarkers will be more sensitive than conventional markers in detecting toxicity signals.

In this paper, we predicted DILI based on microarray data sets provided by the Japanese toxicogenomics project [1], a CAMDA 2013 challenge. We first explored the possibility of replacing the animal model with in vitro assay coupled with toxicogenomics. Previous studies addressed this problem using the agreement of differentially expressed gene lists from *in vivo* and *in vitro* data, and found poor agreement between the two [2]. Pessiot *et al.* then proposed to evaluate the *in vivo-in vitro* agreement using Gene Set Enrichment Analysis (GSEA) on collapsed probesets [2], which has shown success in improving such agreement. Here we took an alternative approach. Instead of comparing features (for example, differentially expressed genes) resulting from *in vivo* and *in vitro* experiments, we evaluated biological consequences such as pathological measurements and observed DILI by comparing the power of gene features to predict these consequences. The underlying hypothesis was that the processes that cause pathology and DILI effects are complicated and may involve many factors; and although *in vivo* and *in vitro* data sets may share many common characteristics, they could also capture different biological information. Because drug toxicity could result from perturbations of biological metabolic pathways, these effects could happen at any level and could be induced by several key players. We also explored the possibility of predicting the DILI potential in humans using the *in vitro* data from rat primary hepatocytes or human primary hepatocytes. Our method focused on the analysis of resulting pathological data and therefore could provide a more fair comparison using downstream effects as our key. Because the pathological data is available for *in vivo* assays only, we assumed that drug will cause similar pathological

results in *in vitro* experiments as it does for *in vivo* assays. For DILI data, we assumed a drug has similar liver injury effect in humans and rats for predicting liver injury using toxicogenomics data from animals.

2 Materials and Methods

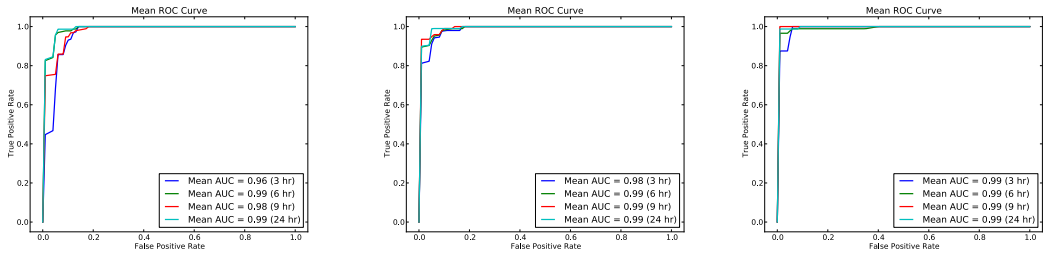
We explored the possibility of predicting the DILI potential in humans using Japanese toxicogenomics project data, which provided close to 20,000 pre-processed Affymetrix microarrays used to measure the effects of 131 drugs on the liver [1]. These included rat *in vivo* data with two experiment designs (single and repeated) and *in vitro* data of both rat and human (using rat and human hepatocytes). Various drug dose levels and sacrifice time after treatments were applied in the experiment design. We used FARMS-summarized and collapsed gene expression values as described previously [3] in modeling and analysis for this paper.

Among the 131 drugs, 101 were associated with one of the following categories: "Most DILI concern", "less DILI concern" and "no DILI concern". We considered DILI prediction as a binary classification problem. Unlike previous work [3] that used two classes of "Most DILI" against "Less DILI" or "No DILI", we used the control microarray data as one class and the microarray data from " Most DILI concern" and "Less DILI concern" drugs as the other class because we found it is difficult to differentiate between these two labels. Each microarray data is represented by 12,088 genes (rats) or 18,988 genes (humans).

CAMDA challenge also provided a total of 5569 summaries of the rat liver pathology reports as previously described [1], which makes supervised training possible to predict pathology. For each pathology finding, we ignore severity and create a dataset for that finding and use binary classification model to classify it. As in previous work [4], only the five most frequent pathology findings, for which the largest data sets were available, were evaluated: hypertrophy, necrosis, cellular infiltration, microgranuloma and cellular change.

As to the classification model used we applied Random Forest (RF) [5], which is an ensemble approach based on the aggregation of a set of decision trees, where each tree is grown from a bootstrap sample (sampling with replacement) of the original data. The average over all of the predictions from the individual trees was considered as the final predicted value. In addition to achieving competing prediction accuracy compared to the state-of-the-art machine learning methods, Random Forest also could cope with high dimensional data and has good model interpretability while incorporating variable selection inside the learning process.

Source code and additional results from the analysis are available at: `https://bitbucket.org/davidzhang/camda2013`.



(a) Low dose level (b) Middle dose level (c) High dose level

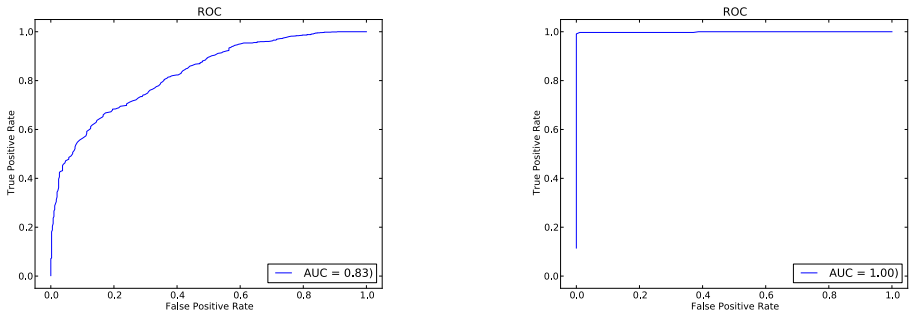
Figure 1: ROC curves in classifying DILI using Rat *in vitro* data

3 Results

We first explored influences of different experiment designs (dose and sacrifice time) on the prediction of DILI potential. Figure 1 demonstrates ROC curves for the classification of DILI categories using rat *in vitro* data sets of different combinations of dose level (low, middle and high) and time point (3, 6, 9 and 24 hour after treatment). The ROC curves are averaged results using 5-fold cross validation. Although various combinations of dose and time points have different sets of differentially expressed genes, they all have good and similar discriminative power in classifying samples as damaged and non-damaged. This observation suggests that time information and dose level are not critical factors in assessing drug toxicity in these data, which is consistent with the prior findings of Pessiot *et al.* [2].

3.1 Comparisons between rat *in vivo* and *in vitro* studies

To evaluate the possibility of replacing an animal *in vivo* study with *in vitro* assay, we built a classifier on available rat *in vitro* data and then used it to predict rat *in vivo* data. Figure 2(a) demonstrates the ROC curve of DILI classification using 1,000 trees in the Random Forest. The result is promising (AUC=0.83). On the other hand, the RF model built on *in vivo* data can perfectly predict DILI potential of *in vitro* assay (AUC=1.00, results not shown). In addition to providing prediction, RF also calculates variable importance (VIM) for each feature when constructing the model. We then compared two gene lists: one list include genes in *in vivo* study with VIMs not larger than 0 (referred to as *in vivo* gene list), and the other list contains genes from *in vitro* assay whose VIMs are not larger than 0 (referred to as *in vitro* gene list). These two gene lists were obtained according to RF VIMs based on RF models constructed on *in vivo* and *in vitro* data sets separately. Figure 4 shows the *Venn* diagram of the comparison. Among 5,845 total important genes, only 1575 (26.88%) genes are shared by two lists, indicating poor agreement when comparison is performed between gene features. Nevertheless, using



(a) Rat *in vitro* classify rat *in vivo*

(b) Rat *in vivo* classify human *in vitro*

Figure 2: ROC curves of rat *in vivo* DILI classification using RF model built on rat *in vitro* data (a) and of human *in vitro* DILI classification using RF model built on rat *in vivo* data (b)

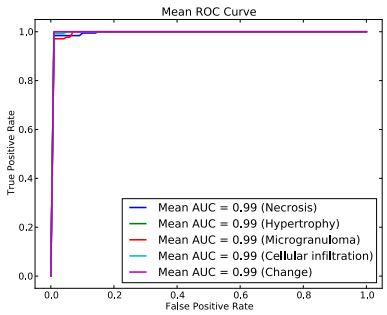
DILI classification as an interpretation of important genes from the two lists is a better approach to compare rat *in vivo* and *in vitro* data sets. Since DILI is derived eventually from pathology and clinical chemistry data, thus we expect classifier using RF model for predicting pathological data also preform good. Figure 3 demonstrates five leading pathologies classification using RF model built on rat *in vivo* data (Figure 3(a)) and rat *in vitro* data (Figure 3(b)) .

3.2 Comparisons between human *in vitro* and rat *in vitro* data

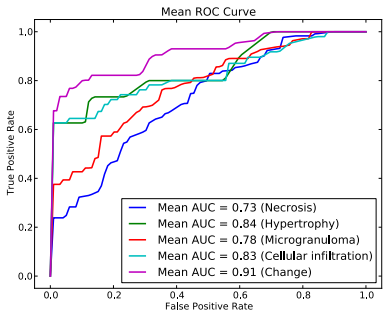
The results as shown in Figure 2(a) demonstrate that replace animal model with *in vitro* assay is possible. Furthermore, we attempted to predict DILI in humans using rat toxicogenomics data. We created an ortholog gene mapping between human genes and rat genes according to their corresponding probe-set common gene names and obtained a list of 9,947 pairs of ortholog genes. A RF classification model was constructed using rat *in vitro* data and such model was then used to predict DILI potential of human *in vitro* gene expression data. The result as shown in Figure 2(b) demonstrates that we could accurately classify human DILI (AUC=1.0) using model built on rat *in vitro* data, which implies it is possible to predict the liver injury in humans using toxicogenomics data from animal *in vitro* assays.

References

1. Uehara T, Ono A, Maruyama T, Kato I, Yamada H, et al. (2010) The japanese toxicogenomics project: application of toxicogenomics. Molecular nutrition & food



(a) Using RF model built on rat *in vivo* data



(b) Using RF model built on rat *in vitro* data

Figure 3: ROC curves for classifying five pathologies

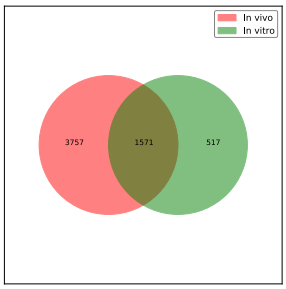


Figure 4: Venn diagram of common differentially expressed genes in *in vivo* and *in vitro* microarray

research 54: 218–227.

2. Pessiotm JF, Wong PS, Maruyama T, Morioka R, Aburatani S, et al. (2013) The impact of collapsing data on microarray analysis and dili prediction. Systems Biomedicine 1: 1–7.

3. Pessiot JF, Wong PS, Maruyama T, Morioka R, Aburatani S, et al. (2013) The impact of collapsing data on microarray analysis and dili prediction. Systems Biomedicine 1: 1–7.

4. Shigetta R, Bowles M (2013) Statistical models for predicting liver toxicity from genomic data. Systems Biomedicine 1: 1–6.

5. Breiman L (2001) Random forests. Machine learning 45: 5–32.

Analyzing the Japanese Toxicogenomics Project Dataset with SVM and RLS Classifiers

Jari Björne, Antti Airola, Tapio Pahikkala and Tapio Salakoski
Department of Information Technology, University of Turku
Turku Centre for Computer Science (TUCS)
Joukahaisenkatu 3-5, 20520 Turku, Finland
firstname.lastname@utu.fi

1 Introduction

The TGP dataset from the Japanese Toxicogenomics Project concerns the response of rats and rat and human *in vitro* cell cultures to a number of drugs [1]. In the CAMDA 2013 challenge this dataset is utilized for analyzing drug-induced liver injury (DILI). Questions include the evaluation of the dataset to determine whether the animal model can be replaced with an *in vitro* cell culture, and whether DILI can be predicted using toxicogenomics data from animals. Both pathology data and microarray genomic expression data are provided in the challenge. We approach these questions as a machine learning task, evaluating the dataset in the context of SVM and RLS classifiers and in defining an experimental setup for automated prediction of DILI.

2 Dataset and Methods

We use the FARMS normalized version of the CAMDA dataset, intended to overcome observed cell culture effects [2]. The dataset consists of a large number of experiments, but in light of the proposed experimental question, predicting the liver injury potential of a drug, there are only 101 distinct examples, each example representing a single drug with a human DILI-concern rating, with features potentially combined from several *in vivo* or *in vitro* experiments. Of these drugs, 8 are in the "no DILI concern", 52 in the "less DILI concern" and 41 in the "most DILI concern" categories.

A *per-drug* example dataset in LibSVM format is provided as part of the CAMDA challenge, for the task of classifying drugs into "no DILI concern" vs. "most DILI concern". With only 8 examples in the "no DILI concern" class, if e.g. 10-fold cross validation were applied to the dataset, each subset would contain on average just a single example of this class, leading to potentially unstable results. We note that Pessiot et. al. [3] performed classification experiments using binary classification into "no or less DILI concern" vs. "most DILI concern", an experimental setup resulting in a more balanced class distribution.

In defining our experimental setup our primary aim was to formulate a question that would result in a larger dataset, potentially producing more reliable results. Therefore, we defined as our

per-individual experiment whether the individual animal or cell culture in a single experiment had been treated with a drug of "no or less DILI concern" or "most DILI concern". This setup is of course very close to classification on the level of drugs, but allows us to explore the classification potential of the individual variation between experiments, and provides us with a larger set of examples. To maximize available data we also combined single and repeated dose rat *in vivo* experiments. In preliminary classification studies, we observed that models trained on high drug dose experiments had best performance, and that 9 hr, 24 hr and 29 day time points for the rat *in vivo* data, as well as 8 hr for the human and 2 hr for the rat *in vitro* data had best performance. Selecting these experiments for further study, we produced datasets with "most"/"no or less" examples at 80/160 for human *in vitro*, 82/120 for rat *in vitro* and 205/215 for rat *in vivo* experiments.

There are of course strong implied dependencies between individual experiments with a single drug, between not only replicates, but potentially also time points and doses. To avoid information leaks, when selecting examples for training and testing, we always put all examples treated with the same drug into either the training or the testing set.

2.1 Features

We defined a number of feature groups to be used for classifying the data. *Pathology features* are the pathology, hematology, biochemistry and liver weight data, available for the *in vivo* rat experiments. *Array features* are the FARMS-processed, non-collapsed microarray expression values available for all experiment types. We also experimented with using *INI scaling*, multiplying the expression values with their reliability estimates ($value * (1 - INI)$) [4].

In addition to these basic features, we also explore refining the dataset with additional data on tissue specificity of gene expression. We retrieve from UniGene[1] known tissues of expression for both rat and human genes. For each tissue-specific group of expression values we define a set of statistical features (minimum, maximum, mean, median and variance) intended to give an overview of expression values. Alternatively, we also select as array features and the tissue-specific statistics only the subset of genes known to be expressed in the *liver*, based on UniGene data.

2.2 Machine learning approach

We apply two state-of-the-art machine learning algorithms: the support vector machine (SVM) and the regularized least-squares (RLS) method, also popularly known as least-squares SVM, or ridge regression [5]. The methods are closely related, and have in numerous experimental comparisons been shown to have quite similar performance. A specific advantage for RLS is the existence of efficient computational short cuts for computing cross-validation estimates. These are especially useful in the considered setting, since due to the small sample size, a central challenge for the evaluation is how to do parameter selection, and at the same time obtain a reliable estimate of the predictive performance. For RLS learning and cross-validation algorithms, we use the implementations in the RLScore[2] software package.

For the initial SVM experiments we applied the SVM^{multiclass} support vector machine[3] [6]

[1] `http://www.ncbi.nlm.nih.gov/unigene`

[2] `http://www.tucs.fi/RLScore/`

[3] `http://svmlight.joachims.org/svm_multiclass.html`

with a linear kernel. Experiments with 10-fold cross-validation proved very unstable, likely due to the small number of examples in each subset, so we adopted a 5-fold cross-validation approach, where two fifths were used to train the classifier, two fifth's to optimize the parameters (opt set) and one fifth to evaluate performance (test set). All examples for the same drug were always placed in the same fold.

To maximize the use of the available data we took advantage of the cross-validation capabilities of the RLScore package. Following the recommendations of [7] we apply a leave-pair-out cross-validation scheme, defined as follows:

$$\frac{1}{|I_+||I_-|}\sum_{i\in I_+}\sum_{j\in I_-} H(f_{\overline{\{i,j\}}}(x_i) - f_{\overline{\{i,j\}}}(x_j)),$$

where $f_{\overline{\{i,j\}}}$ denotes a classifier trained with the whole data set except the i-th and j-th training examples, and $I_+ \subset I$ and $I_- \subset I$ denote the indices of the positive and negative instances in the whole data set Z, respectively. We enumerate all the drug-pair combinations, and on each round of cross-validation leave as test examples all data points corresponding to these two drug pairs. The setup guarantees that we have no information leak between training and test data, since all data points corresponding to same drug are always in the same fold. Further, as shown by [7], the method makes maximal use of the available data, producing an almost unbiased estimate of the AUC, with lower variance than alternative approaches. We perform nested cross-validation, with an inner leave-pair-out loop used for parameter selection, and an outer one for performance estimation.

3 Results and Discussion

In Table 1 we present RLS leave-pair-out classification results for the preprocessed *per-drug* TGP data prepared by the CAMDA organizers. We notice considerable variance in the results: while the best performance achieved on high-dose level at 24 h is 0.71, the lowest one is 0.23 AUC at 2 h on low dosage, which is much worse than a random classifier would be expected to perform (0.5). Therefore we consider it unclear how much predictive power the learned models really have, or if the detected patterns are just due to random chance.

In Table 2 are shown the results for our *per-individual* approach to the CAMDA dataset, testing both SVM and RLS classifiers. We again notice considerable variance on the results. While the RLS cross-validation presents the most efficient way of utilizing the available data for training, the 5-fold SVM cross-validation should result in relatively similar results for truly reliable predictions. We notice the two experiments provide similar results mostly on the rat *in vivo* data, where performance is also the highest. The rat and human *in vitro* datasets show considerably lower performance, with human results slightly more promising. In our experimental setup, the use of INI values did not have much impact on performance. The direct use of the pathology data as features resulted in very unstable models for the *in vivo* data. We note that the UniGene tissue specific expression statistics show some potential on the *in vitro* datasets, achieving on occasion relatively high performance with a much smaller number of features, but due to the variance of the dataset, these observations should be considered highly speculative. Overall, the use of the *in vivo* expression data as features resulted in the most stable models.

Table 1: RLS nested leave-pair-out cross-validation results on preprocessed TGP per-drug data.

Dose	AUC (2 h)	AUC (8 h)	AUC (24 h)	AUC (all timepoints)
All				0.60
Low	0.23	0.48	0.57	
Middle	0.61	0.61	0.63	
High	0.46	0.49	0.71	

Table 2: RLS nested leave-pair-out cross-validation and SVM 5-fold cross-validation results (parameter optimization set and test set) on per-individual TGP data. Results over AUC 0.6 are shown in bold and under AUC 0.5 in italics.

Species	Feature Groups	Features	SVM(opt)	SVM(test)	RLS
human *in vitro*	INI, array	9011	**0.59 ± 0.06**	**0.53 ± 0.07**	0.54
human *in vitro*	INI, array, unigene	9479	**0.60 ± 0.07**	*0.50 ± 0.09*	*0.40*
human *in vitro*	INI, array, unigene(liver)	8049	**0.60 ± 0.06**	**0.52 ± 0.09**	0.53
human *in vitro*	INI, unigene	471	**0.55 ± 0.06**	**0.53 ± 0.07**	0.57
human *in vitro*	array	18980	**0.60 ± 0.07**	*0.49 ± 0.07*	0.54
human *in vitro*	array, unigene	19448	**0.60 ± 0.06**	*0.48 ± 0.08*	*0.40*
human *in vitro*	array, unigene(liver)	12093	**0.60 ± 0.06**	*0.49 ± 0.07*	0.53
human *in vitro*	unigene	471	*0.54 ± 0.05*	*0.52 ± 0.05*	**0.65**
human *in vitro*	unigene(liver)	15	*0.53 ± 0.05*	*0.49 ± 0.01*	0.53
rat *in vitro*	INI, array	7950	0.54 ± 0.03	*0.53 ± 0.06*	0.55
rat *in vitro*	INI, array, unigene	8130	0.54 ± 0.03	*0.52 ± 0.05*	0.53
rat *in vitro*	INI, array, unigene(liver)	4752	0.56 ± 0.03	*0.51 ± 0.06*	0.51
rat *in vitro*	INI, unigene	183	0.55 ± 0.02	*0.53 ± 0.06*	0.56
rat *in vitro*	array	12080	0.54 ± 0.03	*0.53 ± 0.04*	0.55
rat *in vitro*	array, unigene	12260	0.54 ± 0.03	**0.54 ± 0.06**	0.51
rat *in vitro*	array, unigene(liver)	5533	0.56 ± 0.03	*0.52 ± 0.06*	0.51
rat *in vitro*	unigene	183	0.55 ± 0.02	*0.54 ± 0.05*	0.58
rat *in vitro*	unigene(liver)	9	0.50 ± 0.00	0.50 ± 0.00	*0.36*
rat *in vivo*	INI, array	6753	**0.60 ± 0.06**	**0.58 ± 0.04**	**0.61**
rat *in vivo*	INI, array, unigene	6933	**0.58 ± 0.03**	**0.58 ± 0.08**	**0.61**
rat *in vivo*	INI, array, unigene(liver)	4299	**0.57 ± 0.03**	0.55 ± 0.03	**0.60**
rat *in vivo*	INI, unigene	187	**0.59 ± 0.03**	*0.51 ± 0.05*	0.55
rat *in vivo*	array	12084	**0.60 ± 0.06**	**0.61 ± 0.09**	**0.60**
rat *in vivo*	array, pathology	12189	**0.62 ± 0.11**	*0.37 ± 0.11*	*0.49*
rat *in vivo*	array, unigene	12264	**0.59 ± 0.04**	**0.59 ± 0.06**	**0.61**
rat *in vivo*	array, unigene(liver)	5537	**0.58 ± 0.03**	0.56 ± 0.03	**0.60**
rat *in vivo*	pathology	112	**0.62 ± 0.10**	*0.42 ± 0.04*	*0.41*
rat *in vivo*	unigene	187	**0.59 ± 0.03**	*0.50 ± 0.03*	0.55
rat *in vivo*	unigene(liver)	13	0.52 ± 0.01	*0.48 ± 0.02*	0.50

4 Conclusions

We find it notable that mostly the largest drug doses produced data that could be classified the best, possibly indicating that the DILI-related gene expression response is rather faint, pointing to the need for an experimental setup strong enough to produce unambiguous data.

Our Python-based experimental software is built on publicly available tools, depending only on open source classifiers. We will also provide all of our code under an open source license, hopefully useful for further research on the topic.

In testing various feature representations, we observed potential value on refining the expression data with external databases such as UniGene. However, most importantly, performing a large set of experiments with somewhat related feature representations and different classifiers highlighted the disturbingly large variance in classification performance. In understanding the potential of the TGP dataset for building predictive models we therefore consider it highly important that all experimental results are carefully compared and evaluated.

References

[1] T. Uehara, A. Ono, T. Maruyama, I. Kato, H. Yamada, Y. Ohno, and T. Urushidani, "The Japanese toxicogenomics project: Application of toxicogenomics," *Molecular Nutrition Food Research*, vol. 54, no. 2, pp. 218–227, 2010.

[2] S. Hochreiter, D.-A. Clevert, and K. Obermayer, "A new summarization method for affymetrix probe level data," *Bioinformatics*, vol. 22, no. 8, pp. 943–949, 2006.

[3] J.-F. Pessiot, P. S. Wong, T. Maruyama, R. Morioka, S. Aburatani, M. Tanaka, and W. Fujibuchi, "The impact of collapsing data on microarray analysis and DILI prediction," *Systems Biomedicine*, vol. 1, no. 3, pp. 1–7, 2013.

[4] W. Talloen, D.-A. Clevert, S. Hochreiter, D. Amaratunga, L. Bijnens, S. Kass, and H. W. Ghlmann, "I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data," *Bioinformatics*, vol. 23, no. 21, pp. 2897–2902, 2007.

[5] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances in Computational Mathematics*, vol. 13, pp. 1–50, April 2000.

[6] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research (JMLR)*, vol. 6(Sep), pp. 1453–1484, 2005.

[7] A. Airola, T. Pahikkala, W. Waegeman, B. De Baets, and T. Salakoski, "An experimental comparison of cross-validation techniques for estimating the area under the ROC curve," *Computational Statistics & Data Analysis*, vol. 55, pp. 1828–1844, April 2011.

Reasonably integrating data for predicting the drug toxicity by machine learning

Naiyang Guan[1], Zhilong Jia[2], Xiang Zhang[1], Bin Luo[1], Bin Mao[2], Qing Liao[3] and Zhigang Luo[1]

[1] National Laboratory for Parallel and Distributed Processing, School of Computer Science, National University of Defense Technology, Changsha, China
[2] Department of Chemistry and Biology, College of Science, National University of Defense Technology, Changsha, China
[3] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China

E-mail: zhilongjia@gmail.com

The toxicogenomics is usually expected to aid in the risk assessment of drugs. Drug-induced liver injury (DILI) is a leading reason of drugs failing during clinical trials as well as being withdrawn from the market. In this paper, we focused on the two problems in prediction of DILI by analyzing the toxicogenomics data provided by CAMDA2013.

Firstly, we predict the DILI using a more reasonable data-collapsing method although with a relatively lower classification performance. Some of previous work consider the data-collapsing but neglect the following objective situation. During predicting the risk of a new drug, the drug toxicity in any doses and any time-point conditions should be blind to us. More specifically, for predictive machine learning models, it is more reasonable to use some drug profiling data for training while test on another drugs profiling data. Thus, we present a new and more reasonable method for collapsing the multi-doses and multi-time-points expression profiling data of microarray. We use the averaged value and maximum value of the differential expression values in all doses and time-points for each drug within each gene with or without trim to construct the two datasets for learning. By using some classification methods including both LDA and linear SVM with ten folds cross-validation, we found the accuracies of predicting DILI are all about 60% using the rat in vivo single dose type profiling data. This unsatisfactory result may be caused by the fact that different drugs may result in DILI in various approaches. In other words, this problem becomes a small sample problem in spite of many features. Therefore, the differential expression genes have litter in common among the profiling of drugs. This result suggests that subtyping the coarse-grained DILI based on the related pathways of differential expression genes may contribute to the ultimate risk assessment of new drugs.

Secondly, we utilize Canonical Correlation Analysis (CCA) to integrate the profiling data from rat in vivo single and in vitro for predicting DILI. A typical use for CCA is to take two sets of variables, e.g., profiling data from rat in vivo single and in vitro, and see what is common subspace, e.g. DILI, amongst the two sets. So, it is suitable to integrate these data via CCA. Using the integrated data, the prediction result is not better than the result from that of rat in vivo. Our result reveals that the agreement between in vivo and in vitro is poor for predicting the DILI. It seems that the noises in each dataset mask the common profiling pattern of DILI or the subtypes of DILI may result in a less common profiling pattern of DILI.

In summary, we present a more reasonable data-integrating method for the classification of IDLI and the poor agreement between in vivo and in vitro for predicting the DILI. It seems that subtyping the DILI may be essential to better assess the risk of new drugs.

DILI classification model based on *in vitro* human transcriptomics and *in vivo* rat clinical chemistry data.

Danyel Jennen, Jan Polman, Mark Bessem, Maarten Coonen, Florian Caiment, Dennie Hebels, Joost van Delft, Jos Kleinjans

Department of Toxicogenomics, Maastricht University, PO Box 616, 6200 MD Maastricht, the Netherlands
The Netherlands Toxicogenomics Centre, Maastricht University, Po Box 616, 6200 MD Maastricht, the Netherlands

Corresponding author: Dr. Danyel Jennen, Universiteitssingel 40, 6227 ER Maastricht, the Netherlands. Phone +31 433883983, fax +31 433884146, email danyel.jennen@maastrichtuniversity.nl

---

The past decades drug induced liver injury (DILI) is the main cause of drugs to fail during clinical trials or to be withdrawn from the market (Chen *et al.* 2011). Approximately 40% of DILI cases are not detected in preclinical studies based on conventional indicators in *in vivo* rodent studies (Zhang *et al.* 2012). Therefore, alternative methods for predicting the DILI potential in humans are needed and toxicogenomics-based approaches have been considered.
Recently, we developed an *in vitro* transcriptomics-based method in the human hepatic cell line HepG2 for predicting *in vivo* genotoxicity, which showed 89% accuracy, thereby clearly outperforming the standard *in vitro* test battery (Magkoufopoulou *et al.* 2012). For the CAMDA challenge an adapted version of this *in vitro* method was used to develop an *in vitro* classification model for predicting DILI in humans.
The development of the *in vitro* classification model for DILI in human consisted of 3 steps:

1. selecting drugs from the three DILI potential groups (i.e. “no DILI”, “less DILI” and “most DILI”) for the training and validation sets;
2. establishing gene signatures between the different DILI potential groups of the training set using a leave-one-out t-test or ANOVA;
3. using these gene signatures to train and validate the prediction model in PAM (prediction analysis for microarrays) (Tibshirani *et al.* 2002).

Selection of drugs
From each DILI potential group, i.e. “no DILI” (ND), “less DILI” (LD) and “most DILI” (MD), drugs were selected based on the *in vivo* clinical chemistry measurements of alkaline phosphatase (ALP), aspartate aminoptansferase (AST), alanine aminotransferase (ALT), lactate dehydrogenase (LDH) and Y-glutamyltranspeptidase (GTP) from rats with a daily repeated treatment. In particular, 20 MD drugs were selected that showed elevated levels for four or five of the measurements. Six ND drugs that showed decreased or unchanged levels were selected. For 35 LD drugs two or three of the measurements showed elevated levels. Dose and time were not taken into consideration in the selection.

The selected drugs were used in different settings resulting in four training sets:

- all selected drugs; MD, LD and ND or total DILI (D) and ND (61 drugs)
- drugs from MD and ND (26 drugs)
- drugs from LD and ND (41 drugs)
- drugs from MD and LD (55 drugs)

The distribution of drugs over the DILI groups for the training and validation set is summarized in Table 1.

Table 1. Distribution of drugs over the DILI groups for the training and validation set.

	training set	validation set	total
MD	20	21	41
LD	35	13	48
ND	6	2	8
total	61	36	97

Gene signatures

Microarray data from human primary hepatocytes exposed to high doses for 24 hours were used to establish gene signatures for each training set of drugs. The expression data were re-annotated to the MBNI Custom CDF-files and RMA normalized using the web tool arrayanalysis.org (Eijssen *et al.* 2013).

Genes with significantly different expression values ($p<0.01$) between the different DILI groups for each training set were selected from the expression data based on a series of statistical tests (ANOVA with three groups and t-test with two groups). For each test the two replicates of one of the drugs were removed (leave-one-out procedure). The significant genes that were present in all tests (the intersection of all lists) were selected for training the prediction model as signature. The resulting five gene signatures lists contained 31 to 141 genes as indicated in Table 2.

Training and validation of prediction models

PAM analysis (Tibshirani *et al.* 2002) was conducted for each of the signature lists for class prediction (threshold: 0). Misclassification errors (ME) were calculated for each prediction model and were highest (0.25) for ANOVA MD-LD-ND. The other four models had a ME <0.1.

Per prediction model the accuracy for each DILI group was calculated as indicated in Table 2. The accuracy within the training set is >90% for all prediction models except ANOVA MD-LD-ND (accuracy 67%-90%). This model also shows lowest accuracy for the validation (<62%). The other four models, MD-ND, LD-ND, MD-LD and D-ND, had a total accuracy for the validation of 87%, 80%, 50% and 89%, respectively.

The MD-ND and LD-ND models were further examined by testing the LD and MD drugs, respectively. This resulted for the LD drugs that 85% were predicted as MD and for the MD drugs that 95% were predicted as LD. This is also in line with the results (accuracy 89%) of the D-ND model. These findings indicate that both MD-ND and LD-ND models can be used for the prediction of DILI. In addition, the gene signature list from the MD-ND, LD-ND and D-ND models share 36 genes (Figure 1).

These genes were examined for GO processes in DAVID (Huang da *et al.* 2009) and were mainly involved in cell cycle, cell growth & proliferation and signal transduction related processes.

Table 2. Accuracy for training and validation sets for each prediction model. The number of signature genes and misclassification errors (ME) are indicated.

ANOVA MD-LD-ND (105 genes; ME 0.25)

	training	**validation**
MD	90%	33%
LD	88%	62%
ND	67%	0%
tot	87%	42%

t-test MD-ND (83 genes; ME 0.038)

	training	**validation**
MD	95%	95%
ND	100%	0%
tot	97%	87%

t-test LD-ND (79 genes; ME 0.024)

	training	**validation**
LD	97%	92%
ND	100%	0%
tot	98%	80%

t-test MD-LD (31 genes; ME 0.091**)**

	training	**validation**
MD	100%	33%
LD	91%	77%
tot	95%	50%

t-test D-ND (141 genes; ME 0.049)

	training	**validation**
D	95%	94%
ND	100%	0%
tot	95%	89%

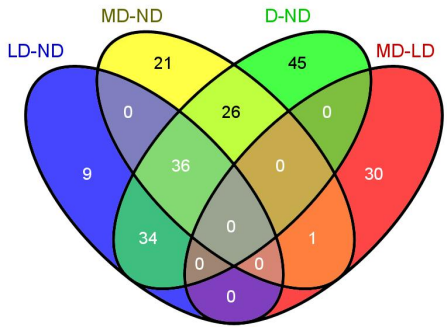


Figure 1. Comparison of the t-test based gene signatures (i.e. number of genes) for the different DILI group combinations.

Conclusions

The results of the *in vitro* human transcriptomics based models are very promising with up to 89% correct prediction for DILI potential. However, it should be noted that the two ND drugs in all validation sets are wrongly predicted and that improvement is definitely needed for distinguishing MD drugs from LD drugs.

Further analyses will be performed in which the following aspects will be considered:

- inclusion of time and dose relationships and/or additional clinical chemical measurement in the selection of drugs for the training set;
- increasing the number of ND drugs from other data repositories;
- performing analysis on transcriptomics data from other time and dose levels;
- enhancing the biological interpretation of gene signature lists.

References

Chen M, Vijay V, Shi Q, Liu Z, Fang H, Tong W (2011) FDA-approved drug labeling for the study of drug-induced liver injury. Drug Discov Today 16 (15-16):697-703

Eijssen LM, Jaillard M, Adriaens ME, Gaj S, de Groot PJ, Muller M, Evelo CT (2013) User-friendly solutions for microarray quality control and pre-processing on ArrayAnalysis.org. Nucleic Acids Res

Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4 (1):44-57

Magkoufopoulou C, Claessen SM, Tsamou M, Jennen DG, Kleinjans JC, van Delft JH (2012) A transcriptomics-based in vitro assay for predicting chemical genotoxicity in vivo. Carcinogenesis 33 (7):1421-1429

Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 99 (10):6567-6572

Zhang M, Chen M, Tong W (2012) Is toxicogenomics a more reliable and sensitive biomarker than conventional indicators from rats to predict drug-induced liver injury in humans? Chem Res Toxicol 25 (1):122-129

Assessing single-nucleotide polymorphism and genotype calling using the KPGP-38 Human Genomes next-generation sequencing data from CAMDA

Wenqian Zhang, Valerii Soika, Jie Shen, Joe Meehan, Zhenqiang Su, Weigong Ge, Hong Fang, Roger Perkins, Vahan Simonyan, Weida Tong, and Huixiao Hong

Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA

Next generation sequencing (NGS) has become the preferred technology in current genetic studies largely because of the potential to identify not only common genetic variants, but also novel and rare variants as well as structural variants. However, reliability of genetic findings based on NGS data relies crucially on accurately calling single-nucleotide polymorphisms (SNPs) and their genotypes. False SNP calls and incorrect genotypes are caused in different ways, including sequencing errors, incorrect base calling, mapping errors, and sampling bias due to insufficient coverage. Analysis methods and quality control metrics to distinguish true SNPs from false variants are urgently needed to realize the full discovery potential from NGS data. The Critical Assessment of Massive Data Analysis (CAMDA) consortium hosts the KPGP-38 Human Genomes NGS data obtained from the Illumina HiSeq 2000 platform with 30x to 40x coverage. Importantly, this dataset's high coverage and the inclusion of two different sets of twins and a Caucasian female provides a suitable opportunity to explore quality control metrics for improving accuracy in SNP and genotype calling and to investigating aspects of population genetics. We first used different pipelines such as SOAP2-SOAPsnp and Hexagon to call SNPs and genotypes as well as structural variants for the KPGP-38 data. Then, twin pairs KPGP88/KPGP89 and KPGP90/KPGP91 were determined to be monozygotic based on SNP and genotype call concordance that was somewhat higher than we've previously observed for technical replicates from the same DNA. Thereafter, discordant SNPs and calls for the pair of twins were presumed to be errors for which causes could be postulated and related to the numerical procedures used. The procedure was then used to define criteria to gauge quality and identify likelihood for a calling error for application in profiling other samples. Interestingly, applying the calling criteria, we identified some SNPs in KPGP-38 that were not identified in the 1000 Genomes Project, suggesting that such metrics could refine population genetics. We will share our methods and results and will further discuss implications of our findings.

Characterization of the Korean Genome

Deepali Jhamb*, Meeta Pradhan*, Premkumar Duraiswamy, Akshay Desai, Mathew J. Palakal**
School of Informatics, Indiana University Purdue University Indianapolis, Indiana, USA, 46202
*First Authors, **Corresponding Author
E-mail Addresses: DJ: djhamb@iupui.edu, MP: mpradhan@iupui.edu, PD: premdura@iupui.edu, AD: akdesai@iupui.edu, MJP: mpalakal@iupui.edu

Introduction

Genome wide association studies have identified a large number of common variants associated with complex diseases but despite such efforts a very small fraction of disease heritability and phenotypic variation has been explained by these studies. Whole genome sequencing (WGS) provides the most comprehensive view of an organism's genetic variation. A major challenge is to understand these variants with respect to their functions and interactions with other variants. In this study, we developed an innovative systems biology pipeline for the analysis of next generation sequencing data to discover functional and pathway modules specific for the Korean population. We used the SNV data from the KPGP-38 Human Genomes dataset to understand the interrelationships between the genes with the rare variants. The methodology developed here can also be easily applied to include other variants identified by WGS such as INDELS, CNV, and SV and also other types of next generation sequencing data. Our analysis identified genes related to neurodegeneration, cancer, and hypertension associated with the KPGP-38 Human Genome.

Methodology

SNV data for 37 Korean samples (one sample for Caucasian female, KPGP10, was excluded from our analysis since we focused on the analysis of Korean samples) was extracted in the vcf format. GATK CombineVariants was used to merge these samples into one vcf file (korean_merge). To compare Korean SNVs with other populations, 1000 genome data (1000G) was downloaded. Rare variants in the Korean population were identified (with an allele frequency (AF) less than three in 1000G and those that were unique in the korean_merge were also included) and further analyzed using systems biology approaches. SNPEff was used to annotate the rare variants. The physical interactions among the rare variant genes were extracted using BioGRID. A novel algorithm was designed to identify significant modules from the rare variant network. To initiate this process, a "seed matrix" was constructed which was based on the number of variants in all the regions of a gene. Due to the low representation, we ignored following regions from our analysis: Intragenic, Synonymous Stop, Non Synonymous Start, Stop Lost and Stop Gained. The variants were further normalized based on the remaining nine regions. The gene with highest normalized score was selected as the seed to proceed for constructing the modules. This normalized score was defined as the NodeWeight. The module was expanded based on the NodeWeight, Gene Ontology biological process $\geq$ 70% and maximum pathway similarity between the seed and the leaf node. We constructed modules for the top 100 seeds from seed matrix based on the above three conditions: The top scoring gene was selected as the seed and the module expansion was performed based on the biological process and pathway match between the connecting nodes. We constructed such modules for the top 100 seeds from the seed matrix. The disease information was overlaid on these modules using the OMIM database.
Below we describe the pseudo code for the identification of seed nodes:

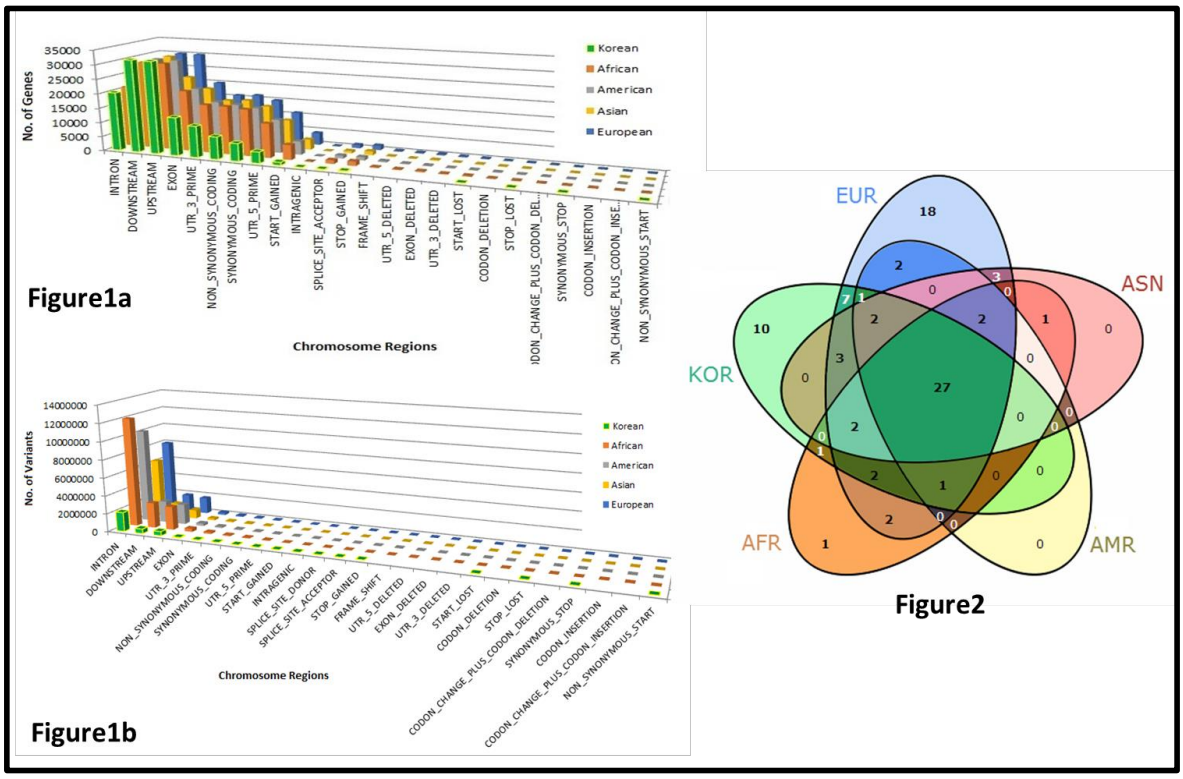
For (i in range (1, total genes))
Normalized i
For (j in range of (1, total genes)

$$score\ j = \sum_{k=1}^{9} region\ of\ k$$

sort j extract top 100 genes

Results

Figure1a and Figure1b show the distribution of genes and rare variants in the Korean vs. other populations respectively. Rare variants in the 1000G data were calculated based on AF<3 for the respective populations. Twenty five different regions were identified in other populations from 1000G data (26 in European) however genes in the Korean population were distributed across 17 regions. Regions such as "Frame shift" and "Exon deleted" were not found in the Korean rare variants. Even though the number of genes identified were similar for the top chromosome regions, they were relatively less for Koreans in rest of the genomic regions. The number of variants identified in the Korean population were significantly lower than the rest of the populations. This could be due to the low sample size of Korean population as compared to the 1000G data.



Biological processes (BPs) for genes of all the populations were extracted using DAVID and the comparison of top 10 biological processes is shown in **Figure2**. A total of ten unique BPs were identified for the Korean population (KOR) however no unique BPs were identified in the Asian (ASN) and American (AMR) population. Eighteen unique BPs were identified for the European (EUR) population and one for the African (AFR) population. Twenty seven BPs were common

to all the populations. Interestingly, the BPs unique to KORs were all related to brain such as cerebellum morphogenesis, and cerebellar cortex development among others. A recent study shows the prevalence of autism in South Korea which is estimated to be 2.6 percent and 1child in every 38 children is affected with this disorder [1]. It has also been previously reported that widely known dysbindin gene, DTNBP1, in schizophrenia is not associated in the Korean population [2]. Our analysis also did not identify this gene.

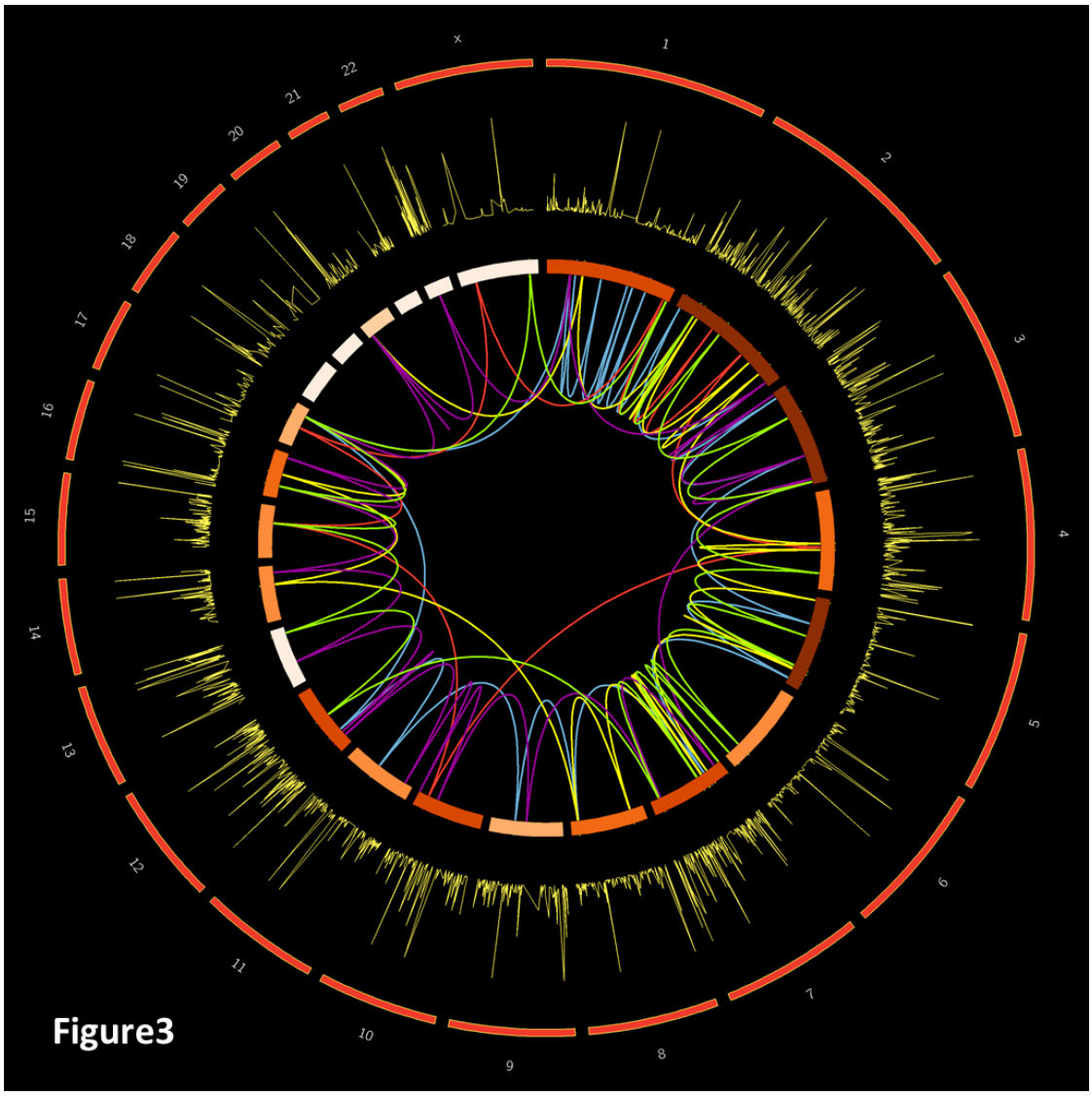


Figure3 depicts the overall distribution of the rare variants and KEGG pathways across the chromosomes. Only genes with greater than 200 variants were included in this analysis. The outermost circle in the figure represents the chromosomes, second circle shows the number of variants for genes in each chromosome, third circle indicates the frequency of pathway distribution (deeper orange color – higher number of variants), and fourth circle in the center shows the interconnected genes from different chromosomes participating in a pathway.

Chromosomes 3, 1, 2, 7, and 5 were found to have the highest number of rare variants in the Korean population. However, chromosome 16 has the maximum no. of genes (57) with variants above average followed by chromosomes 5, 2, 3, and 1. Chromosomes 2, 3, 5, 1, and 7 have the highest no. of pathways. Top ten pathways of these genes were plotted in the center of this image. It can be seen that several genes from different chromosomes interact together to participate in pathways. Neuroactive ligand-receptor interaction (colored blue in **Figure3**), and axon guidance (colored red) pathways have genes distributed across 12 chromosomes each. It has been indicated in the literature that genes associated to these chromosomes are correlated with neurodegenerative diseases in Koreans as well as other populations [3, 4, and 5].

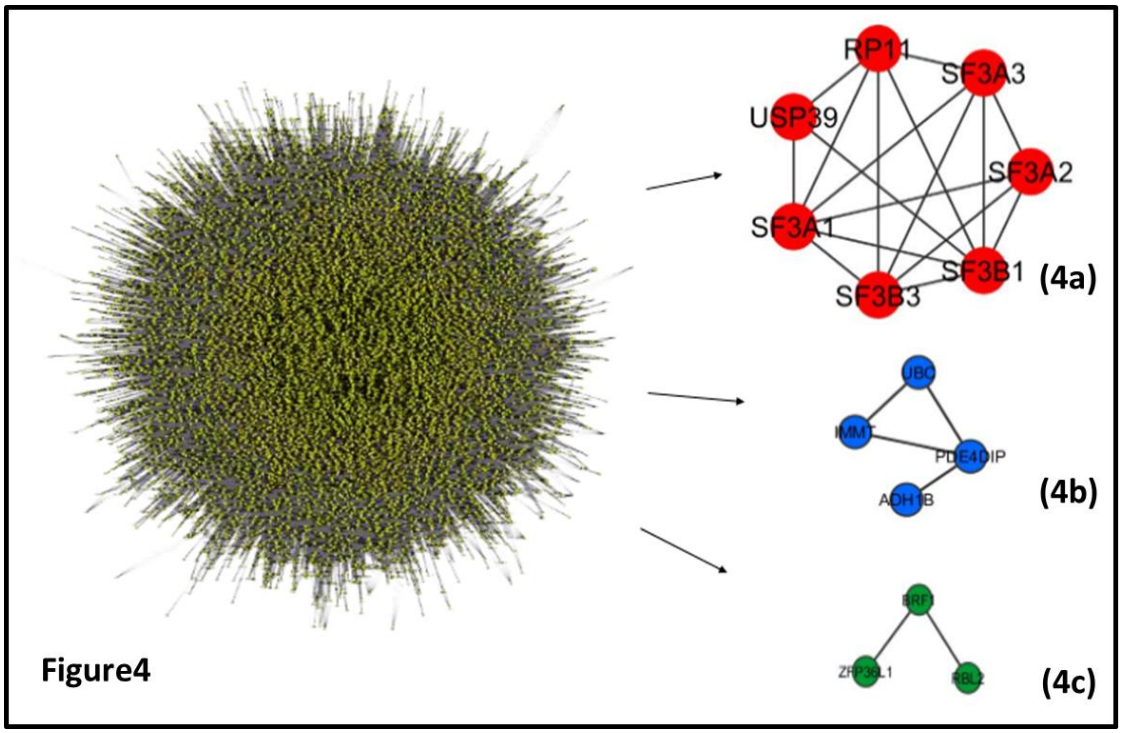


Figure4

Systems biology analysis was further performed to identify significant genes and functional modules which are present in the Korean population. Functional modules were extracted as described in the methodology section. **Figure4** represents the network and few important modules of different sizes. We obtained modules of size 3-9 and these were ranked based on their node property and edge property. The genes in the modules were analyzed for their association with any disease or a disease reported in literature associated with Korean population. One of the most important modules was RP11 module (**Figure 4a**): RP11 SF3A1 SF3A3 SF3A2 USP39 SF3B3 SF3B1 PRPF3. A novel mutation in the RP11 gene is known to cause retinitis pigmentosa in the Chinese population [6], SF3A1 is known to be associated with myleodysplasia and retinoblastoma associated binding protein [7], SF3A2 with CNV aberrations in Korean contribute to AML [8], USP9 with gastric cancer [9], SF3B3 Gastric cancer cell [10], SF3B1 with tumors [11]. This analysis shows that the genes in this module are all correlated with tumor. Another top scored module identified in this work consisted of PDE4DIP, IMMT, UBC ADH1B genes (**Figure 4b**). Detailed analysis of genes in this module associated PDE4DIP with

psychiatric disorders [12], IMMT with spinocerebellar ataxia [13], UBC with hungtinton disease [14], ADH1B as ethnic variant in Asian population [15], correlating this module with neurodegenerative diseases. In addition, genes associated with the height of Korean population and hypertensions were also found in these highly ranked modules. Several other modules which were identified with the GO biological processes unique to Koreans (**Figure2**) were also analyzed and found to correlate to the brain or neuro diseases. Overall it was observed that modules extracted in our study using the system's biology approach could be classified into two major domains: (i) Neurodegenerative diseases and (ii) Tumor related genes. This observation suggests that the Korean population might be more susceptible to these two major classes of complex diseases.

Conclusions

In this work, we identified rare variants unique to the Korean population which were mainly distributed across chromosomes 1, 2, 3, 5, and 7. The novel systems biology approach developed here identified modules whose genes were reported in the literature to be correlated as markers of neurodegenerative diseases and tumor. Systems biology can help elucidate the major variants in a population. The power of systems biology should be exploited to reduce the "big data" generated by next generation sequencing studies and identify significant variants.

Acknowledgements

We would like to acknowledge National Center for Genome Analysis Support, Indiana University, for providing the supercomputing support for the data analysis. We would also like to acknowledge Sammed Mandape for annotating the genes with PharmGKB knowledge base and Shruti Sakhare for literature analysis.

References

1. http://www.autismspeaks.org/about-us/press-releases/new-study-reveals-autism-prevalence-south-korea-estimated-be-26-or-1-38-chil.
2. Joo, E.Y. et. al., Neurosci lett. 2006 Oct 23; 407(2):101-6. Epub 2006 Sep 7.
3. Shin, J. H. et. al., Chest. 2005 Oct; 128(4):2999-3003.
4. Lee, H.R. et. al., Cancer Genet Cytogenet. 2010 Dec; 203(2):193-202.
5. Genetics Home reference (http://medline.gov/)
6. Xia et al., Mol.Vis. 2004, May 20; 10:361-5
7. Wang et. al., J. Proteome Res. 2009 Oct;8(10):4428-40
8. Yun H.J. et. al., Bio Chem. 2008 Oct; 389(10):1313-8
9. Yang, S. et. al., Genomics 2007 Apr;89(4):451-9.
10. Kang, H.C. et. ,al. Clin. Cancer Res. 2004 Jan 1:10(1 Pt 1):272-84
11. Eun Mi Je et al., IJC 2013, Vol. 133, Issue no. 1, pages 260-265.
12. Kim, S. et. al., Transl Psychiatry 2012 May; 2(5)e113.
13. Lee, L.C. et. al., Clin Chim Acta. 2009 Feb; 400(1-2):56-62.
14. Bett, J.S. et. al., J Cell Mol. Med. 2009 Aug;13(8B):2645-57
15. Li, H. et. al., PLoS One. 2008 Apr 2;3(4):e1881

Application of next-generation genome and transcriptome based methods for the exploration of secondary metabolites from marine fungi for the treatment of cancer

Abhishek Kumar and Frank Kempken

Department of Genetics & Molecular Biology in Botany, Institute of Botany,

Christian-Albrechts-University at Kiel, Kiel, Germany

Email: akumar@bot.uni-kiel.de | fkempken@bot.uni-kiel.de

Fungi of marine origin are potent groups of secondary metabolite producers. However, they are not well characterized and underutilised in terms of biotechnological applications. We aim for sustainable exploration of marine fungal isolates and their encoding natural products as drugs against cancer under the EU-funded project marine fungi (www.marinefungi.eu). Besides isolation of new fungal strains from unique marine habitats, the molecular development of effective producer strains is in the focus. Genomes of selected candidate strains originating from our unique strain collection of marine fungi are currently characterized with respect to secondary metabolite production.

Next-generation sequencing (NGS) techniques have changed the facets of genomics and its application. We have established the genomic sequences from three marine isolates, *Scopulariopsis brevicaulis, Pestalotiopsis* sp. and *Calcarisporium* sp. by the use of different next-generation sequencing methods (Roche 454, Illumina and ion-torrent).

We report on different properties of genome assemblies and annotations for these fungi. Several gene families and superfamilies have been analyzed to explore genetic peculiarities of these species along with repeats and transposable element contents. The assembled genome of *Scopulariopsis brevicaulis* is ~32 Mb in size with N50 equals to 88 kb and 935 contigs containing 16298 genes with average intron length equals to 129.4. During the annotation process, we were able to annotate 9340 genes (57.31 %) while 6958 genes (43.69 %) remained non-annotated in *Scopulariopsis brevicaulis* genome. 17 genes encoding for non-ribosomal peptide synthetases (NRPSs), 18 polyketide synthases (PKSs) and one gene encoding a hybrid NRPS-PKS were found. Similarly, the genome size for *Pestalotiopsis sp.* is ~46 Mb with N50 equals to 71.9 kb and 4186 contigs containing 23492 genes, which is surprisingly high for an ascomycete. The average intron length and the average intron per gene are 126.8 and 2.2, respectively. During annotation process, we annotated 60% genes of *Pestalotiopsis* genome with 44 NRPSs, 62 PKSs and 7 hybrid NRPS-PKS genes. The assembled genome size of *Calcariosporium sp.* is about 35 Mb genome with N50 equals to 91.9 kb and 2464 contigs containing 15459 genes. The percentage GC% for this genome is 50.7%. The average intron length and the average intron per gene are 121 and 2.1, respectively. During annotation process, we annotated 72% genes, while 28% genes remained non-annotated for *Calcariosporium* genome with 52 NRPSs, 66 PKSs and 7 hybrid NRPS-PKS genes.

Predicted genes are presently in process of validation using illumina based RNA-seq. We are also comparing wild type phenotypes with higher-yielding mutants of these fungi with special interest on specific natural compounds.