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).

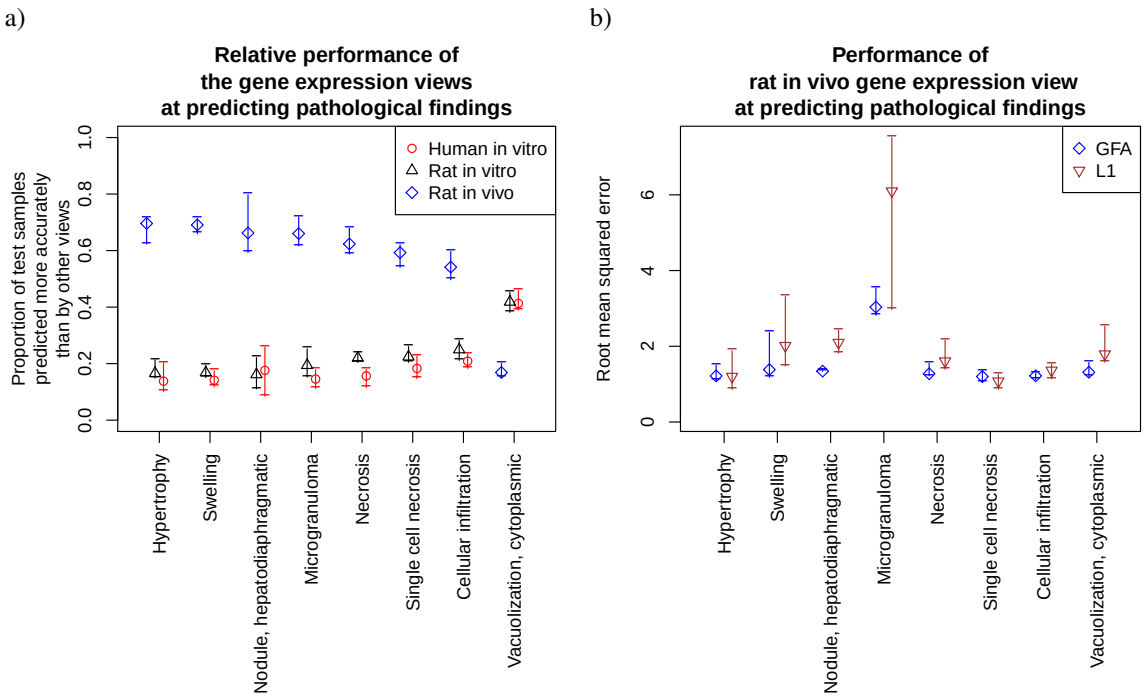

Figure 1: GFA-predictor based on gene expression of rat *in vivo* samples yields a superior prediction on pathological findings in the test data compared to gene expression of the *in vitro* samples (left). The absolute performance of GFA is comparable to or better than the performance of the ℓ_1-regularized 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.

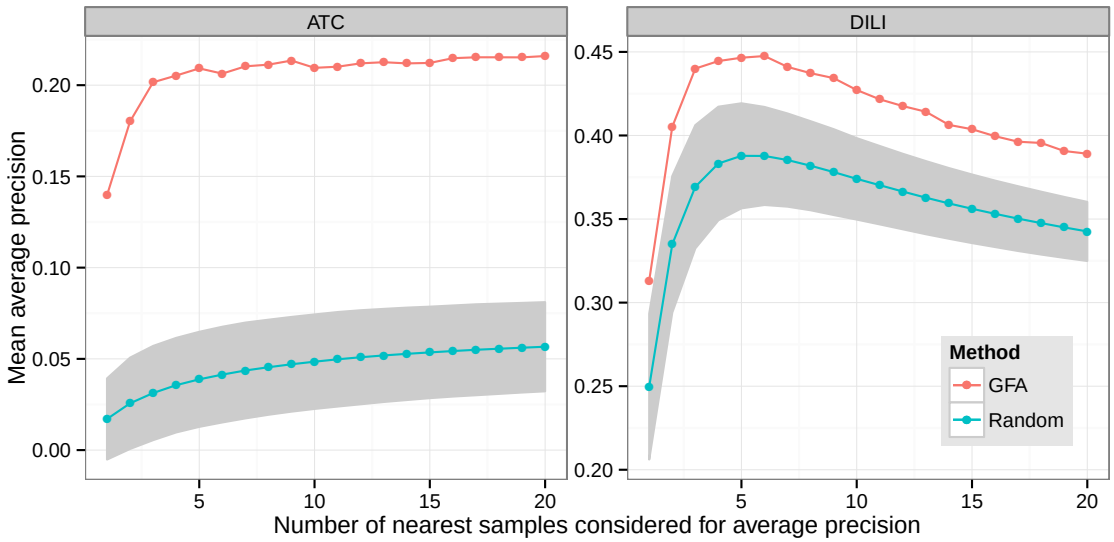

Figure 2: Similar drug compounds are significantly aggregated in the latent space in terms of both the ATC codes and DILI labels (left and right, respectively). The aggregation is quantified as mean average precision score of the retrieval of similar compounds in the latent space of GFA. The retrieval 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.