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

Tommi Suvitaival

Juuso A. Parkkinen Seppo Virtanen Samuel Kaski

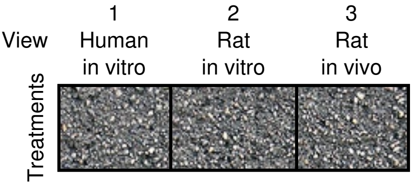
HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

Aalto University School of Science

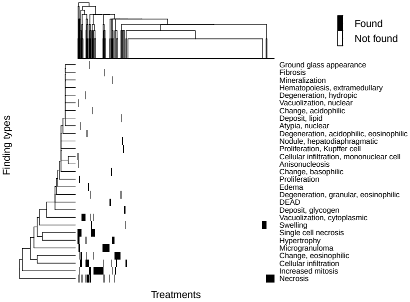
July 19-20, 2013 – CAMDA

Starting point

High-dimensional gene-expression data from 3 types of organisms

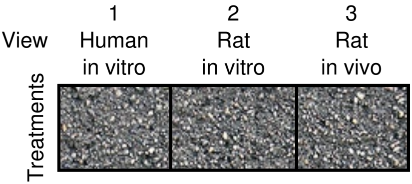


Sparse pathological data of rat *in vivo*

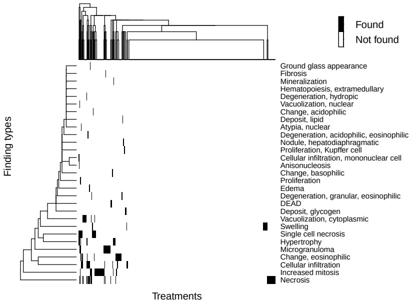


Starting point

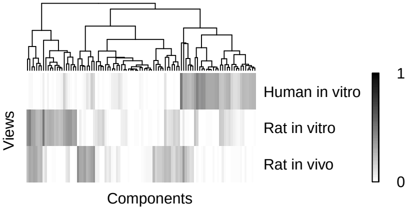
High-dimensional gene-expression data from 3 types of organisms



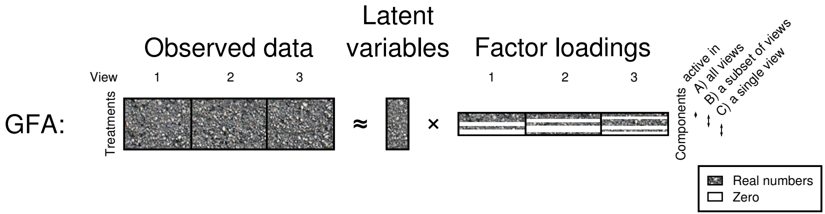
Sparse pathological data of rat *in vivo*



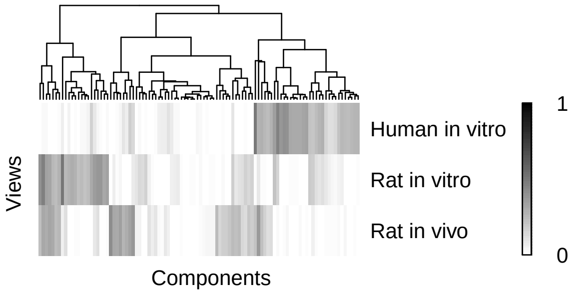
1. Can we replace the animal study with *in vitro* assay?
2. Can we predict the liver injury in humans using toxicogenomics data from animals?



Group factor analysis (GFA)



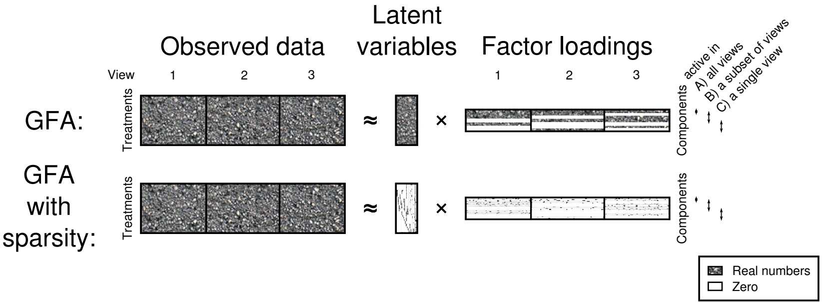
Making generalizations across organisms



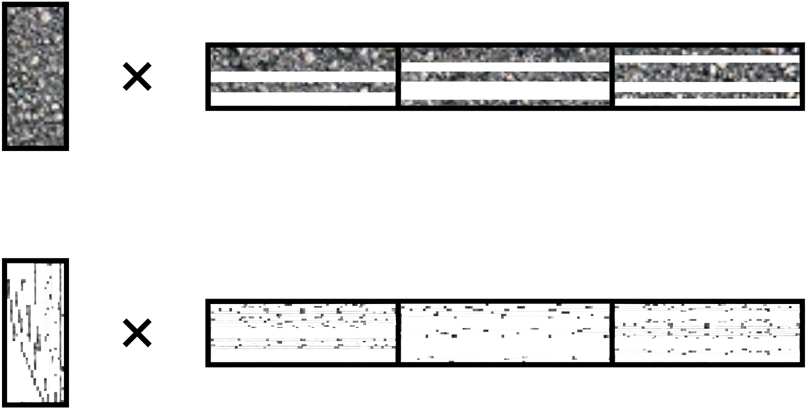
Shared components

- associations between views
- cross-view prediction

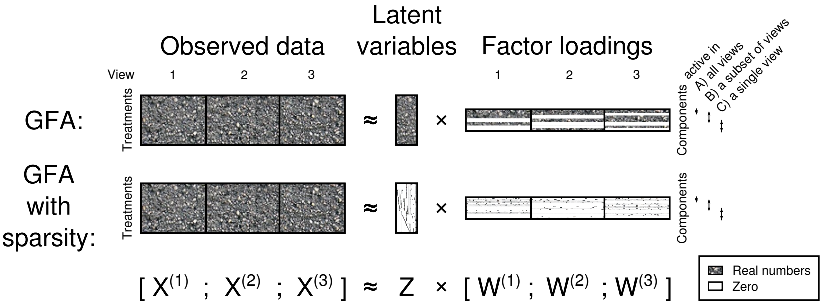
GFA with sparsity (1)



GFA with and without sparsity



GFA with sparsity (2)



$$[\, X^{(1)} \;;\; X^{(2)} \;;\; X^{(3)} \,] \approx Z \times [\, W^{(1)} \;;\; W^{(2)} \;;\; W^{(3)} \,]$$

Sparsity – why

Sparsity in the model is encouraged due to

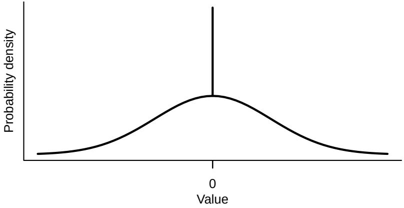
1. High dimensionality of the gene expression microarray data sets 2. Strong sparsity of the pathology data	$\Rightarrow$	Sparsity in terms of variables
3. Treatments heterogeneous by their effects	$\Rightarrow$	Sparsity in terms of samples

Sparsity – how

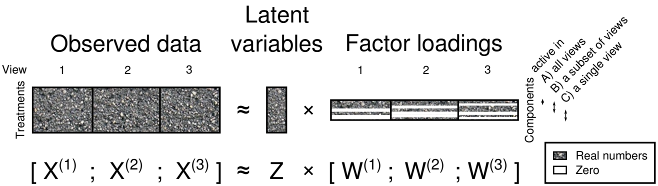
1. Sparsity in terms of variables $\Rightarrow$ *Spike-and-slab prior** for factor loadings matrix $\mathbf{W}$

2. Sparsity in terms of samples $\Rightarrow$ Spike-and-slab prior for latent variables $\mathbf{Z}$

*



GFA – model



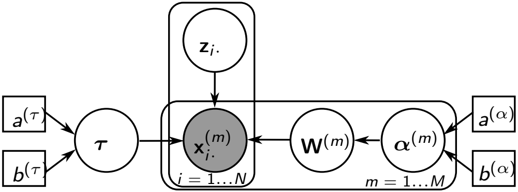
$$\mathbf{x}_{i\cdot}^{(m)} \sim \mathcal{N}\left(\mathbf{z}_{i\cdot}\mathbf{W}^{(m)}, \tau_m^{-1}\mathbf{I}\right)$$

$$\mathbf{z}_{i\cdot} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)$$

$$\mathbf{w}_{k\cdot}^{(m)} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\boldsymbol{\alpha}_k^{(m)}}\mathbf{I}\right)$$

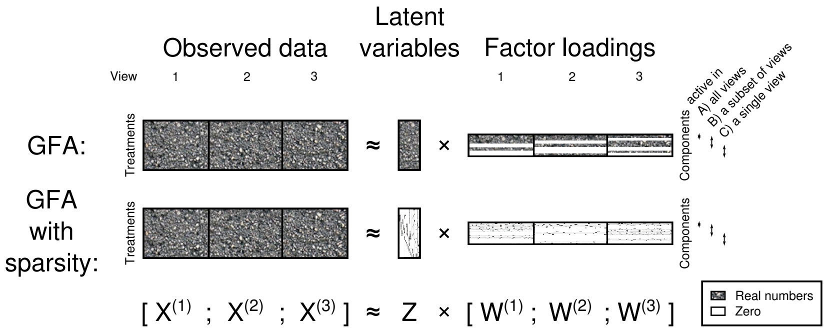
$$\boldsymbol{\alpha}_k^{(m)} \sim Gamma\left(a^{(\alpha)}, b^{(\alpha)}\right)$$

$$\tau_m \sim Gamma\left(a^{(\tau)}, b^{(\tau)}\right)$$



i: samples, m: views

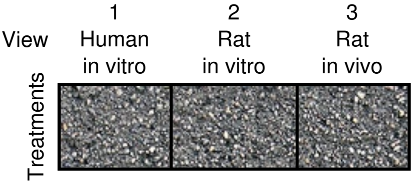
GFA with sparsity – model



GFA	GFA with sparsity
$\mathbf{x}_{i\cdot}^{(m)} \sim \mathcal{N}\left(\mathbf{z}_{i\cdot}\mathbf{W}^{(m)}, \tau_m^{-1}\mathbf{I}\right)$	$\mathbf{x}_{i\cdot}^{(m)} \sim \mathcal{N}\left(\mathbf{z}_{i\cdot}\mathbf{W}^{(m)}, \left(\mathbf{\Lambda}^{(m)}\right)^{-1}\right)$
$\mathbf{z}_{i\cdot} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$	$\mathbf{z}_{ik} \sim \mathbf{H}_k^{(z)}\mathcal{N}\left(0, \frac{1}{\boldsymbol{\alpha}_{ik}^{(z)}}\right) + \left(1 - \mathbf{H}_k^{(z)}\right)\delta_0$
$\mathbf{w}_{k\cdot}^{(m)} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\boldsymbol{\alpha}_k^{(m)}}\mathbf{I}\right)$	$\mathbf{W}_{dk}^{(m)} \sim \mathbf{H}_{dk}^{(m)}\mathcal{N}\left(0, \frac{1}{\boldsymbol{\alpha}_{dk}^{(m)}}\right) + \left(1 - \mathbf{H}_{dk}^{(m)}\right)\delta_0$

Data representation – gene expression

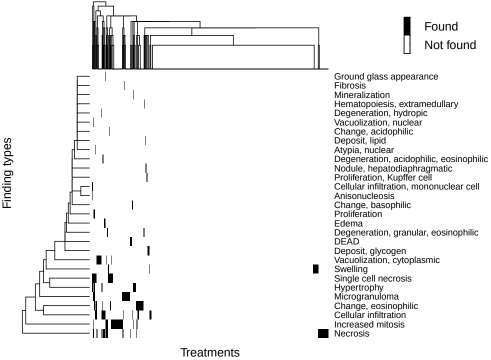
- Treatments that occur in all 3 types of organism:
  - 119 compounds
  - dosage levels *middle* & *high*
  - time points *8/9 h* & *24 h*
- Average differential expression over the replicates of each treatment
  - ⇒ Treatment = sample for the model
  - ⇒ Matching treatments between the 3 transcriptomic *views* $\mathbf{X}^{\text{human}}_{\text{in vitro}}$, $\mathbf{X}^{\text{rat}}_{\text{in vitro}}$ and $\mathbf{X}^{\text{rat}}_{\text{in vivo}}$



Data representation – histopathology of the liver

Grade-weighted count of each pathological finding type over the replicates of a treatment

$\Rightarrow$ Pathology view $\mathbf{Y}^{\text{rat}}_{\text{in vivo}}$ with matching treatments to the 3 transcriptomic views



Results

Our tasks:

1. Predict liver damage of rats *in vivo* based on cell-level transcriptomic responses in the 3 types of model organisms
2. Test how well the transcriptomic cell-level responses generalize to known effects of the compounds on humans

Analysis: model organisms' generalizability to organ level

Training: Learn associations between the views

- 3 transcriptomic views $\mathbf{X}^{\text{human}}_{\text{in vitro}}$, $\mathbf{X}^{\text{rat}}_{\text{in vitro}}$ and $\mathbf{X}^{\text{rat}}_{\text{in vivo}}$
- Pathology view $\mathbf{Y}^{\text{rat}}_{\text{in vivo}}$

Testing: Predict the pathological findings $\mathbf{Y}^{\text{rat}}_{\text{in vivo}}$

- Given one of the transcriptomic views

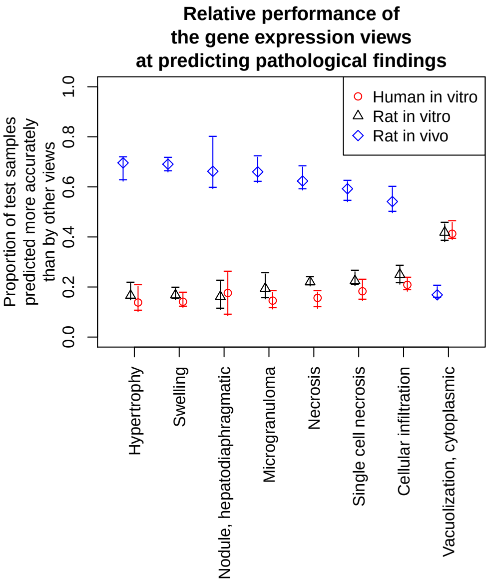
Analysis: model organisms' generalizability to organ level

Training: Learn associations between the views

- 3 transcriptomic views $\mathbf{X}^{\text{human}}_{\text{in vitro}}$, $\mathbf{X}^{\text{rat}}_{\text{in vitro}}$ and $\mathbf{X}^{\text{rat}}_{\text{in vivo}}$
- Pathology view $\mathbf{Y}^{\text{rat}}_{\text{in vivo}}$

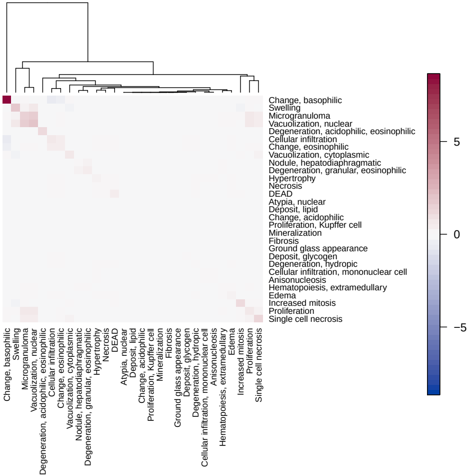
Testing: Predict the pathological findings $\mathbf{Y}^{\text{rat}}_{\text{in vivo}}$

- Given one of the transcriptomic views



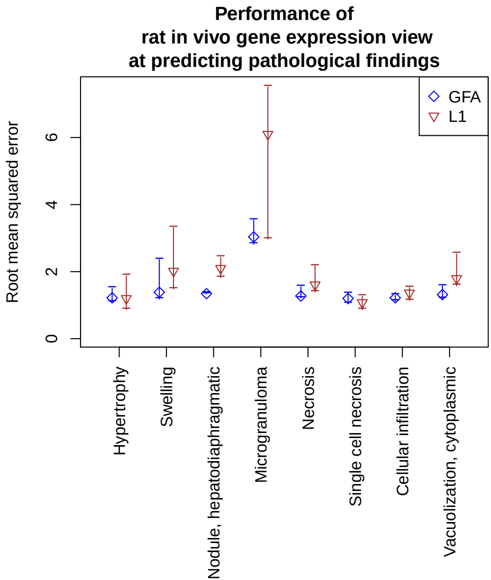
Sparsity in the target view

- $\mathbf{W}^T\mathbf{W}$ reveals the similarity of component activities between the variables
- Thanks to sparsity, projections to many variables are 0
- The model automatically decides which variables to explain by
  A. coherent components
  B. noise parameter



Prediction: drug hepatotoxicity based on gene expression

- Given $\mathbf{X}^{\text{rat}}_{\text{in vivo}}$, predict $\mathbf{Y}^{\text{rat}}_{\text{in vivo}}$
- Same prediction task using ℓ_1-regularized linear regression



Translation over model organisms to humans

- How do the transcriptional changes in model organisms generalize system-level effects in humans?
- Can the model learn structure relevant to the properties of the compounds in an unsupervised way?

Translation over model organisms to humans (1)

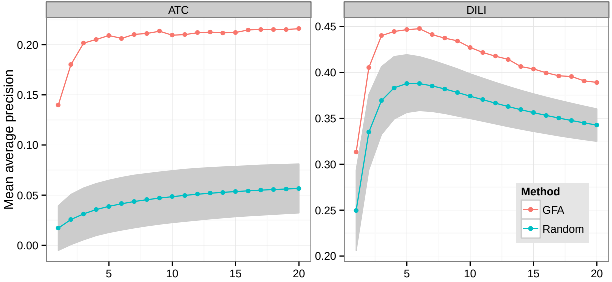
We quantify the success of translation by the retrieval of similar compounds

- Ground-truth:
  A. Anatomical Therapeutic Chemical (ATC) Classification System's labels (level 4)
  B. Drug-induced liver injury (DILI) labels
- Model: GFA with sparsity for the transcriptomic views of the model organisms, $\mathbf{X}^{\text{human}}_{\text{in vitro}}$, $\mathbf{X}^{\text{rat}}_{\text{in vitro}}$ and $\mathbf{X}^{\text{rat}}_{\text{in vivo}}$
- Measure: Average precision in the retrieval of similar compounds in the latent space

Translation over model organisms to humans (2)

We quantify the success of translation by the retrieval of similar compounds

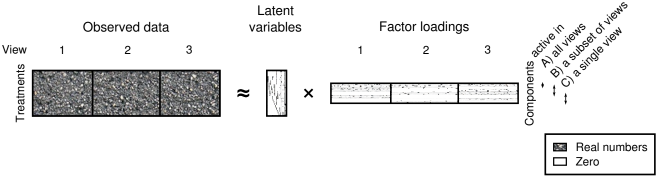
- Ground-truth:
  - A. Anatomical Therapeutic Chemical (ATC) Classification System's labels (level 4)
  - B. Drug-induced liver injury (DILI) labels
- Model: GFA with sparsity for the transcriptomic views of the model organisms, $\mathbf{X}^{\text{human}}_{\text{in vitro}}$, $\mathbf{X}^{\text{rat}}_{\text{in vitro}}$ and $\mathbf{X}^{\text{rat}}_{\text{in vivo}}$
- Measure: Average precision in the retrieval of similar compounds in the latent space



Size of the neighborhood for retrieval

Conclusions

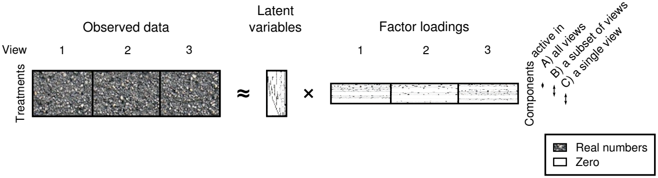
- GFA reveals associations between the views
- Associations indicate what generalizes between the views
- Sparsity helps in this decision
- Latent representation allows us to explore structure in the data in an unsupervised way



Discussion

We can

- analyse the similarity of model organisms
- learn what generalizes from the model organisms to humans



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
- Helsinki Doctoral Programme in Computer Science