Multivariate Data Analysis and Knowledge Extraction from Biological Data

Ali Mohammad-Djafari
Jérôme Lapuyade, Ghazaleh Khodabandelou

Groupe Problèmes Inverses
Laboratoire des Signaux et Systèmes
UMR 8506 CNRS - SUPELEC - Univ Paris Sud 11
Supélec, Plateau de Moulon, 91192 Gif-sur-Yvette, FRANCE.

djafari@lss.supelec.fr
http://djafari.free.fr
http://www.lss.supelec.fr

Annual meeting of C5Sys_ERASysBio, Mars 30-31, 2011

Content

- Project and collaborations
- Context and Data
- Time series and their visualization
- Gene expression data

- Different Questions to Answer:
  - Dimensionality Reduction and Factor Analysis
  - Discriminent Analysis and Classification
  - Graph of links between variables
  - Directed Graph of links between variables

- Questions and Discussion

Project and collaboration

- C5Sys of ERASysBio+ : Circadian and cell cycle clock systems
- Coordinator: Francis Lévi: INSERM
- Partner 2: Gilbertus Van der Horst,
  ERASMUS, The Netherlands
- Partner 3: David Whitmore, Ucollege, United Kingdom
- Partner 4: Franck Delaunay, CNRS, France
- Partner 5: Ali Mohammad-Djafari,
  L2S, CNRS-SUPELEC-UNIV PARIS SUD, France
- Partner 6: Jean Clairambault, INRIA, France
- Partner 7: David Rand,
  Warwick Systems Biology Centre (WSBC), United Kingdom

Context and Data

- **Genes expressions Time series data in two organs:**

  Colon:

Clock:	Rev, Per2, Bmal1
Metabolism:	CE2, Top1, UGT, DBP
CC:	Wee1, Ccna2, Ccnb2
Apoptose:	Bcl2, Mdm2, Bax, P53

  Liver:

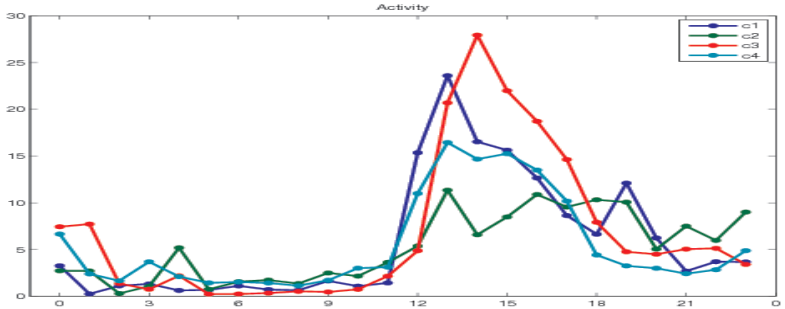
Clock:	Rev, Per2, Bmal1
Metabolism:	CE2, Top1, UGT, DBP
CC:	Wee1, P21
Apoptose:	Bcl2, Mdm2, Bax, P53

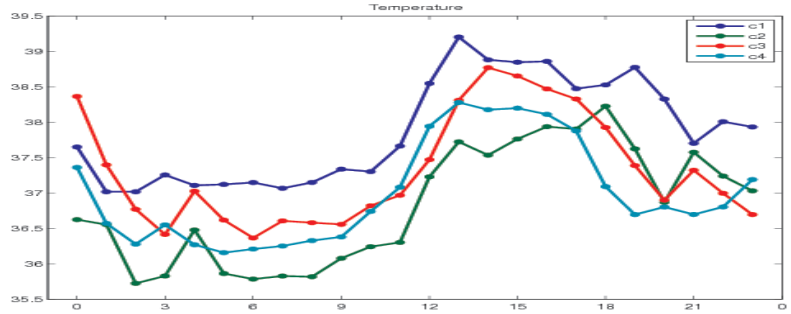
- **Physiological Time series data:**

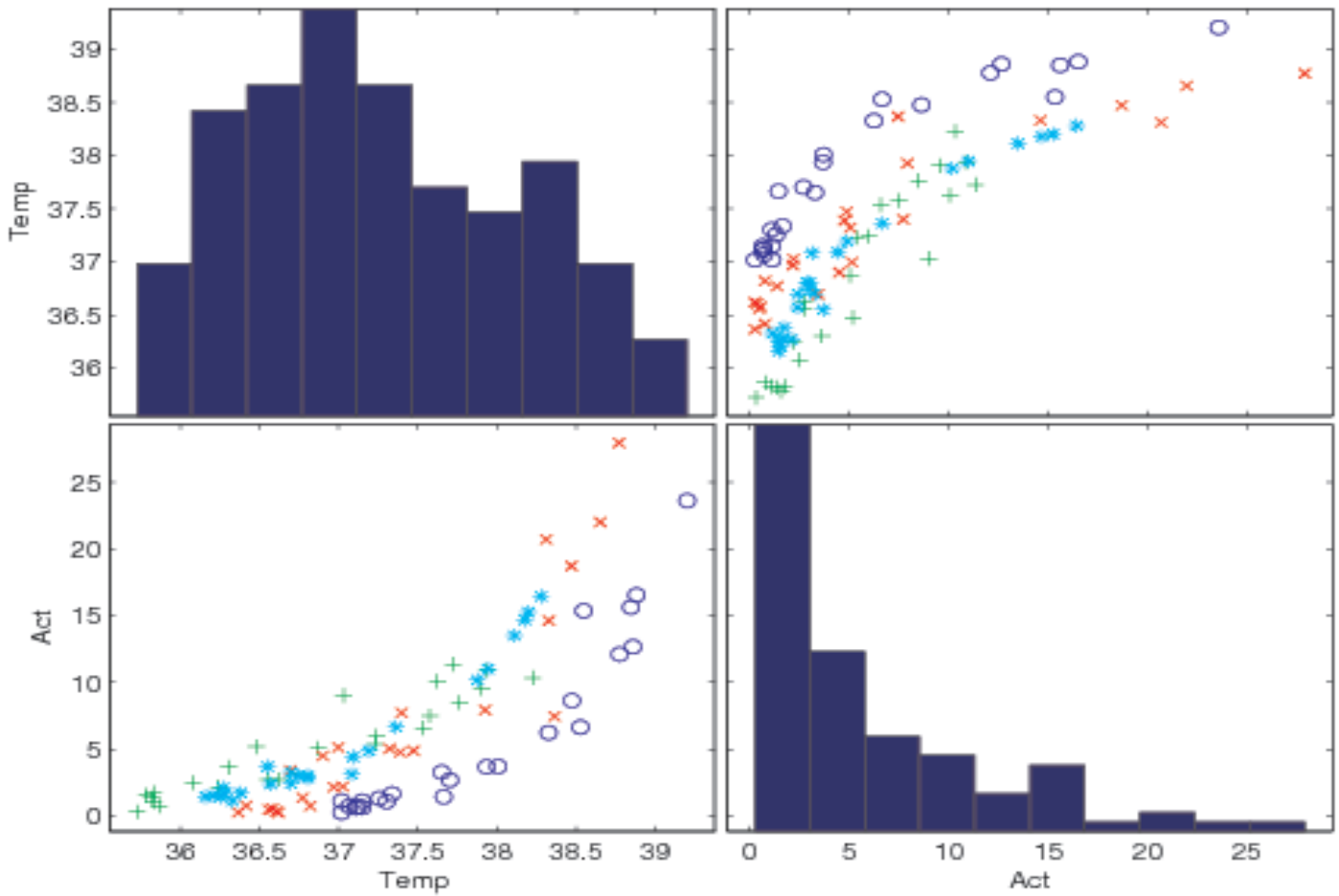
  Temperature, Activity

  Hormons: Cortico, Melato

Time series and their visualization: Ex: Temp/Act







Genes expression time series

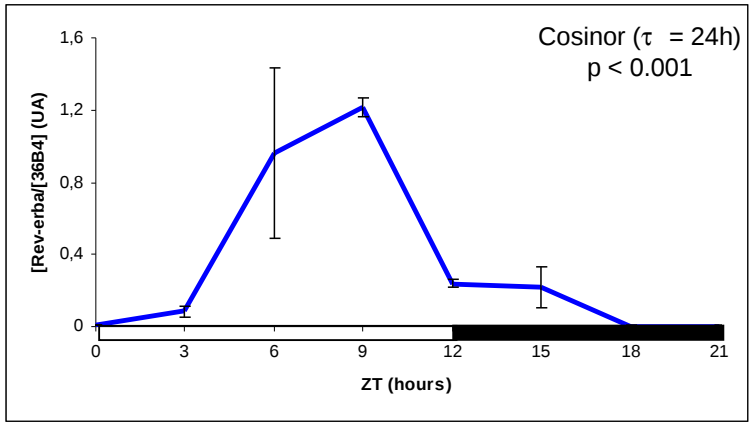
- B6D2F1 male mice synchronized with LD 12:12 for 3 weeks.
- Liver and colon mucosa sampled every 3 h for 24 h
- Circadian expression of mRNA quantified with RT-PCR:
  - Clock genes: Reverb α, Per2 and Bmal1 (liver & colon).
  - Cell cycle genes: Wee1, P21 (liver).
    Wee1, Ccna2 and Ccnab2 (colon).
  - Apoptosis genes: Bcl2, Mdm2, Bax and P53 (liver & colon).

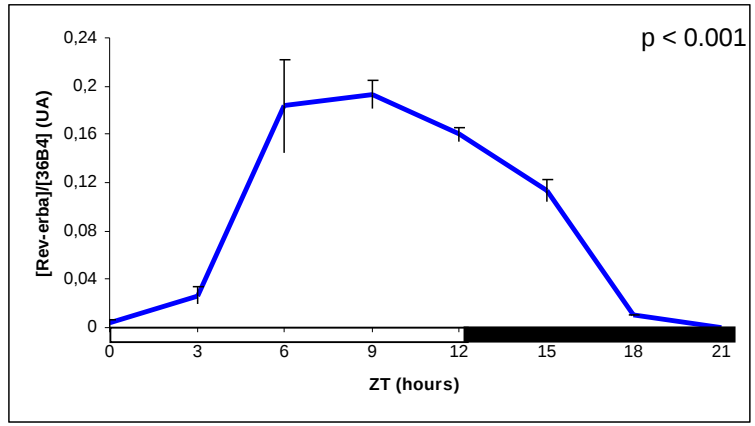
Clock

Liver

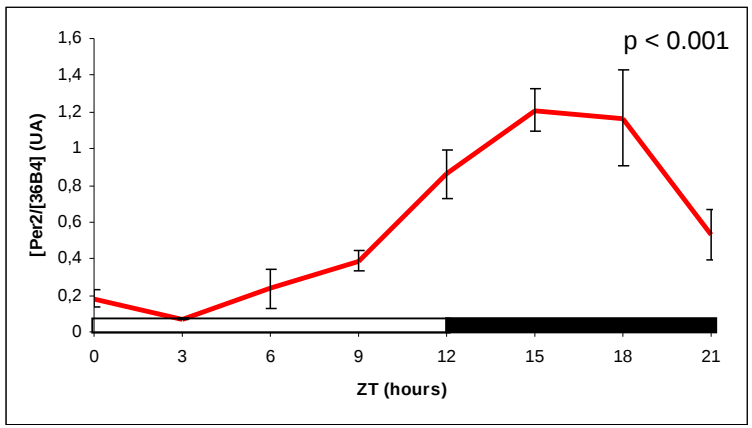
Colon

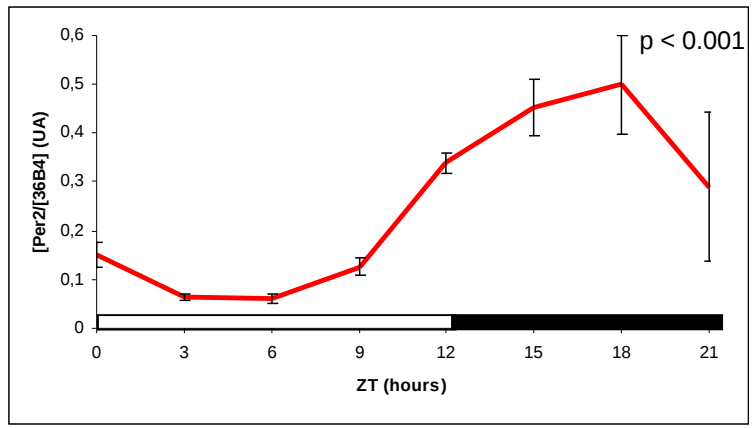
Rev-erbα



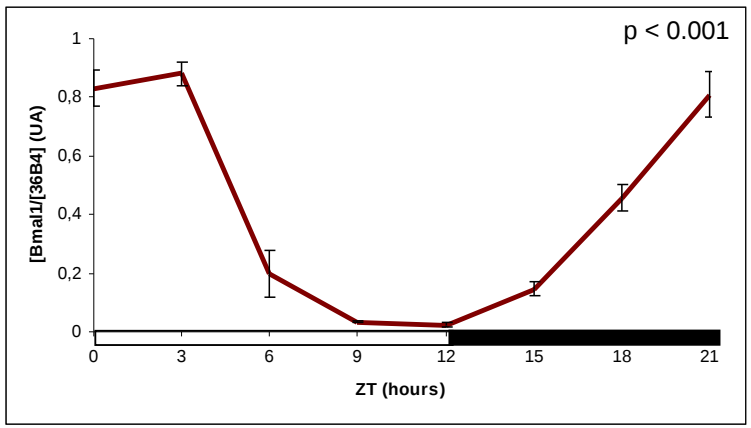


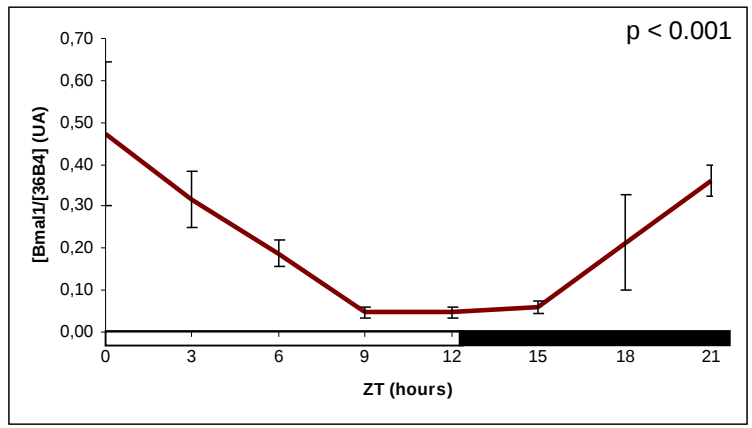
Per2





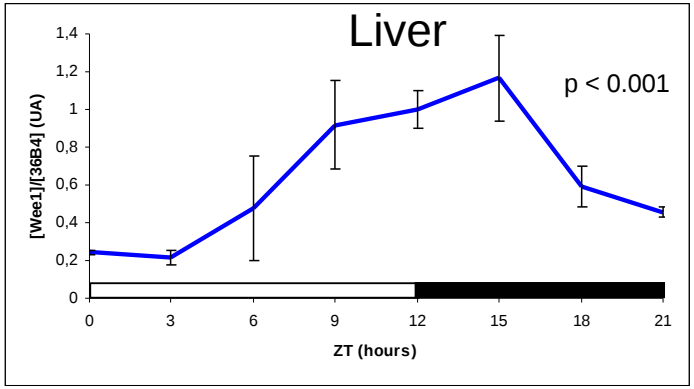
Bmal1

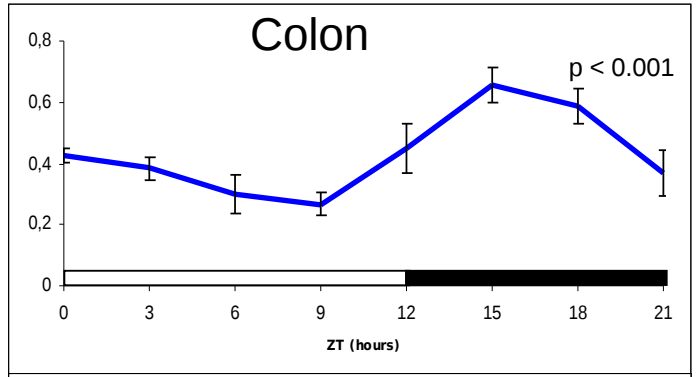




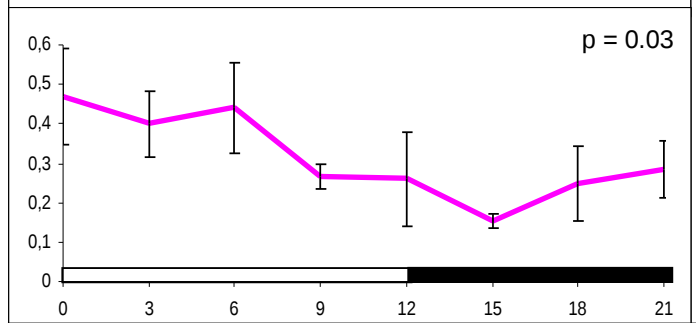
Cell Cycle

Wee1

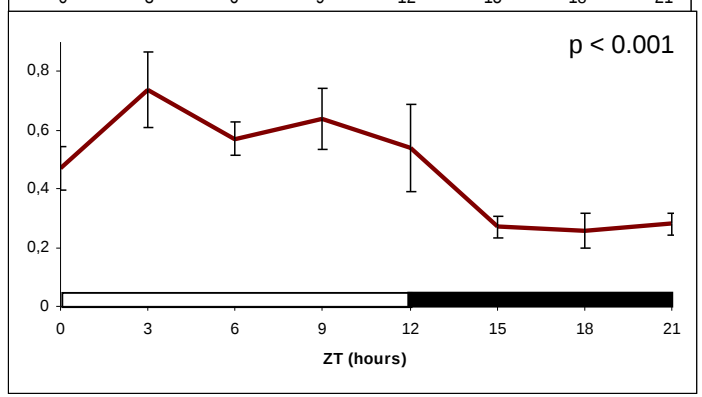




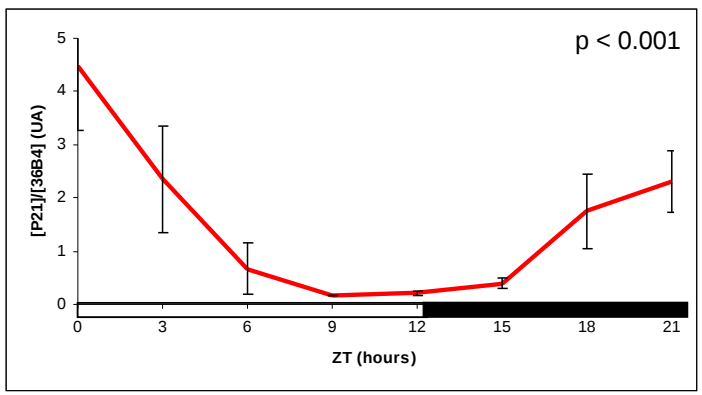
Ccna2



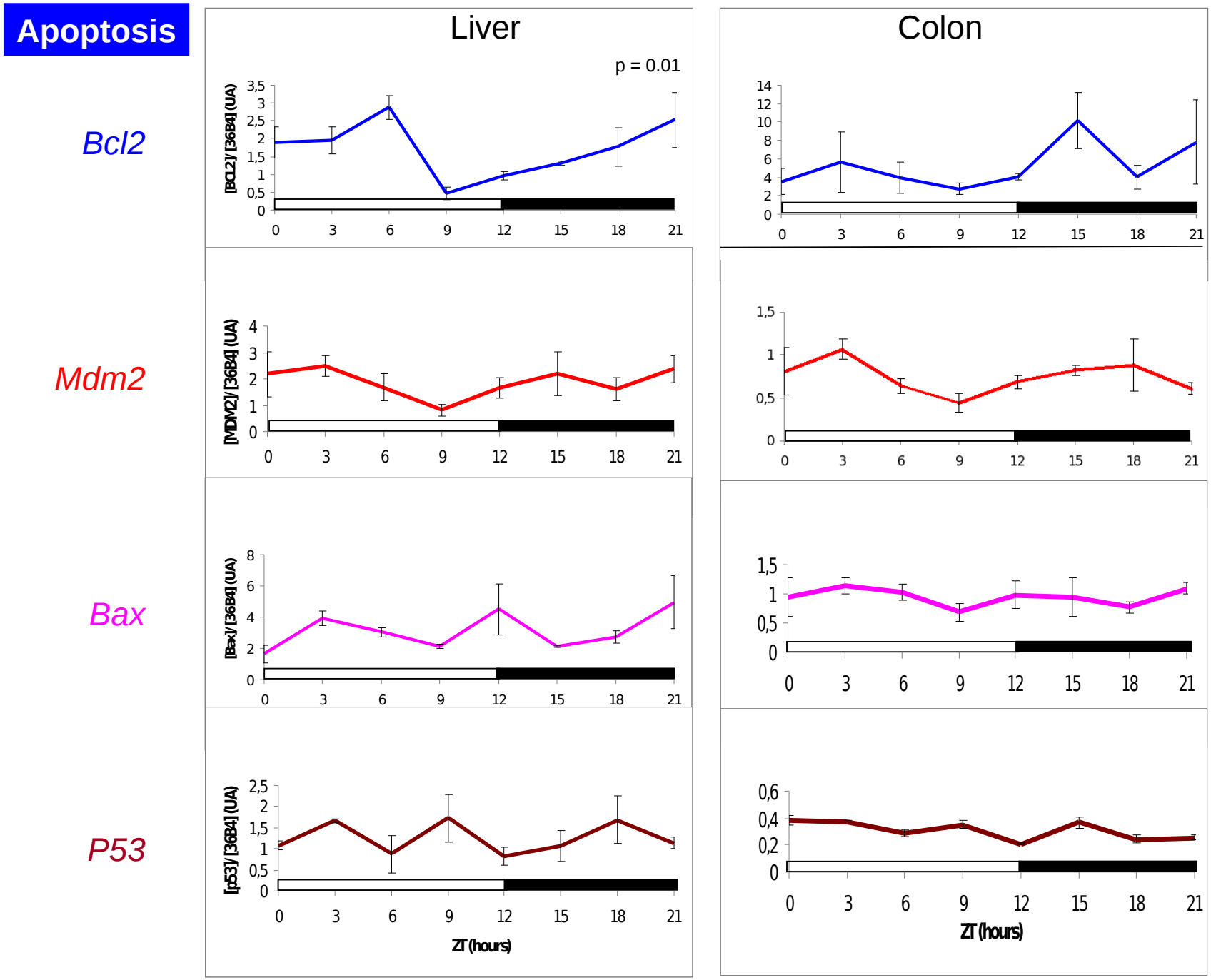
Ccnb2



P21



Apoptosis

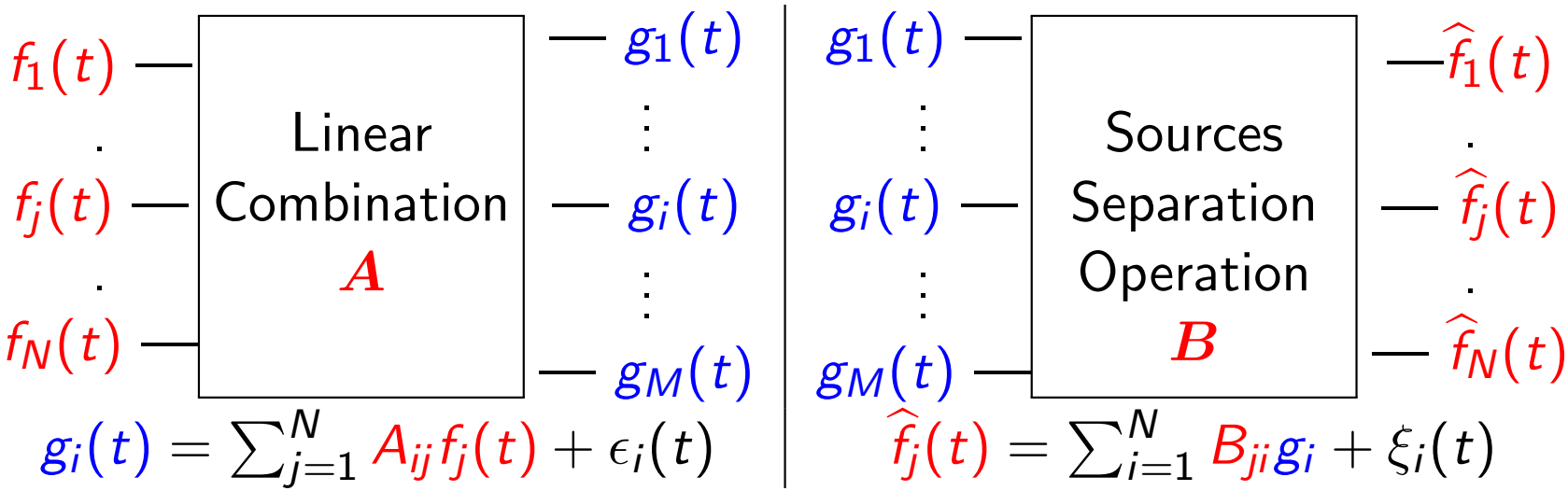

Liver
Colon
p = 0.01
Bcl2
Mdm2
Bax
P53
[BCL2]/[36B4] (UA)
[MDM2]/[36B4] (UA)
[Bax]/[36B4] (UA)
[p53]/[36B4] (UA)
ZT (hours)

Different Questions to Answer

- What are the most important variables?
  - Principal Component Analysis (PCA) and Factor Analysis (FA)
  - Independent Component Analysis (ICA)
- What are the most discriminent variables?
  - Linear Discriminent Analysis (LDA)
  - Quadratic Discriminent Analysis (QDA), EDA, RDA, XDA, SVM, ...
  - Different classification and clustering methods (supervised or unsupervized)
- What are the most important links between variables?
  - Pearson Correlation measure ρ
  - Spearman Correlation measure ρ_s
  - Kendall Correlation measure τ
- What are the most important directed links between variables?
  - Directed Graph
  - Bayesian Network
  - Copula based Networks

What are the most important variables?

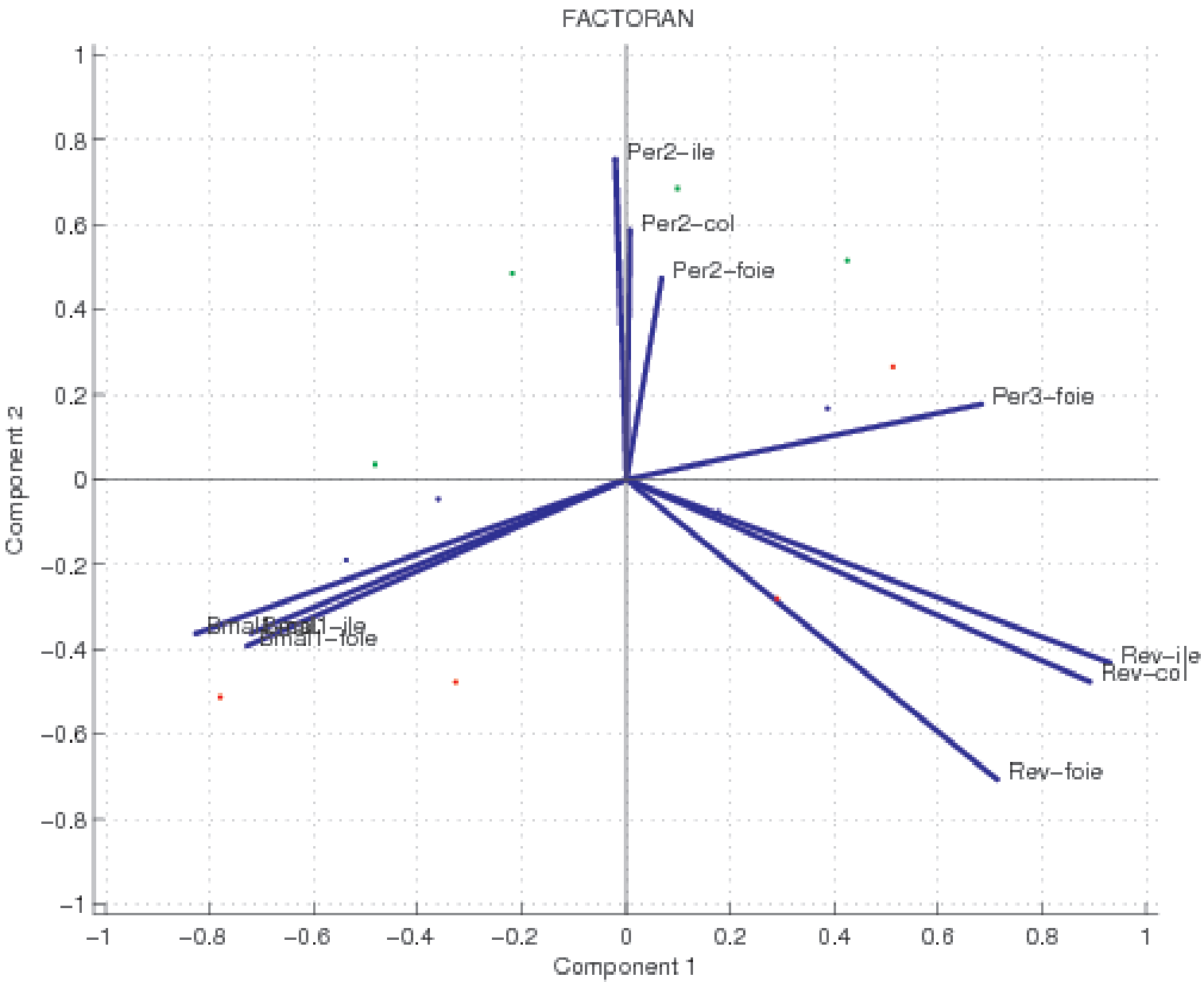
Basic idea: the data $g_i(t)$ can be
obtained by a linear combination of some basic sources (factors) $f_j(t)$



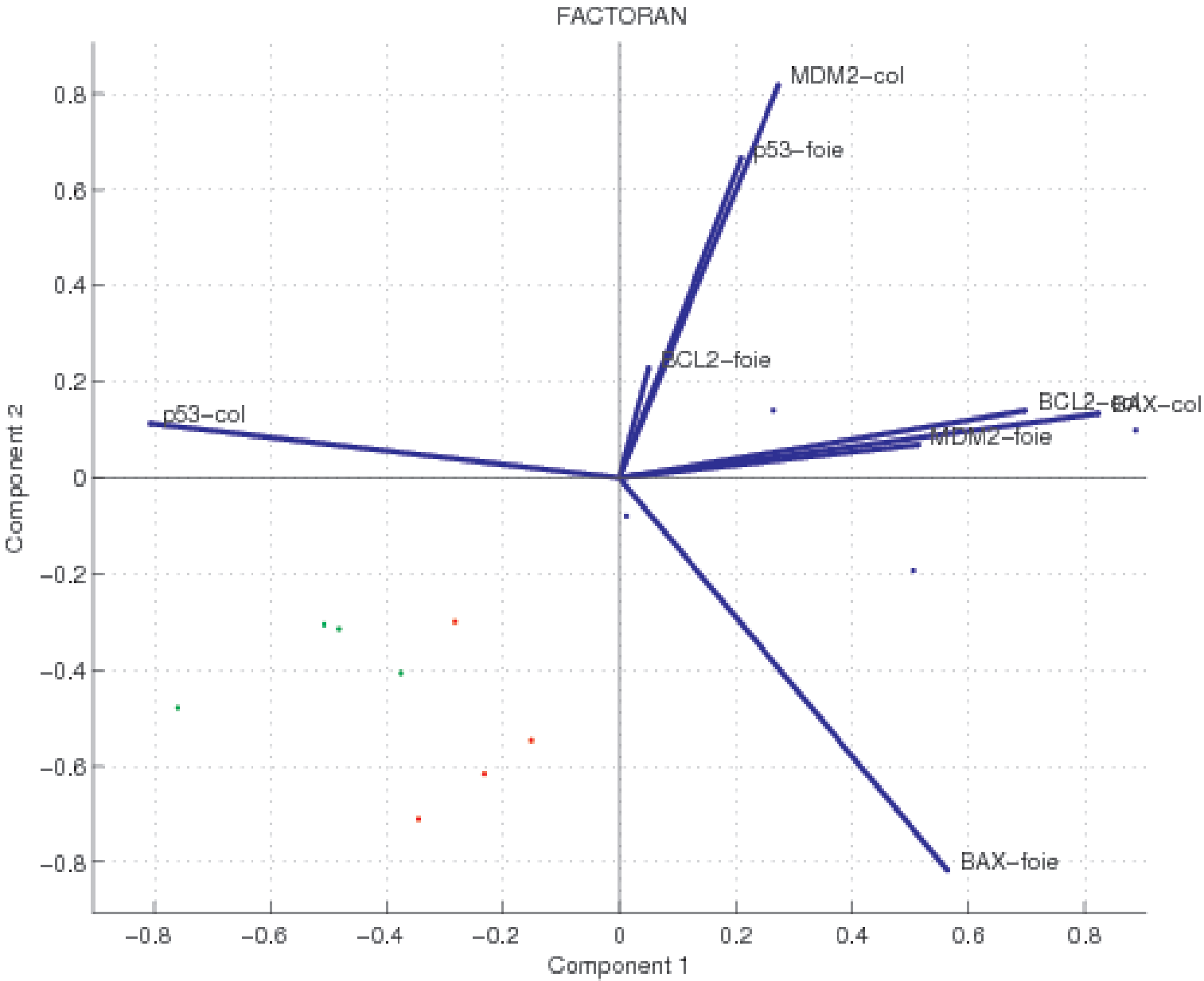
$$g_i(t) = \sum_{j=1}^{N} A_{ij} f_j(t) + \epsilon_i(t) \qquad \widehat{f}_j(t) = \sum_{i=1}^{N} B_{ji} g_i + \xi_i(t)$$

- $\boldsymbol{A}$ is called Mapping or Mixing Matrix
- Columns of $\boldsymbol{A}$ are called Factors Loading.
- PCA and FA are the tools to obtain: $\Rightarrow \boldsymbol{A}, \boldsymbol{f}(t)$
- $\boldsymbol{B}$ is called Separation Matrix
- ICA methods focus on determining $\boldsymbol{B}$
- Sources Separation methods try to estimate $\boldsymbol{B}$ and $\boldsymbol{f}(t)$

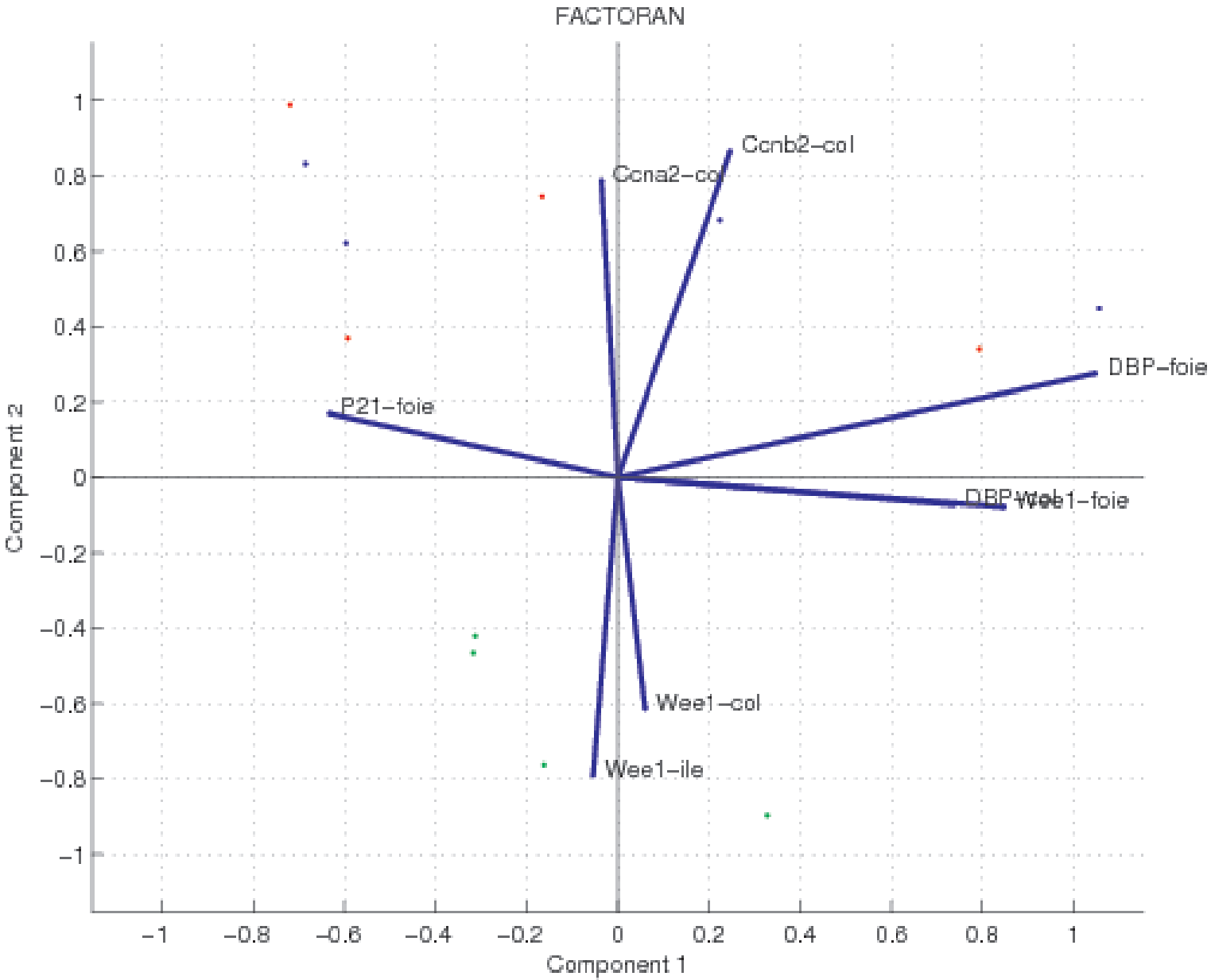
Most important Clock Genes



Most important Apoptosis Genes

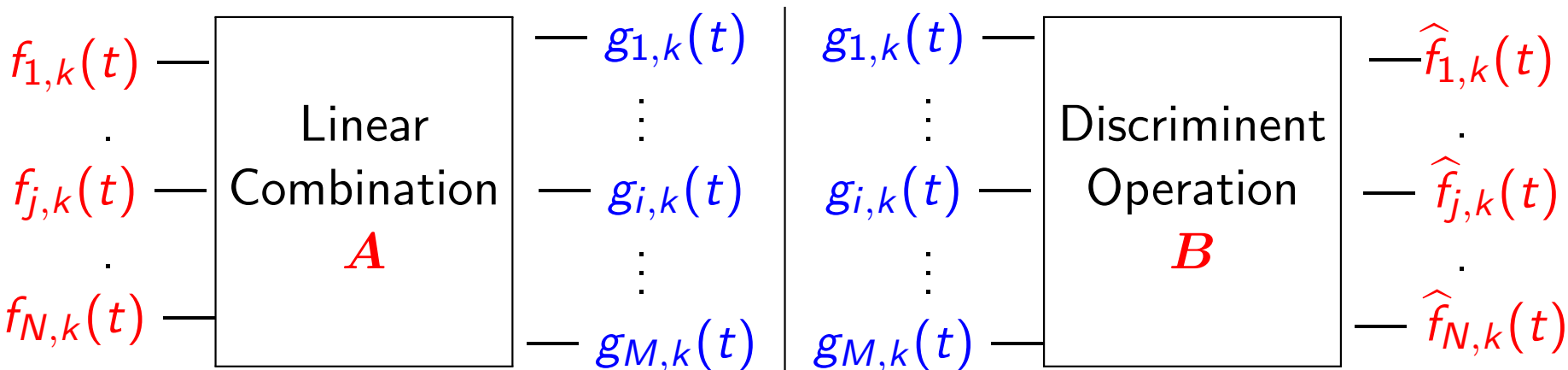


Most important Cell Cycle Genes



Discriminent Analysis

- Very often the data are gathered for different cases (classes)
- Discriminent Analysis try to find the minimum number of factors which are the most discriminent
- The data are
  $g_{i,k}(t), \quad i = 1, \cdots, M, \ k = 1, \cdots, K, \ t = 1, \cdots, T$



$g_{i,k}(t) = \sum_{j=1}^{N} A_{ij} f_{j,k}(t) + \epsilon_{i,k}(t)$ | $\widehat{f}_{j,k}(t) = \sum_{i=1}^{N} B_{ji} g_{i,k} + \xi_{i,k}(t)$

- Discriminent Analysis methods try to find $\boldsymbol{B}$

Discriminent Analysis and Classification

- DA methods try to obtain the most discriminent factors via the matrix $\boldsymbol{B}$.
- Linear Discriminent Analysis (LDA) tries to find hyperplanes equations between any two classes. Need the inter $\boldsymbol{\Sigma}_b$ and intra $\boldsymbol{\Sigma}_w$ classes covariance matrices

$$
\begin{aligned}
\bar{\boldsymbol{g}} &= \frac{1}{MK}\sum_i \sum_k \boldsymbol{g}_{i,k} && \text{global mean} \\
\bar{\boldsymbol{g}}_k &= \frac{1}{M}\sum_i \boldsymbol{g}_{i,k} && \text{means of each class} \\
\boldsymbol{\Sigma}_w &= \frac{1}{MK-K}\sum_k \sum_i (\boldsymbol{g}_{i,k}-\bar{\boldsymbol{g}}_k)(\boldsymbol{g}_{i,k}-\bar{\boldsymbol{g}}_k)' && \text{within} \\
\boldsymbol{\Sigma}_b &= \frac{K}{K-1}\sum_k (\bar{\boldsymbol{g}}_k-\bar{\boldsymbol{g}})(\bar{\boldsymbol{g}}_k-\bar{\boldsymbol{g}})' && \text{between}
\end{aligned}
$$

- The solution is described as

$$\widehat{\boldsymbol{B}} = \arg\max_{\boldsymbol{B}} \left\{ \frac{|\boldsymbol{B}'\boldsymbol{\Sigma}_b\boldsymbol{B}|}{|\boldsymbol{B}'\boldsymbol{\Sigma}_w\boldsymbol{B}|} \right\}$$

  which is obtained via SVD: $[\boldsymbol{\Sigma}_w]^{-1}\boldsymbol{\Sigma}_b\boldsymbol{B}_{*i} = \lambda_i \boldsymbol{B}_{*i}$
- Many extensions via Mixture Modelling: LDA, QDA, RDA, EDA, ...

Discriminent Analysis and Classification via Mixture Modelling

- Mixture modelling:

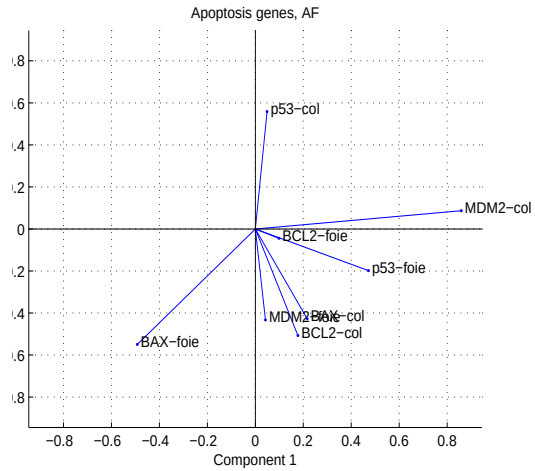
$$p(\boldsymbol{g}|z=k),\ p(z=k) \longrightarrow p(\boldsymbol{g}) = \sum_{k=1}^{K} p(z=k)\ p(\boldsymbol{g}|z=k)$$

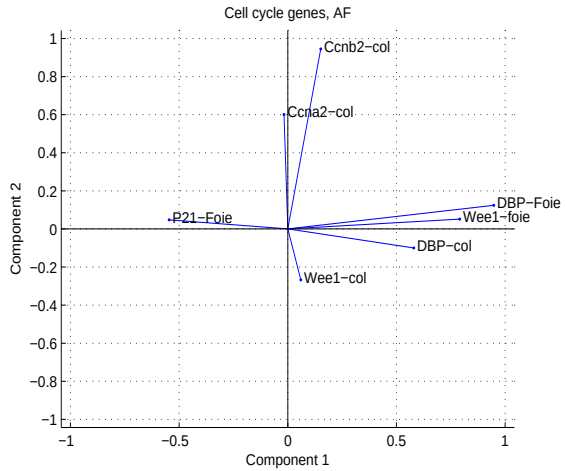
- Mixture of Gaussians (MoG) modelling:

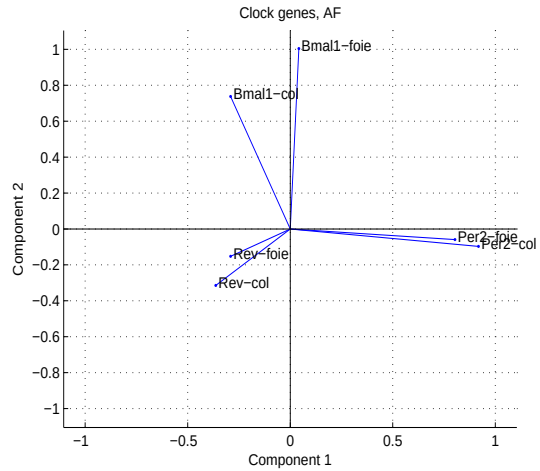
$$p(\boldsymbol{g}) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(\boldsymbol{g}|\boldsymbol{m}_k, \boldsymbol{\Sigma}_k) \longrightarrow \begin{cases} p(\boldsymbol{g}|z=k) = \mathcal{N}(\boldsymbol{m}_k, \boldsymbol{\Sigma}_k) \\ P(z=k) = \alpha_k \end{cases}$$

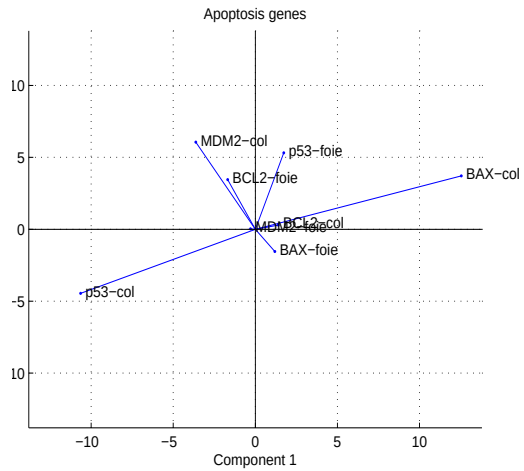
- Learning: Estimation of the parameters $\boldsymbol{\theta} = \{\alpha_k, \boldsymbol{m}_k, \boldsymbol{\Sigma}_k\}$
  Classical or Bayesian EM algorithms
- Classification:
  - Supervised: $p(z=k|\boldsymbol{g}, \boldsymbol{\theta}, K)$
  - Semi-Supervised: $p(z=k|\boldsymbol{g}, K)$
    Needs estimation of parameters $\boldsymbol{\theta}$
  - Unsupervised: $p(C=k|\boldsymbol{g})$
    Needs estimation of K and the parameters $\boldsymbol{\theta}$

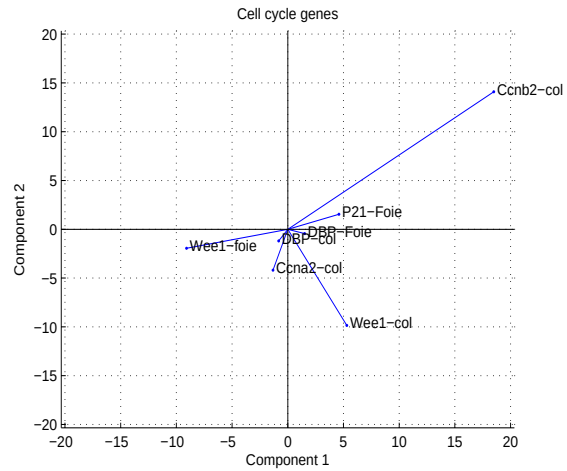
Factor Analysis Vs. Discriminant Analysis

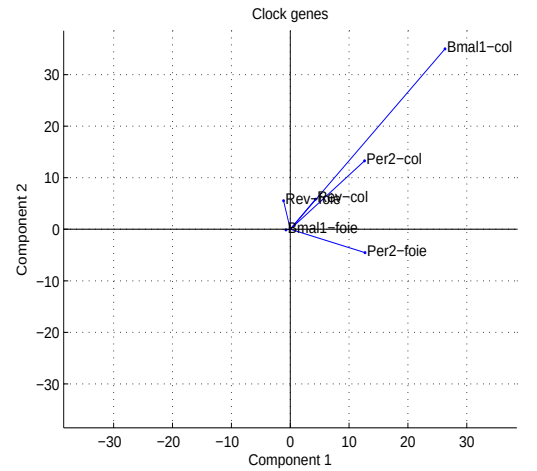












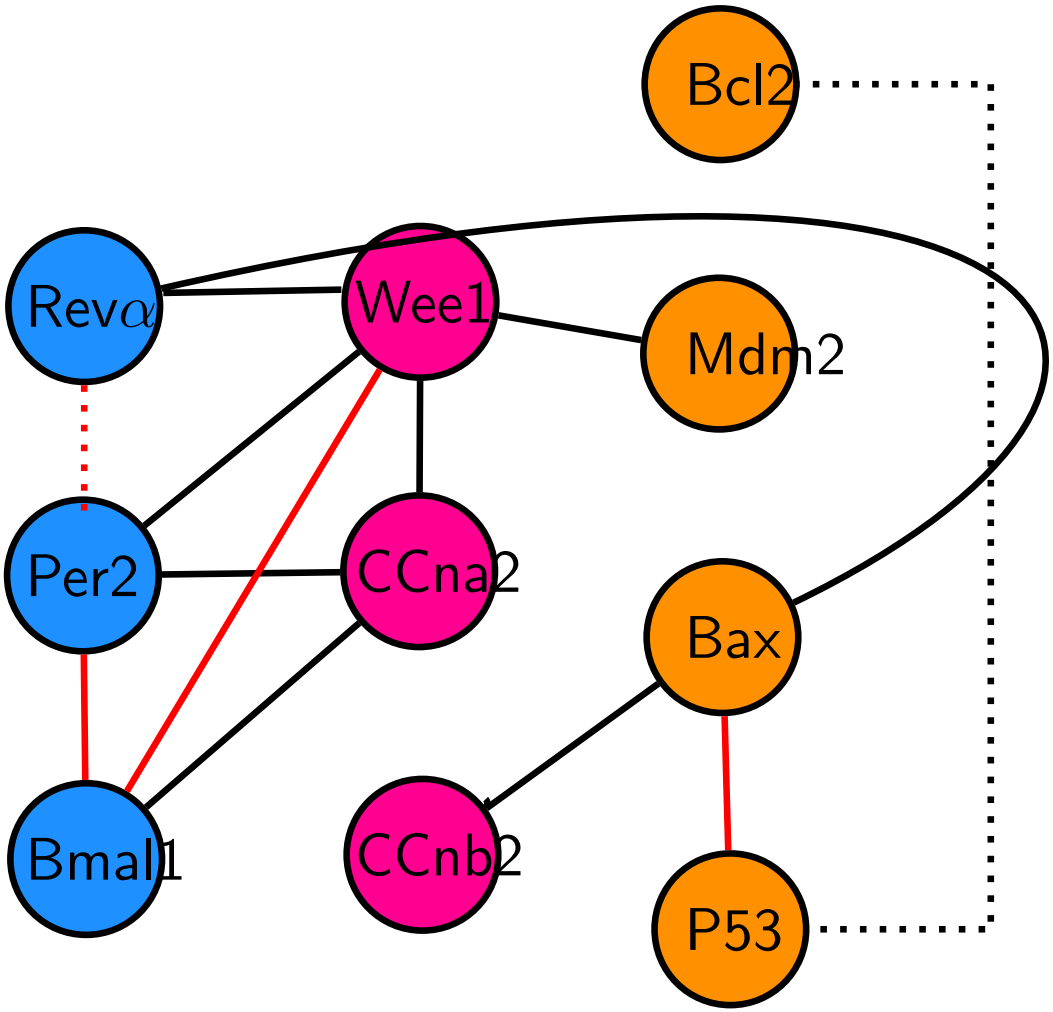
Most important Dependency Graph between variables

- Question: How to measure the dependency between two variables X and Y?
- Classical Pearson correlation ρ

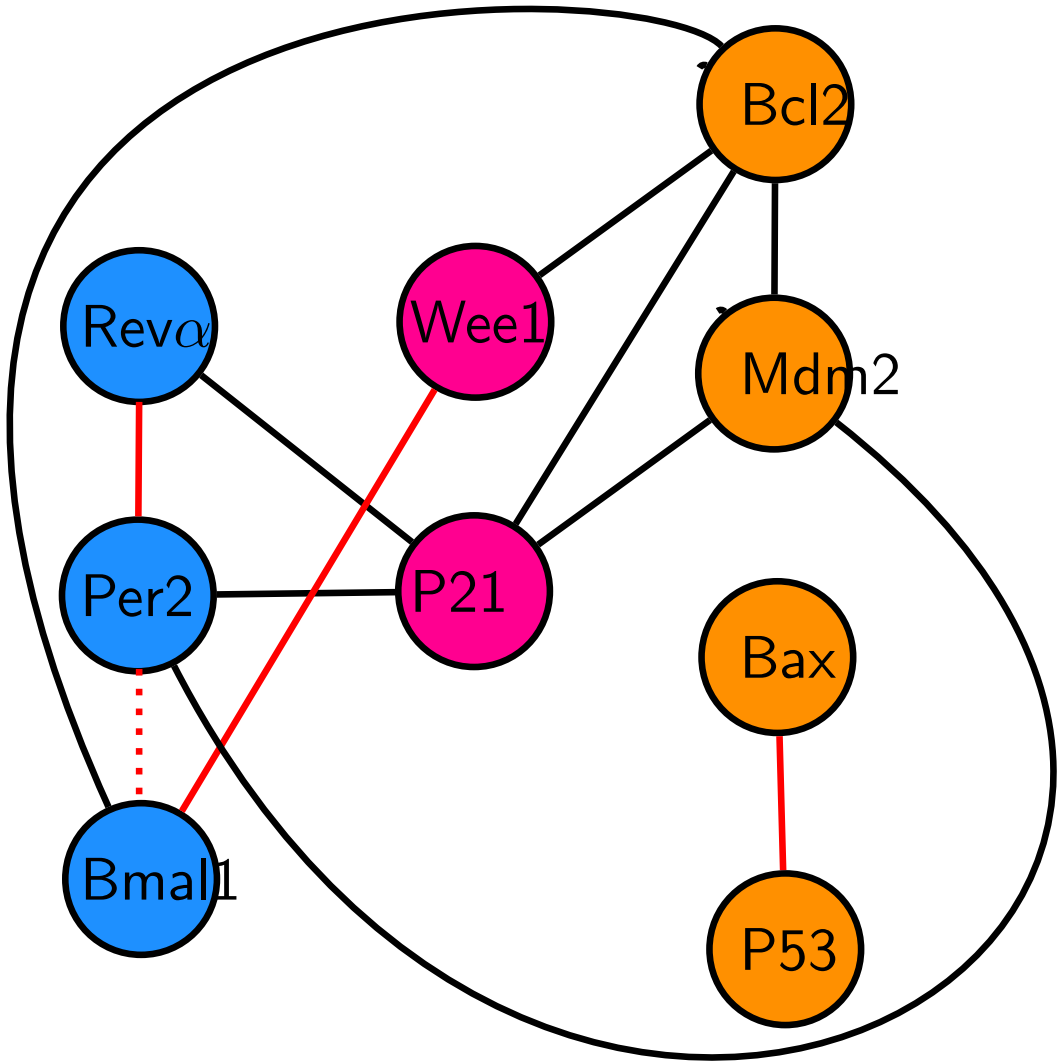
$$\rho = \frac{\mathrm{E}\{X\}\,\mathrm{E}\{Y\}}{\sqrt{\mathrm{E}\{(X-m_x)^2\}\,\mathrm{E}\{(Y-m_y)^2\}}}$$

  measures only the linear dependency
- Two other mesaures are: Spearman ρ_s and Kendall τ
- When $\rho = 0$ does not mean that the variables are not dependent !
- We used a combination of these measures to creat a graph of dependencies between the variables.
- We followed two directions:
  - For each tissue (Liver & Col) and each class, creat separate graphs and then combine the results (Decision fusion)
  - Use the whole data and creat one common graph

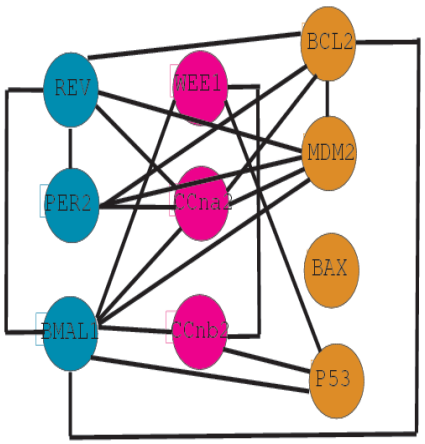
Directed Graph of links between variables



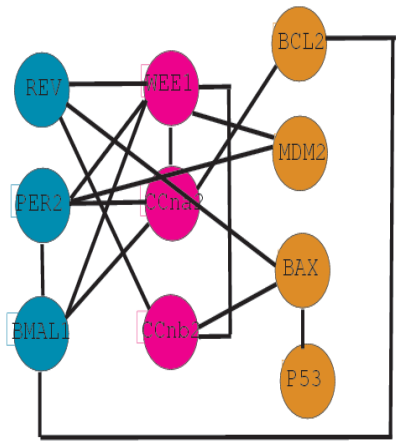
Directed Graph of links between variables



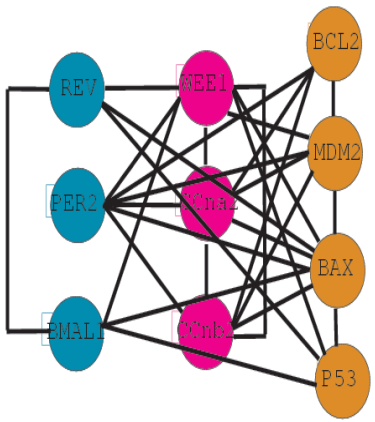
Most important Dependency Graph between variables



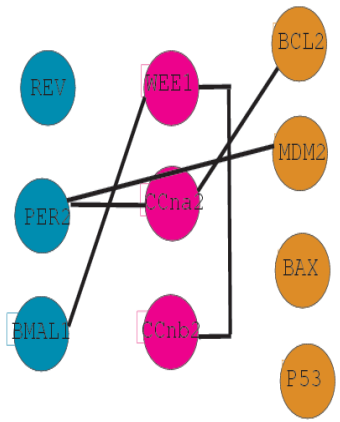
Class 1



Class 2

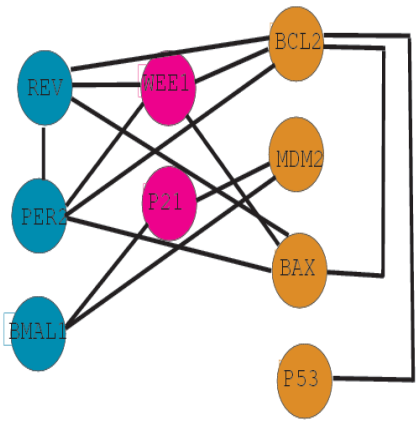


Class 3

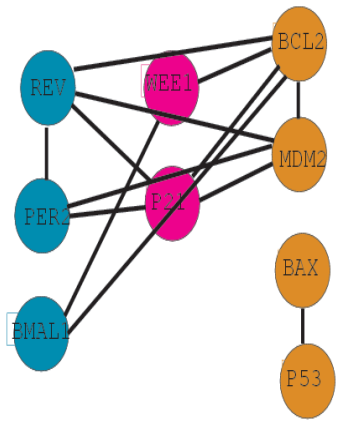


Common

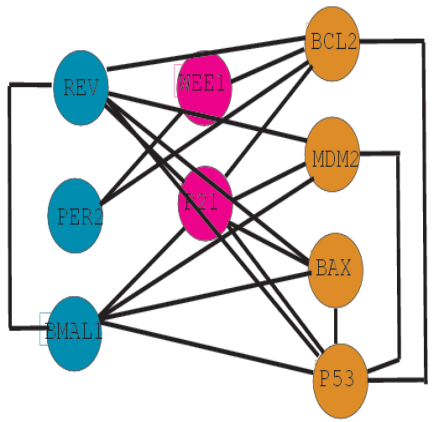
Most important Dependency Graph between variables



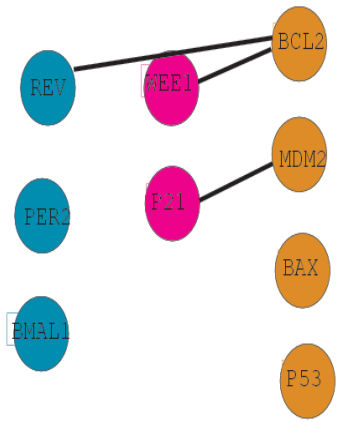
Class 1



Class 2

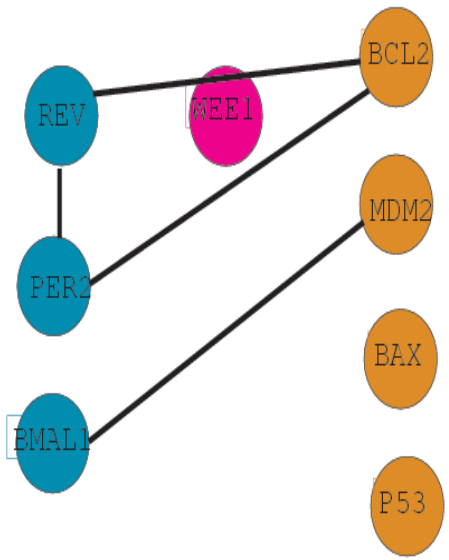


Class 3

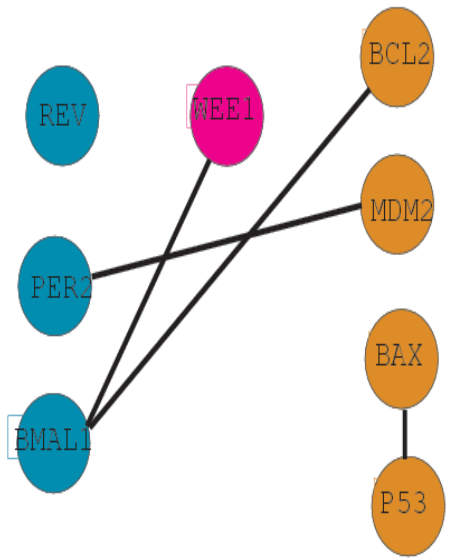


Common

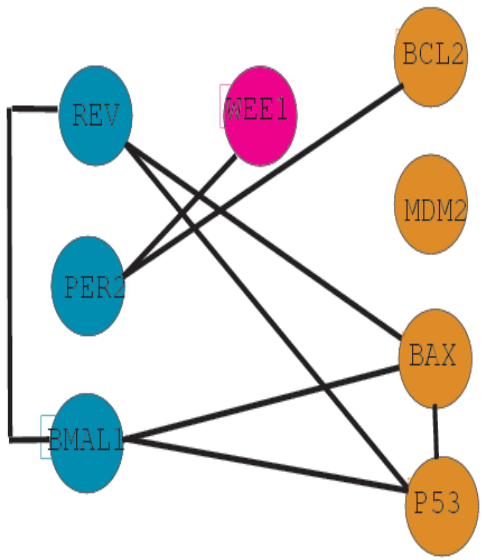
Most important Dependency Graph between variables



Class 1



Class 2



Class 3

Directed Graph of links between variables

- Question: Which genes are affecting the others ? (Causality)
- Directed graphs
- Bayesian networks
- Needs to have more data
- We plan to have more dense time series data (1h) and longer (48h), in normal and blocked genes situations.
- Hope to be able to answer the question: which varaibles are the causes and which expressions are the effects

Questions and Discussion