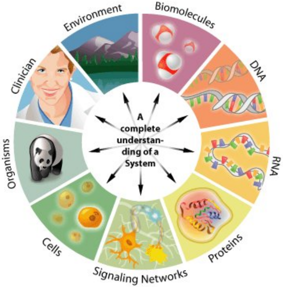
Data integration of highly dimensional biological data sets with multivariate analysis

Kim-Anh Lê Cao

Queensland Facility for Advanced Bioinformatics
The University of Queensland

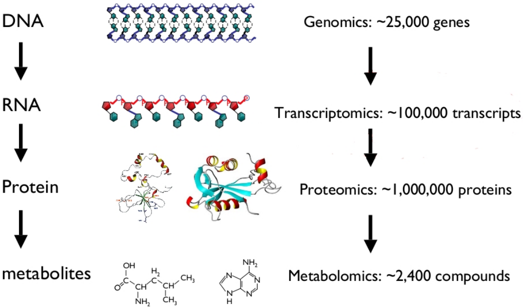
- Study of complex interactions in biological systems
- Holism vs. reductionism

  'Systems biology [...] requires that we develop ways of thinking about integration that are as rigorous as our reductionist programmes, but different [...].It means changing our philosophy, in the full sense of the term', Denis Noble
- Biology-based inter-disciplinary study field



→ Understand better the entirety of processes that happen in a biological system
→ Model and discover emergent properties, properties of cells, tissues and organisms functioning as a system

The biological dogma and the ‘omics’ cascade

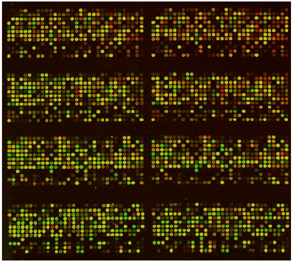


DNA → RNA → Protein → metabolites

Genomics: ~25,000 genes → Transcriptomics: ~100,000 transcripts → Proteomics: ~1,000,000 proteins → Metabolomics: ~2,400 compounds

→ Integrative systems biology: understand the relationships between these functional levels

Transcriptomics: DNA microarray technology



- Measures the expression of thousands of genes on a single individual
- 1 spot = 1 gene
- Gene expression measure = signal intensity
- Spots are ‘on’ (activated) or ‘off’ (silent) across biological conditions

→ **Identify biomarkers or regulated genes to understand the processes of cellular differentiation or carcinogenesis (Supervised analysis)**

High-throughput sequencing

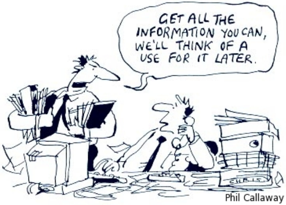
DNA sequencing: methods for determining the order of the nucleotide bases A-C-G-T in a molecule of DNA.

High-throughput sequencing: parallelize the sequencing process, produces thousands or millions of sequences at once

- inexpensive genome-wide sequence and fast
- provides insights into genome variation and evolution
  - genotyping, genome resequencing, de novo genome assembly projects and metagenomics
  - need of efficient methodologies to process and analyse the data

Challenges

Close interaction between statisticians, bioinformaticians and molecular biologists


GET ALL THE INFORMATION YOU CAN, WE'LL THINK OF A USE FOR IT LATER.
Phil Callaway

- Understand the biological problem
- Irrelevant (noisy) variables
- $n << p$ and n very small
  → **limited statistical validation**
- Is the statistical approach is biologically relevant?
- Keep up with new technologies
- Anticipate computational issues

Some research questions

Now, consider the framework of longitudinal 'omics' studies ...

1. Do we observe a 'natural' separation between the different groups of patients across time?
2. Cluster the times profiles for:
   -same type of biological features
   -different type of biological features
   → Identify subsets of correlated features across time
3. Do several assays performed on the same samples contain the same information?

Linear Multivariate approaches

- Dimension reduction
  → project the data in a smaller subspace
- To handle multicollinear, irrelevant, missing variables
- To capture experimental and biological variation

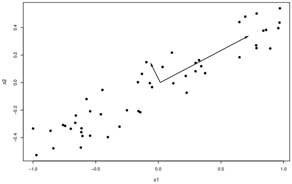
In the R package `mixOmics`, focus is on:

- Data integration
- Variable selection
- Computationally efficient methodologies for large biological data sets
- Interpretable graphical outputs

Principal Component Analysis: PCA

Seek the best directions in the data that account for most of the variability

→ principal components: artificial variables that are linear combinations of the original variables:



$$c_1 = w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$$

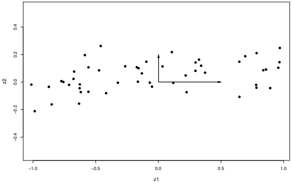
where

- where c_1 is the **first** principal component with max. variance
- $\{w_1, \ldots, w_p\}$ are the weights in the linear combination
- $\{x_1, \ldots, x_p\}$ are the gene expression profiles.

All PCs are mutually orthogonal. $(c_1, c_2, \ldots)$

The new PCs form a a smaller subspace of dimension $< p$

Project the data on these new axes to summarize the information related to the variance.



→ approximate representation of the data points in a lower dimensional space

PCA is an (almost) compulsory first step in exploratory data analysis to:

- Have a first understanding of the underlying data structure
- Identify bias, experimental errors, batch effects

Problem with PCA: interpretation can be difficult with very large number of (possibly) irrelevant variables.

Remember that the principal components are linear combinations of the original variables:

$$\boldsymbol{c} = w_1 \boldsymbol{x}_1 + w_2 \boldsymbol{x}_2 + \cdots + w_p \boldsymbol{x}_p$$

A clearer signal could be observed if some of the variable weights $\{w_1, \dots, w_p\}$ could be set to 0 for the irrelevant variables:

$$\boldsymbol{c} = 0 * \boldsymbol{x}_1 + w_2 \boldsymbol{x}_2 + \cdots + 0 * \boldsymbol{x}_p$$

These variables weights are defined in the loading vectors.
Important weights = important contribution to the PC.
Similar weights = correlated variables.

sparse Principal Component Analysis: sPCA

sparse PCA: sparse loading vectors to remove noisy or irrelevant variables which determine the principal components.

→ Solving PCA through least squares problem (SVD) allows to include regularization parameters

$$\min_{\mathbf{v_h}} ||X_h - \mathbf{u}_h \mathbf{v}_h^T||_F^2 + P_\lambda(\mathbf{u}_h)$$

P_λ is a penalty function with tuning regularization parameter λ

→ use Lasso penalization, or soft-thresholding
→ obtain sparse loading vectors, with very few non-zero elements

Shen, H., Huang, J.Z. 2008. Sparse principal component analysis via regularized low rank matrix approximation, *J. Multivariate Analysis*.

Why PCA can ‘fail’ to summarize the data?

- In some time course experiments, the subject variation can be larger than the time variation
- PCA makes the assumption that samples are independent of each other
- In univariate analysis we use a paired t-test instead of a t-test
- In multivariate analysis we use a multilevel approach:
  - different sources of variation can be separated (treatment effect within subjects and differences between subjects)
  - gain in power

Multilevel approach

- The variation in the data is separated: within matrix and between matrix
- Multivariate tools can then be applied on the within matrix (Westerhuis, 2008)
- We can take into account the repeated measures design of the experiment

VEGFC Study: Human lymphatic endothelial cells were treated in vitro with recombinant VEGF-C for 16 time points: 0min, 15min, 30min, 45min, 60min, 80min, 100min, 2h, 2.5h, 3h, 3.5h, 4h, 5h, 6h, 7h, or 8h) in triplicates, CAGE data (FANTOM5, Riken Institute).

Liquet*, B., Lê Cao*, K-A., Hocini, H., Thiébaut, R. A novel approach for biomarker selection and the integration of repeated measures experiments from two platforms, *BMC Bioinformatics*, accepted (25/09/2012).

VEGFC study: high individual effect

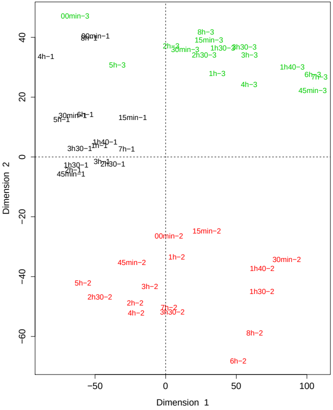


Figure: PCA on original data, color = patient

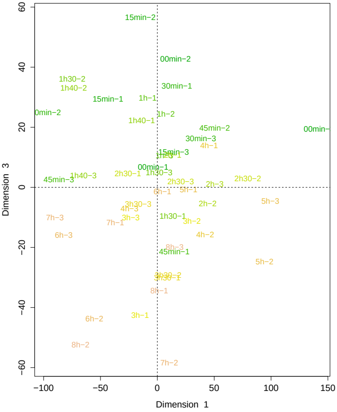
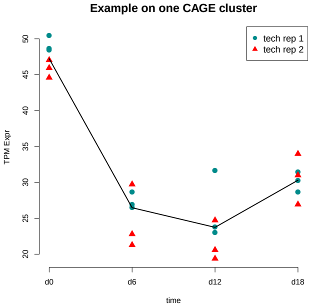


Figure: PCA within matrix, color = time

Modelling trajectories: cubic smoothing splines

Aim: summarize the trajectory of each variable

- Use cubic smoothing splines to summarize each profile
- The derivative between each time point can be estimated
- Fit a non-supervised algorithm to cluster the profiles based on the derivative (k-means, SOM)



Déjean et al. (2008), Clustering Time-Series Gene Expression Data Using Smoothing Spline Derivatives *Eurasip J.*

Example with K-means

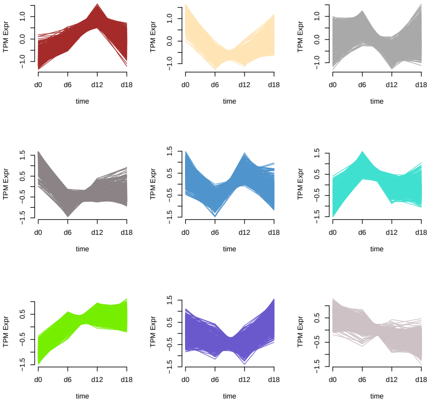


Figure: K-means on derivatives

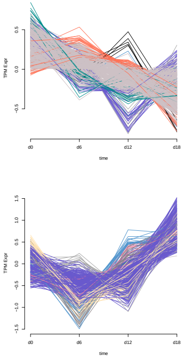


Figure: K-means on original data

How about with Self Organising Maps (SOM)?

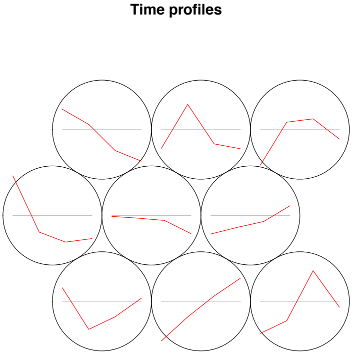


Figure: SOM summary

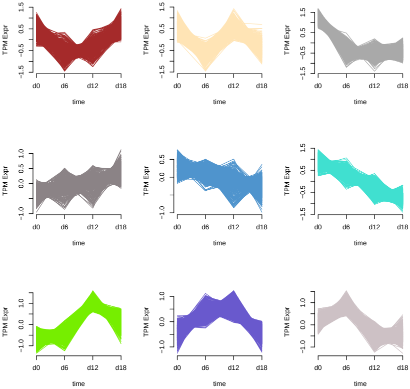


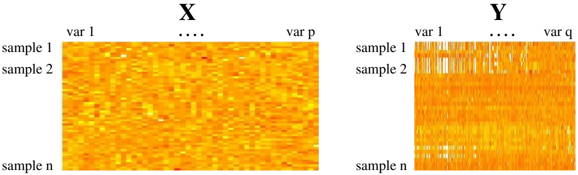
Figure: SOM on original data

Link between smoothing splines and LMM

- Be able to assess the variability for each feature:
  - technical variability
  - biological variability
- Well fitted for correlated repeated measures
- Fits into a linear mixed model framework (not parameters to tune, Verbyla et al. 1999)
- Can take into account random intercepts, bio reps as random effects ...
- Enables interpolation of missing values
- Enables to model the shape of the trajectories
- Variance components and estimates of fixed and random effects can be obtained

Integrating two large data sets

Aim: integrate two data sets and select the relevant features simultaneously:



- Two large data sets X and Y
- Measurements of two types of *variables* on the same samples

- Partial Least Squares regression maximises the covariance between each linear combination (components) associated to each data set

$$\max_{||u_h||=1, ||v_h||=1} cov(X_h\mathbf{u}_h, Y_h\mathbf{v}_h), \qquad h = 1 \dots H$$

where X (n x p) is the transcriptomics data set and Y (n x q) is the proteomics data set

- Similarly to PCA, the PLS components indicate the similarities between samples (useful plots!)
- The loading vectors indicate the contribution of the variables of the same type to the PLS component (useful for variable selection)

sparse PLS-SVD

Use the PLS-SVD variant that directly gives the latent variables and loading vectors and low rank rank approximation.

Let $M_h = X_h^T Y_h$, sparse PLS solves the optimization problem:

$$\min_{\mathbf{u_h},\mathbf{v_h}} ||M_h - \mathbf{u}_h \mathbf{v}'_h||_F^2 + P_{\lambda_1}(\mathbf{u}_h) + P_{\lambda_2}(\mathbf{v}_h)$$

where P_λ is a penalty function

→ obtain simultaneously sparse loadings $\mathbf{u}_h$ and $\mathbf{v}_h$
→ simultaneous select variables from both data sets which are correlated across samples

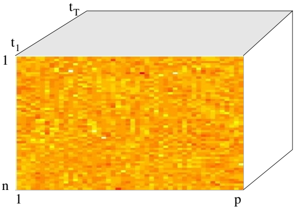
Lê Cao K-A., Rossouw D., Robert-Granié C. and Besse P. 2008. A Sparse PLS for Variable Selection when Integrating Omics data. *SAGMB* **7**(1).

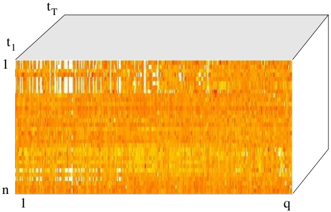
Parameters to tune

- Number of PLS components:
  -Q_h^2 index
  -graphical outputs
- Lasso penalizations λ_1^h, λ_2^h $(h = 1, \dots, H)$:
  -error prediction with cross-validation
  -maximisation of the covariance, stability analysis, permutations(?)

→ the biologist will also help choosing these parameters!

Integration of two longitudinal studies





- Select correlated profiles across time, between and within each data set.
- But difficult to deal with 3D data sets!
- PLS can integrate 2 data sets of 2 dimensions.

Step 1: use cubic smoothing splines to reduce one dimension (samples dimension)

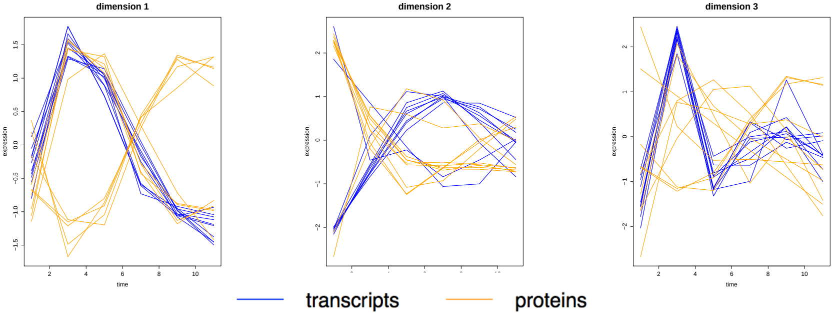
Step 2: apply sPLS on the estimated splines to identify correlated profiles both within and between the two data sets

Two illustrative studies:

Kidney transplant study: Transcriptomics and proteomics study of 40 patients with kidney transplant, rejecting ($n_1 = 20$) or not ($n_2 = 20$) the transplant. Follow up on 5 time points (weeks), PROOF Centre, UBC.

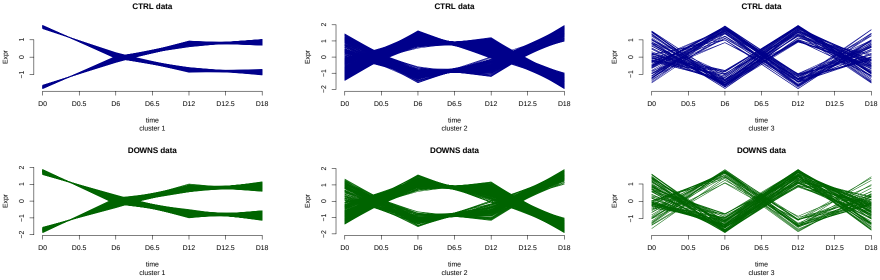
Neuronal study: Human induced pluripotent stem cells from Downs syndrome patients and controls differentiated to neurons (CAGE data). 2 bio reps and 3 tech reps per genotype (control, down syndrome), 4 time points (days), FANTOM5, Riken Institute .

Profile clusters on kidney transplant study



- sPLS selects both transcripts and proteins which are positively or negatively correlated across time
- Quality of clusters decreases with the number of PLS components (dimensions) as obvious patterns cannot be extracted anymore

Concordant profile clusters on Neuronal study

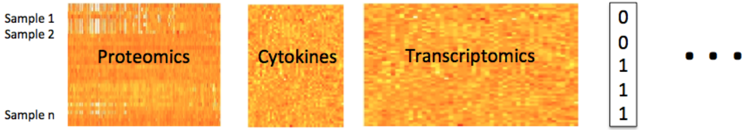

CTRL data

DOWNS data

- Each cluster of profiles corresponds to a PLS component.
- Selection of different features per condition and per component.

Integration of multiple data sets

Integrate heterogeneous data sets



- Need to define the relationships between the different data sets
- Select relevant biological entities which are correlated across the different data sets

Regularized CCA

Classical Canonical Correlation Analysis solves the problem

$$\max cor_{\mathbf{a}_h,\mathbf{b}_h}(\boldsymbol{X}\mathbf{a}_h, \boldsymbol{Y}\mathbf{b}_h) \text{ s.t. } \quad var(\boldsymbol{X}\mathbf{a}_h) = var(\boldsymbol{Y}\mathbf{b}_h) = 1$$

For $n << p + q$, the empirical covariance matrices are ill-conditionned → canonical correlations close to 1.

In regularized CCA the covariance matrices are replaced by:

$$Cov(\boldsymbol{X}) + \lambda_1\mathbf{Id} \text{ and } Cov(\boldsymbol{Y}) + \lambda_2\mathbf{Id}$$

González I., Déjean S., Martin P.G.P., Goncalves O., Besse P. and Baccini A. 2009 Highlighting Relationships Between Heteregeneous Biological Data Through Graphical Displays Based On Regularized Canonical Correlation Analysis, *Journal of Biological Systems*, 17 (2).

Multi-block analysis: Regularized Generalised CCA

- RGCCA generalizes rCCA to more than 2 data sets
- Constitutes a general framework for many multi-block data analysis methods
- Objective: seeks linear combinations of block variables:
  (i) block components explain their own block well and/or
  (ii) block components that are assumed to be connected are highly correlated.

Tenenhaus, A., Tenenhaus, M (2011) Regularized Generalised Canonical Correlation Analysis, *Psychometrika*, 76 (2).

RGCCA

For J blocks of variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_j$, the design matrix $\boldsymbol{C} = \{c_{j,k}\}$, the function g and the shrinkage constants $\tau_1, \ldots, \tau_J$,

RGCCA optimizes the problem:

$$\max_{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_J} \sum_{j,k=1, j \neq k}^{J} c_{kj} g(Cov(\boldsymbol{X}_j \boldsymbol{a}_j, \boldsymbol{X}_k \boldsymbol{a}_k))$$

subject to the constraints $\tau_j ||a_j||^2 + (1 - \tau_j) Var(\boldsymbol{X}_j a_j) \quad j = 1, \ldots, J$, where the $\boldsymbol{a}_j$ are the loading vectors associated to each block j.

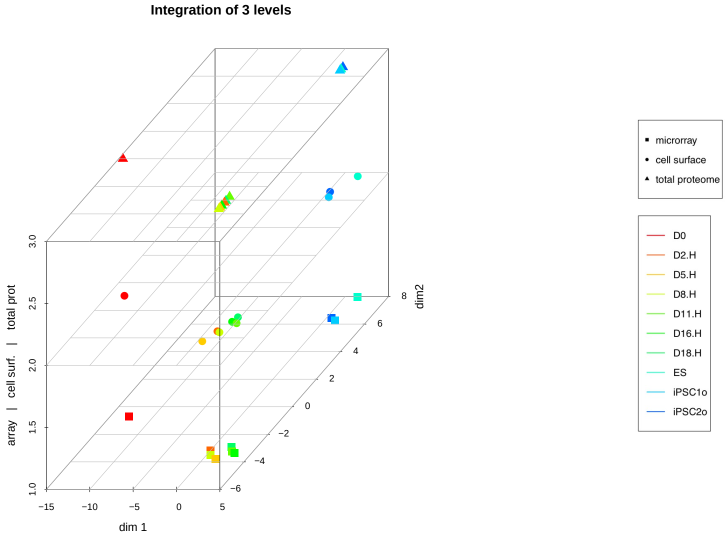
sGCCA

- Similar to the sPLS, L_1 penalizations can be applied to the loading vectors to obtain a sparse version of RGCCA to select different types of biological entities across different functional levels

Grandiose project: Longitudinal study of cell reprogramming. In this study: 8 time points are considered. Multi platform study involving: 6 platforms: microarray, cell surface proteome, total proteome, RA-seq isoform, RNA-seq genes, miRNA.

Tenehaus, A., et al. Variable Selection For Generalized Canonical Correlation Analysis, *submitted.*

Results on Grandiose project



Conclusions

- Statistical exploratory and integrative tools to extract patterns in time course data
- Can be applied to a variety of problems
- Does not provide p-values but can help generating new hypotheses, further statistical tests can then be applied
- Future directions: biological interpretation of the gene lists, time delay, generalised multi-way analysis, identifying discordant clusters across data sets for the same genes ...

Un GRAND MERCI à ...

Neuronal time course

Christine Wells	UQ
Ernst Wolvetang	UQ

Kidney transplant study

Oliver Günther	UBC
Scott Tebutt	UBC

Grandiose proj. and Nagy group

Andras Nagy	Univ. Toronto

`mixOmics` team

Sébastien Déjean	Univ. Tlse
Ignacio González	Univ. Tlse
Xin Yi Chua	QFAB

VAC18 Project and multilevel

Benoît Liquet	Univ. Bdx2

Multiple data integration

Arthur Tenenhaus	Supelec Paris

Time course developments

Kathy Ruggiero	Univ. Auckland
Sébastien Gadat	Univ. Toulouse
Christèle Robert	INRA Toulouse

FANTOM consortium

sample providers	RIKEN, Japan