Extensions of CCA and PLS to unravel relationships between two data sets

S. Déjean[1] - I. González[2] - K-A. Lê Cao[3]

1. Institut de Mathématiques de Toulouse, UMR 5219
Université de Toulouse et CNRS
sebastien.dejean@math.univ-toulouse.fr
2. Plateforme Biopuces, Genopôle Toulouse Midi-Pyrénées
Institut National des Sciences Appliquées
ignacio.gonzalez@insa-toulouse.fr
3. ARC Centre of Excellence in Bioinformatics
Institute for Molecular Bioscience, University of Queensland, Australia
k.lecao@uq.edu.au

useR! Conference 2009 – Rennes, July 8-10

History

Once upon a time in Toulouse, a city in South West of France, two groups of scientists lived nearly together without talking to each other.
But one day, they decided to do so and to work together.
They had Ph.D students, wrote articles and built R packages...

« Stat »

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

$$(X'X)^{-1}X'Y$$

DNA
RNA
ATGCC
TACCAGT

« Bio »

ATGCC

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

J. Stat. Soft.

J. Biol. Syst.

CCA

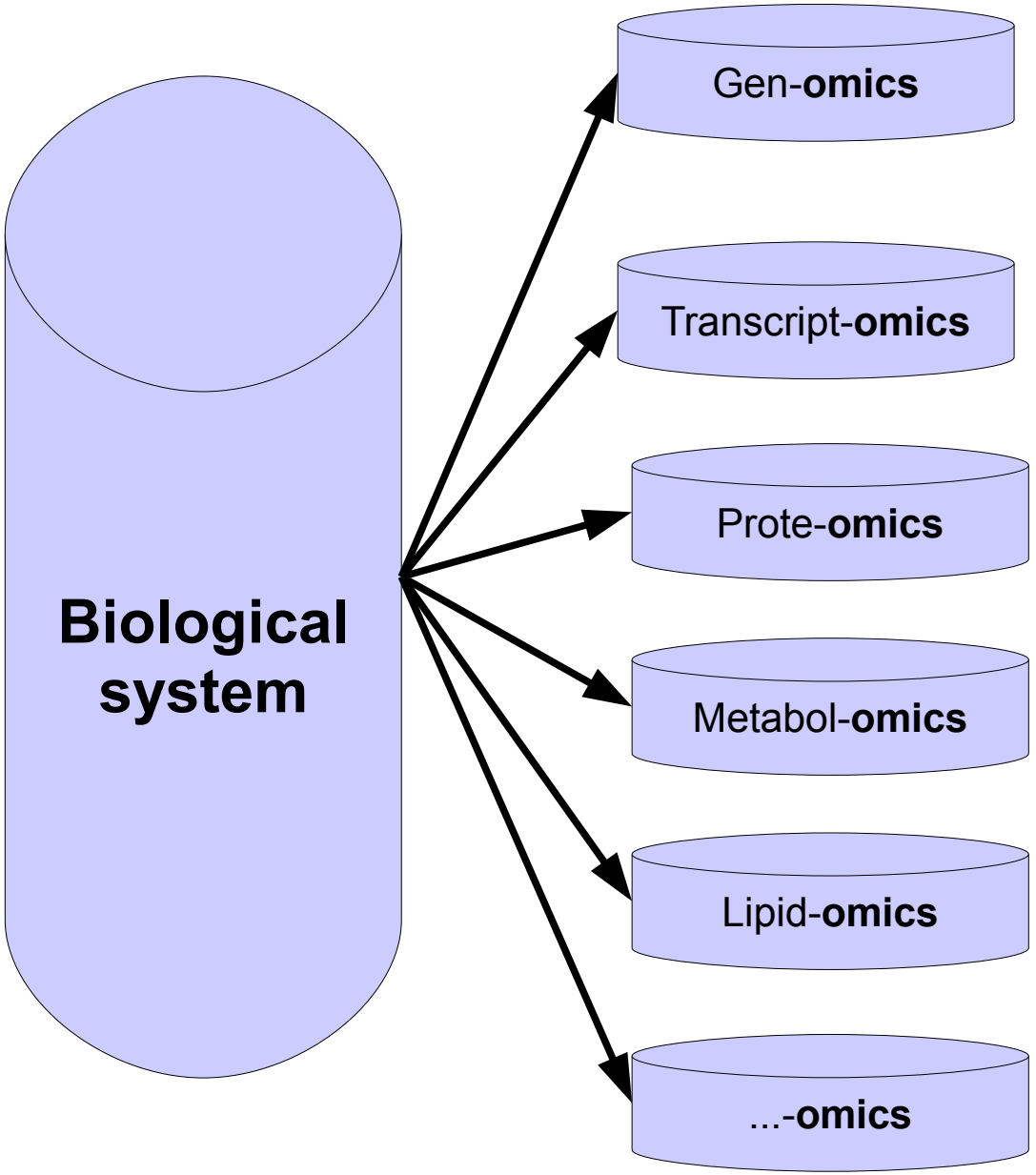
ofw

SAGMB

BMC Bioinformatics

integrOmics

Why integrOmics ?



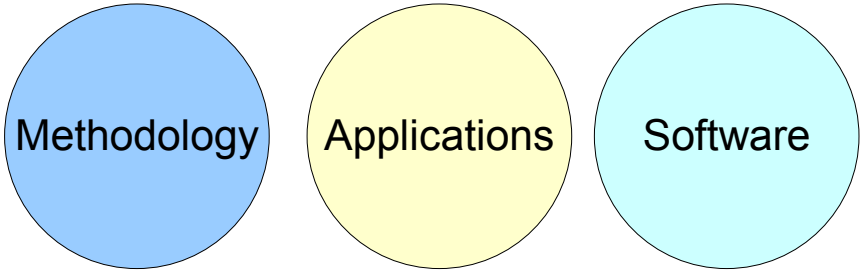
- Each « -omics » data set can be studied separately, but
- A great part of relevant information can be extracted from joint analysis of 2 or several datasets, so

⇒ Integrate omics data project

in short :

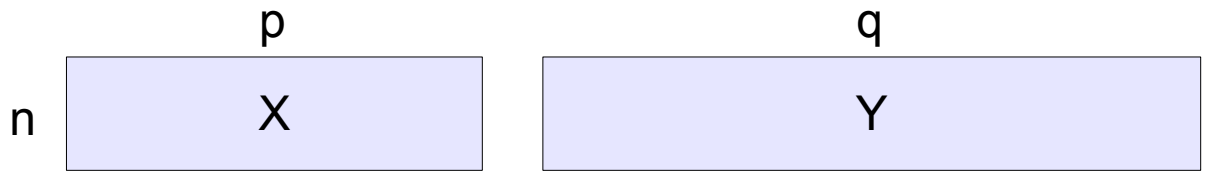
integrOmics

math.univ-toulouse.fr/biostat



Methodology

Two ways to deal with the 'large p - small n' problem in the classical framework provided by Canonical Correlation Analysis and Partial Least Squares regression.



CCA / regularization

- Maximize correlation between linear combination of variables in X and Y
- Requires inversion of XX' and YY'
- **Regularization** of CCA

$$(XX')^{-1} \Rightarrow (XX' + \lambda_X I_n)^{-1}$$

PLS / selection

- Maximize covariance between linear combination of variables in X and Y
- **Selection** obtained through Lasso penalization of loading vectors

Applications

- **E. Yergeau, S.A. Schoondermark-Stolk, E.L. Brodie, S. Déjean, T.Z. DeSantis, O. Gonçalves, Y.M. Piceno, G.L. Andersen and G.A. Kowalchuk** (2008). Environmental microarray analyses of Antarctic soil microbial communities. *The International Society for Microbial Ecology Journal*, 3, 340-351

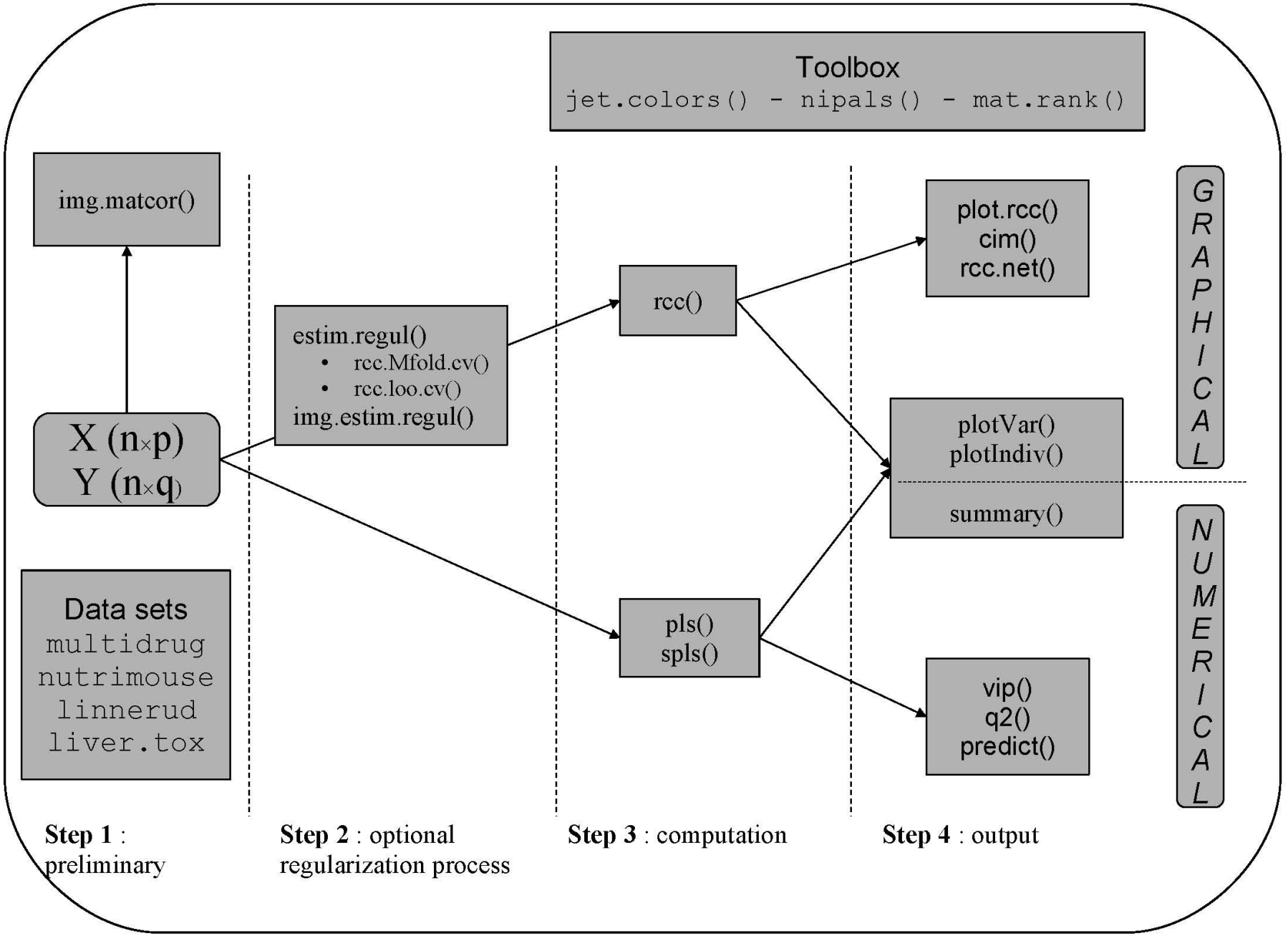
- **S. Combes, I. González, S. Déjean, A. Baccini, N. Jehl, H. Juin, L. Cauquil, B. Gabinaud, F. Lebas, C. Larzul** (2008). Relationships between sensory and physicochemical measurements in meat of rabbit from three different breeding systems using canonical correlation analysis. *Meat science*, 80(3), 835-841

- **I. González, S. Déjean, P.G.P. Martin, O. Gonçalves, P. Besse, A. Baccini** (2009). Highlighting Relationships Between Heterogeneous Biological Data Through Graphical Displays Based On Regularized Canonical Correlation Analysis. *Journal of Biological Systems*, 17(2), 173-199

- **K. A. Lê Cao, D. Rossouw, C. Robert-Granié, P. Besse** (2008). A sparse PLS for variable selection when integrating Omics data, *Statistical Applications in Genetics and Molecular Biology*, 7(1), article 35

IntegrOmics R package

Software



Using integrOmics

Software

- From X and Y two matrices
- Preliminary view of the correlations matrices

```
R> imgCor(X, Y, type = "separate")
```

- Classical CCA

```
R> res.rcc = rcc(X, Y)
```

- Regularized CCA

```
R> res.rcc = rcc(X, Y, 0.05, 0.01)
```

- PLS

```
R> res.pls = pls(X, Y)
```

- Sparse PLS

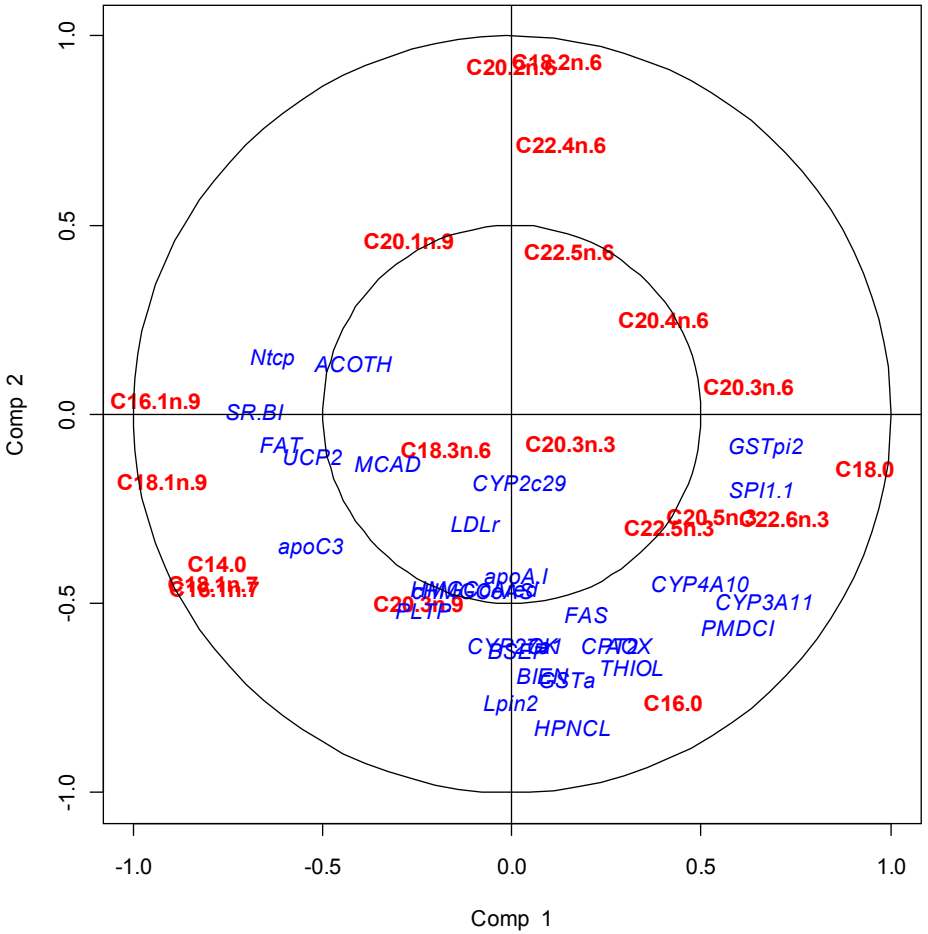
```
R> res.pls = spls(X, Y, mode=c("regression", "canonical"),
+ keep.X=c(10, 10, 10), keep.Y=c(10, 10, 10))
```

Graphical display

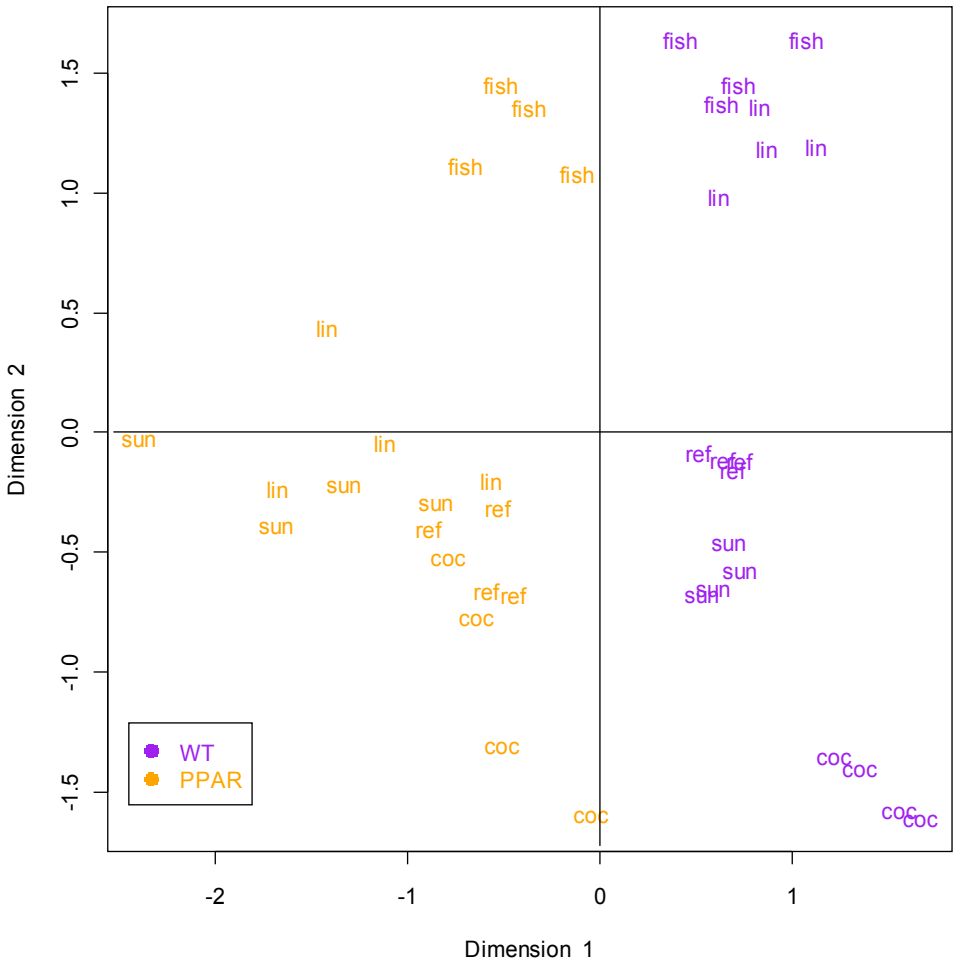
Software

```
R> plotVar(res.rcc,
+ X.label=T,Y.label=T)
```

```
R> plotIndiv(res.rcc),
+ ind.names=nutrimouse$diet
```

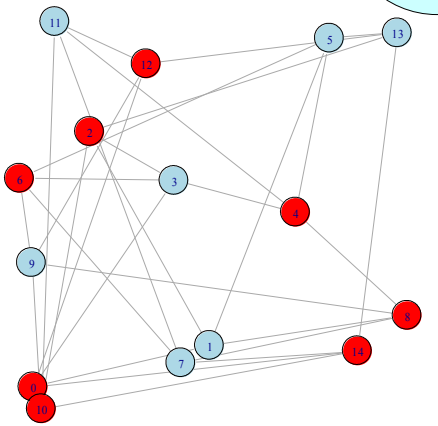


Variables plot

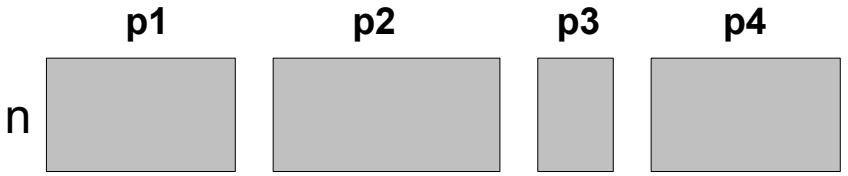


Individuals plot

Future work

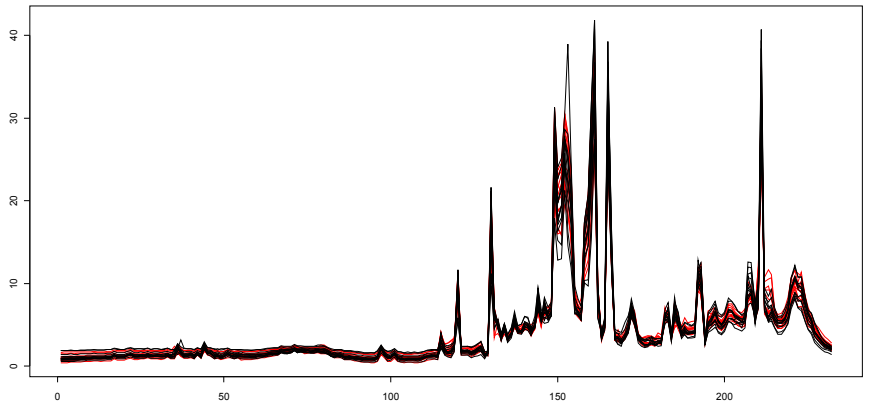


- Provide new graphical display using graphs

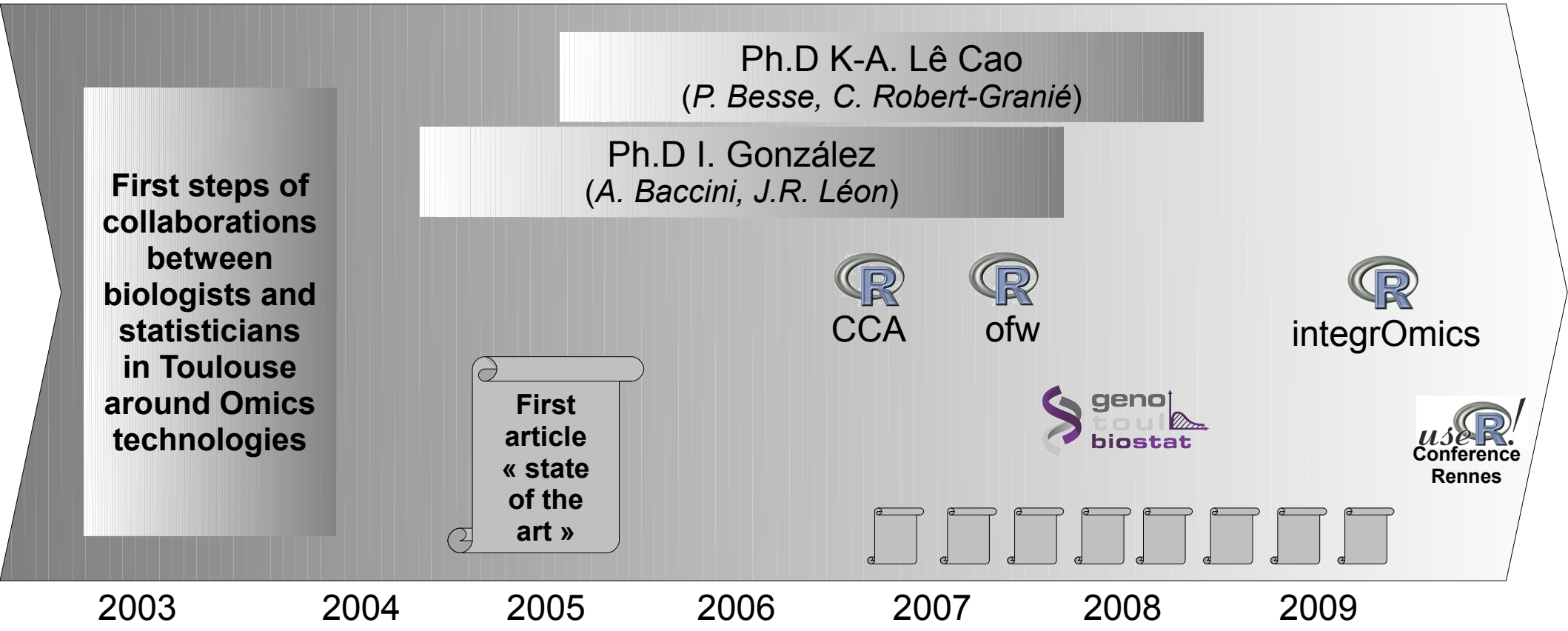


- Methodologies to deal with more than 2 data sets

- Functional statistics to deal with metabolomics or proteomics data



Summary


Ph.D K-A. Lê Cao
(P. Besse, C. Robert-Granié)
Ph.D I. González
(A. Baccini, J.R. Léon)
First steps of collaborations between biologists and statisticians in Toulouse around Omics technologies
First article « state of the art »
CCA
ofw
integrOmics
geno toul biostat
useR! Conference Rennes
2003
2004
2005
2006
2007
2008
2009