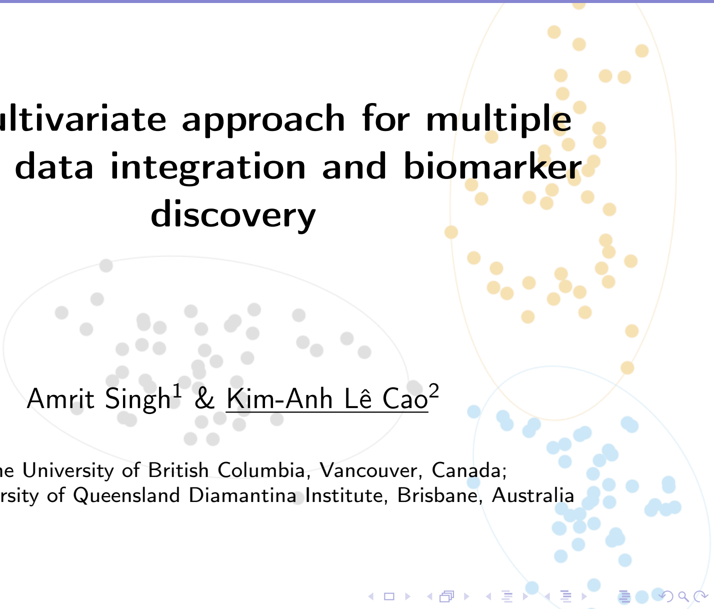
A multivariate approach for multiple ‘omics data integration and biomarker discovery

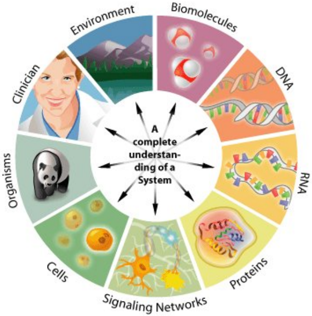


Amrit Singh[1] & Kim-Anh Lê Cao[2]

[1] The University of British Columbia, Vancouver, Canada;
[2] The University of Queensland Diamantina Institute, Brisbane, Australia

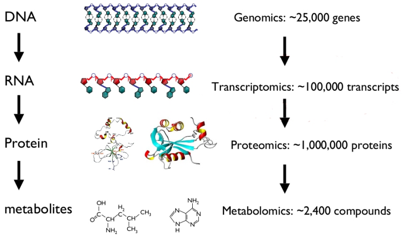
Systems biology is the study of complex interactions in biological systems

- Holistic approach instead of a reductionist approach
- Multi-disciplinary field
- Integration of heterogeneous data

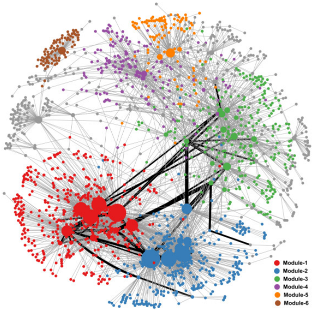


→ we need to develop new ways of thinking and of analysing biological data

How to make sense of biological ‘big data’?


DNA
RNA
Protein
metabolites
Genomics: ~25,000 genes
Transcriptomics: ~100,000 transcripts
Proteomics: ~1,000,000 proteins
Metabolomics: ~2,400 compounds

⟹



from PMID: 22548756

‘What is the key information that can be extracted from heterogeneous data sets?’

Linear multivariate approaches

Linear multivariate approaches use latent variables (e.g. variables that are not directly observed) to reduce the dimensionality of the data.

A large number of observable variables are aggregated in linear models to summarize the data.

- Dimension reduction
  → project the data in a smaller subspace
- Handle highly correlated, irrelevant, missing values
- Capture experimental and biological variation

Some projection-based multivariate methods for data dimension reduction

	Aims	Single ‘omics	Multiple ‘omics
Unsupervised	Data mining Exploration Correlated features	PCA	CCA & PLS MCA (talk: A Bernard) GCCA ($>$ 2 ‘omics)
Supervised	As above **Biomarker discovery**	PLS-DA (talk: F Rohart)	GCC-DA ($>$ 2 ‘omics)

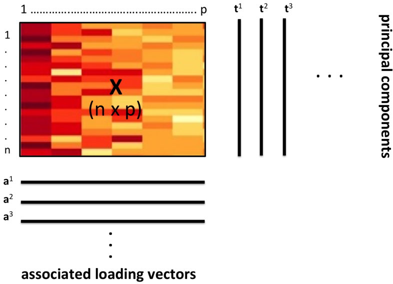
PCA: Principal Component Analysis
PLS: Projection on Latent Structures
DA: Discriminant Analysis
(G)CCA: (Generalised) Canonical Correlation Analysis
MCA: Multiple Correspondence Analysis

Principal Component Analysis (PCA)

Objective function for the first component:

$$\max_{||\boldsymbol{a}||=1} var(\boldsymbol{Xa})$$

- $\boldsymbol{X}$ is a matrix $(n \times p)$,
- $\boldsymbol{a}$ is the loading vector,
- $\boldsymbol{t} = \boldsymbol{Xa}$ is the first principal component (linear combination of p variables)


1 .... p
1
n
X
(n x p)
t1
t2
t3
. . .
principal components
a1
a2
a3
associated loading vectors

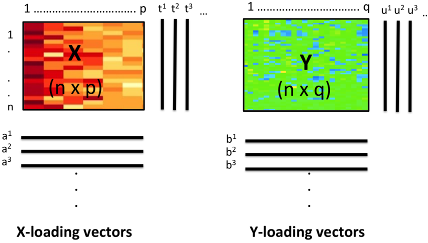
Other principal components follow with the condition that they are orthogonal to each other.

Projection on Latent Structures (PLS)

Objective function for the first set of variates:

$$\arg \max_{||\boldsymbol{a}||=1,\, ||\boldsymbol{b}||=1} \operatorname{cov}(X\boldsymbol{a}, Y\boldsymbol{b}),$$

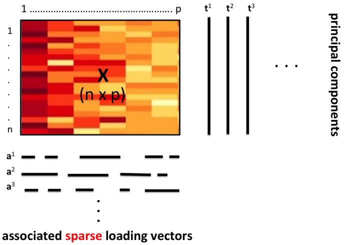
- Matrices: $\boldsymbol{X}$ $(n \times p)$ and $\boldsymbol{Y}$ $(n \times q)$
- Loading vectors: $\boldsymbol{a}$, $\boldsymbol{b}$
- Latent components: $\boldsymbol{t} = \boldsymbol{Xa}$ and $\boldsymbol{u} = \boldsymbol{Yb}$
(linear combination of each set of variables)



Other latent variables follow with the condition that they are orthogonal to each other.

Variable selection: example with sparse PCA

- sPCA is solved iteratively with NIPALS algorithm (Wold 1987) to fit into a least squares framework
- Lasso penalisation removes irrelevant variables when calculating principal components

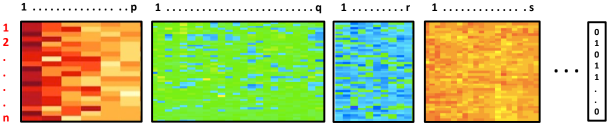


→ **component-wise variable selection**

→ Similar idea for sparse PLS

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *JRSSB*;
Shen, H., Huang, J.Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation, *J. Multivariate Analysis*.
Lê Cao K-A. et al. (2009) A Sparse PLS for Variable Selection when Integrating Omics data, *Stat Appl Gen Mol Biol*, 7(1).

Biomarker discovery when integrating multiple data sets


1 . . . . p
1 . . . . q
1 . . . . r
1 . . . . s
1
2
.
.
.
.
n
. . .
0
1
0
1
1
.
.
0

- Data sets measured on the same samples
- Aim: select relevant biological features that are correlated within and between heterogeneous data sets
- Extends integrative multivariate analysis for more than 2 data sets

Tenenhaus A, Lê Cao K-A. et al. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics*.
Günther O., Lê Cao K-A. et al. (2014) Novel multivariate methods for integration of genomics and proteomics data: Applications in a kidney transplant rejection study, *OMICS: A journal of integrative biology*, 18(11), 682-95.

Generalised Canonical Correlation Analysis

Maximizes the sum of covariances between latent components associated to 2 data sets.

For J blocks of variables $\boldsymbol{X}_1(n \times p_1), \ldots, \boldsymbol{X}_J(n \times p_J)$,

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

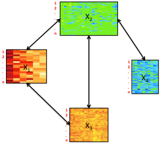
$$\text{s.t. } ||\boldsymbol{a}^j||_2 = 1 \quad \text{and} \quad ||\boldsymbol{a}^j||_1 \leq \lambda_j,$$

with $\boldsymbol{C} = \{c_{kj}\}$ the design matrix, $\boldsymbol{a}^j$ the loading vectors associated to each block j, λ_j the lasso parameter for each data set $\boldsymbol{X}_j$.

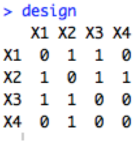
Parameters to choose in sGCCA

1. The design matrix $\boldsymbol{C}$ (user input)
2. The number of components H (cross-validation)
3. The lasso parameters $\sim$ number of variables to select on each component of each data set (cross-validation)

The design matrix $\boldsymbol{C}$ determines which pairwise covariance matrix to maximize:



is coded as



```
> design
   X1 X2 X3 X4
X1  0  1  1  0
X2  1  0  1  1
X3  1  1  0  0
X4  0  1  0  0
```

Prediction in supervised sGCC-Discriminant Analysis

The outcome to predict is the dummy matrix $\boldsymbol{Y}$.

GCC-DA models each data set X_j as:

$$Y_1 = X_1\beta_1 + E_1, \quad Y_2 = X_2\beta_2 + E_2, \quad \dots \quad Y_J = X_J\beta_J + E_J$$

β_j is the matrix of the regression coefficients for each data set X_j and defined w.r.t GCCA constraints, E_j is the residual matrix.

The prediction of a new sample X_j^{new} is:

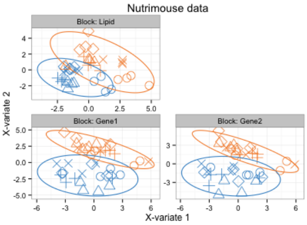
$$\hat{Y}_1 = X_1^{new}\hat{\beta}_1, \quad \hat{Y}_2 = X_2^{new}\hat{\beta}_2, \quad \dots \quad \hat{Y}_J = X_J^{new}\hat{\beta}_J$$

$\hat{\beta}_j$ obtained from the loading vectors $(\boldsymbol{a}_j^1, \boldsymbol{a}_j^2, \dots, \boldsymbol{a}_j^H)$, with H the components.

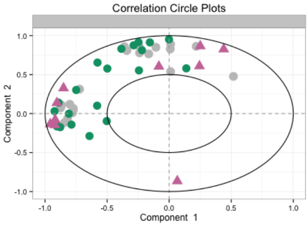
→ Prediction based on majority vote or average

Data visualisation

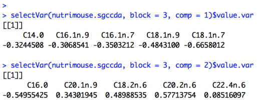
Visualisation to make sense of those large data sets by projection onto the subspace spanned by the latent components



Sample plots



Variable plots



```
>
> selectVar(nutrimouse.sgccda, block = 3, comp = 1)$value.var
[[1]]
     C14.0   C16.1n.9   C16.1n.7   C18.1n.9   C18.1n.7
-0.3244508 -0.3068541 -0.3503212 -0.4843100 -0.6658012

> selectVar(nutrimouse.sgccda, block = 3, comp = 2)$value.var
[[1]]
      C16.0    C20.1n.9    C18.2n.6    C20.2n.6    C22.4n.6
-0.54955425  0.34301945  0.48988535  0.57713754  0.08516097
```

List of selected biomarkers

Breast cancer study (The Cancer Genome Atlas)

Breast cancer is a heterogeneous disease with respect to molecular alterations, cellular composition, and clinical outcome.

- Develop tumor classifications clinically useful for prognosis or prediction
- Intrinsic classifier based on a signature of 50 genes (PAM50 classifier[1])

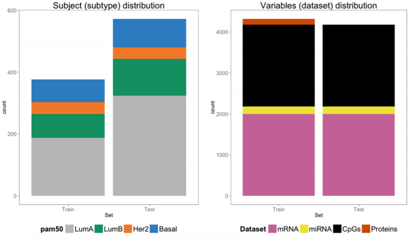
Can we expand the gene signature to other 'omics data types, increase prediction accuracy, and understand breast cancer at a systems biology level?

[1] Tibshirani R, et al. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS* **99**

Amrit Singh, University of British Columbia, Canada

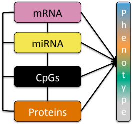
The multi 'omics data

- Four intrinsic subtypes of breast cancer luminal A, luminal B, HER2-enriched, basal-like
- Training set $n = 377$, test set $n = 573$
- mRNA, miRNA, proteomics and methylation data (up to 2,000 features each)

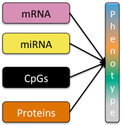


Comparisons with other methods

	Single 'omics	Multiple 'omics
Unsupervised	PCA	Concatenation + PCA
Supervised	sPLS-DA[1] eNet[2]	Concatenation + eNet/sPLS-DA Ensemble + eNet/sPLS-DA sGCC-DA null design sGCC-DA full design



Full Design



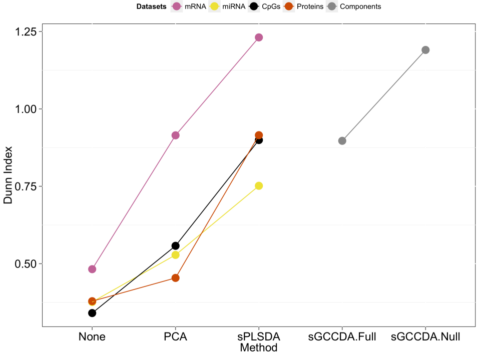
Null Design

[1] Lê Cao, K.-A. et al (2011). Sparse PLS Discriminant Analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC bioinfo*, 12(1).

[2] Zou, Hastie (2005). Regularization and Variable Selection via the Elastic Net. *JRSSB*

Unsupervised clustering to understanding the data types

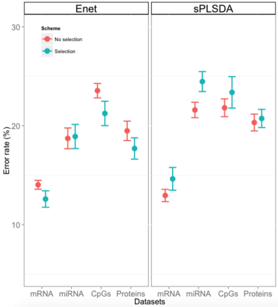
Dunn Index: evaluate clustering based on the known tumour subtypes

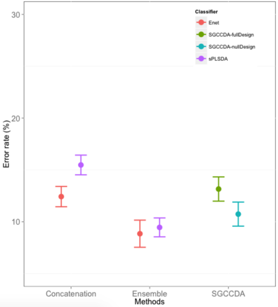


- mRNA data set clusters tumour subtypes well
- sGCCA null-design clusters as well as mRNA while integrating all 4 data sets

Kevin Chang, University of Auckland, NZ

Classification error rates on training set (50 x 5-fold CV)





Single 'omics:

- eNet >> sPLS-DA
- variable selection overlap ∼ 10-30%

Multi ‘omics:

- Ensemble > sGCC-DA
- sGCC-DA design matters for performance
- variable selection overlap ∼ 20-50%

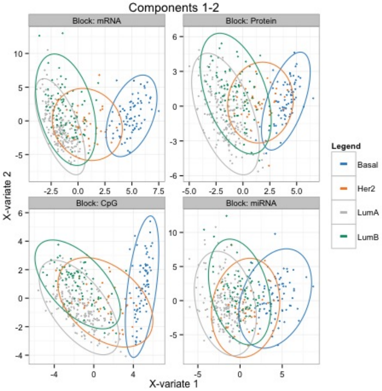
Performance of sGCC-DA with variable selection

	Basal	Her2	LumA	LumB	Overall
Training	0.00 (0.00)	11.3 (2.17)	7.71(0.84)	49.09 (2.72)	15.01 (0.76)
Test	3.23	13.51	8.64	58.82	18.50

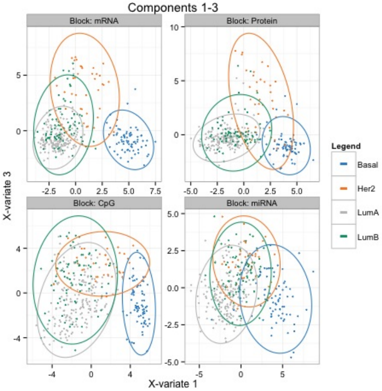
Table : Mean classification error rate based on sGCCA full design with 3 components and a selection of 20 variables per component

- Similar error rates between training and test set.
- LumB subtype difficult to classify.

Samples projected in each 'omic subspace: integration is not an easy task!



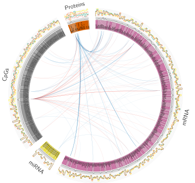
Comp 1 vs. 2



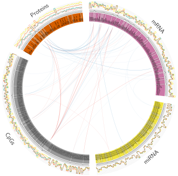
Comp 1 vs 3

Integrative methods are better at unravelling associations between variables of different types

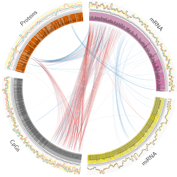
	Concatenation	Ensemble	sGCC-DA null design	sGCC-DA full design
# associations ($\|r\| > 0.6$)	752	458	1,343	1,671



Concatenation



Ensemble



sGCC-DA full design

Dr Michael Vacher, The University of Western Australia

A highly connected biomarker signature

Gene Ontology analysis: selection of 60 genes and 60 proteins highlight estrogen response pathway.

Known: Estrogen receptor can cause changes in the expression of specific genes, which can lead to the stimulation of cell growth, particularly in luminal breast cancers.

In addition,

- many oncogenic genes identified in our signatures
- mRNAs and proteins part of the estrogen response pathway are distinct
  → investigate whether those come intra and extra cellular components across data types

Dr Casey Shannon, PROOF Centre of Excellence, Vancouver

Conclusions

Multivariate linear methods enables to answer a wide range of biological questions via

- Data exploration
- Classification
- Integration of multiple data sets
- Variable selection

Multivariate methods presented here are part of the mixOmics R package dedicated to the exploration and integration of (large) biological data sets.

Integration of heterogeneous data set is a difficult challenge: this is only the beginning! (see next talks)

http://www.mixOmics.org

Acknowledgements

mixOmics development

Sébastien Déjean	Univ. Toulouse
Ignacio González	Univ. Toulouse
Francois Bartolo	Univ. Toulouse
Xin-Yi Chua	QFAB Bioinformatics
Benoît Gautier	UQDI
Florian Rohart	AIBN, UQ

Methods development

Amrit Singh	UBC, Vancouver
Casey Shannon	UBC, Vancouver
Oliver Günther	UBC, Vancouver
Kevin Chang	Univ. Auckland
Michael Vacher	Univ. Western Austra
Arthur Tenenhaus	Supelec Paris

Australian Government
Australian Research Council

Australian Government
National Health and
Medical Research Council