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

Amrit Singh[1], Benoît Gautier[2], Kim-Anh Lê Cao[2]

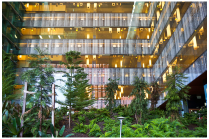
[1]The University of British Columbia, Vancouver, Canada;
[2]The University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, Australia;

Where do I live? ... and work!

Arafura Sea
Guinea
Coral Sea
NT
QLD
Australia
WA
SA
NSW
Great Australian Bight
ACT
VIC
Tasman Sea
TAS
New Zealand

In Brisbane, Australia since late 2008

In 2014 I moved to the Translational Research Institute, the Australian-first initiative of 'bench to bedside' medical research to build my own research group.



TRI partnering for better health

Outline

1. Introduction
2. Multivariate analysis for biological data
3. Integration for multiple data sets
4. Results

Outline

1 Introduction

2 Multivariate analysis for biological data

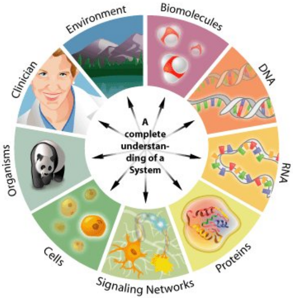
3 Integration for multiple data sets

4 Results

Systems biology is the 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 (2006)



→ an inter-disciplinary field enabling a better understanding of the entirety of processes that happen in a biological system

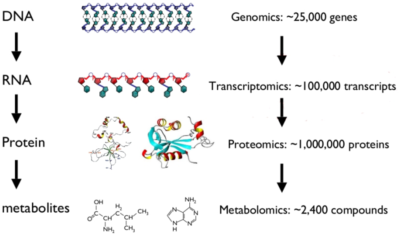
Challenges

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 or noisy variables
- # samples small << # variables
  → **statistical validation limited**
- Rely on biological interpretation
- Keep up with new technologies
- Anticipate computational issues

How to make sense of biological ‘big data’?



DNA → RNA → Protein → metabolites

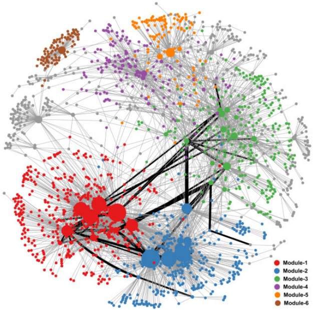
Genomics: ~25,000 genes

Transcriptomics: ~100,000 transcripts

Proteomics: ~1,000,000 proteins

Metabolomics: ~2,400 compounds

⟹



from PMID: 22548756

‘What is the key information I can extract 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

Multivariate methods (briefly) presented today

	Aims	Single ‘omics	Multiple ‘omics
Unsupervised	Data mining Exploration Correlated features	PCA	CCA (2 ‘omics) PLS (2 ‘omics) GCCA ($>$ 2 ‘omics)
Supervised	Biomarker discovery Data mining Exploration Correlated features	PLS-DA	GCC-DA ($>$ 2 ‘omics)

A bit of algebra: a linear combination of variables

$\mathbf{X}$ =

	Height	Weight
1	174.0	65.6
2	175.3	71.8
3	193.5	80.7
4	186.5	72.6
5	187.2	78.8
6	181.5	74.8
7	184.0	86.4
8	184.5	78.4
9	175.0	62.0
10	184.0	81.6

We assign two coefficients $a_1 = 0.5$ and $a_2 = 2$ to the variables Height and Weight respectively: $\boldsymbol{a} = \binom{0.5}{2}$

Merci Sébastien Déjean

A bit of algebra: a linear combination of variables

A linear combination of Height and Weight with the coefficients $a_1 = 0.5$ (associated to Height) and $a_2 = 2$ (associated to Weight) is defined as:

	Height			Weight		Linear combination
	174.0			65.6		218.20
	175.3			71.8		231.25
	193.5			80.7		258.15
	186.5			72.6		238.45
$0.5 \times$	187.2	$+$	$2 \times$	78.8	$=$	251.20
	181.5			74.8		240.35
	184.0			86.4		264.80
	184.5			78.4		249.05
	175.0			62.0		211.50
	184.0			81.6		255.20

We can write the linear combination as a matrix product:
Linear combination $= \mathbf{Xa}$, with X is a matrix of size $(n \times p)$ and $\boldsymbol{a}$ is a vector of length p

Merci Sébastien Déjean

Outline

1 Introduction

2 Multivariate analysis for biological data

3 Integration for multiple data sets

4 Results

Principal Component Analysis

PCA objective function for the first component:

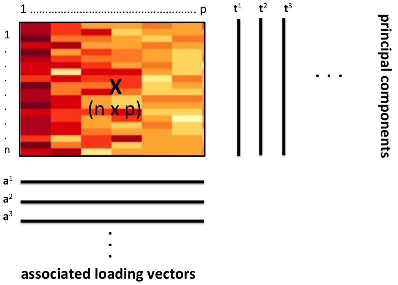
$$\max_{||\boldsymbol{a}||=1} var(\boldsymbol{Xa})$$

wihere $\boldsymbol{X}$ is a matrix $(n \times p)$, $\boldsymbol{a}$ is the loading vector of length p and $\boldsymbol{t} = \boldsymbol{Xa}$ is the first **Principal Component**.

Other Principal Components follow with the condition that they are orthogonal to each other.

Principal Component Analysis

PCA as a matrix decomposition



PCA solved with Singular Value Decomposition: $X = U\Delta A^T$

- Δ diagonal matrix with $\sqrt{\delta_h}$ (eigenvalues)
- $T = U\Delta$, T contains the PCs $\boldsymbol{t}^h$
- A contains the loading vectors $\boldsymbol{a}^h$ (eigenvectors)
- $h = 1..H$ is the number of PCs

The variance of the first principal component $\boldsymbol{t}^1$ is equal to its associated eigenvalue δ_1, and so fourth for the other PCs. The eigenvalues δ_h decrease and correspond to the explained variance per component.

Canonical Correlation Analysis

CCA objective function for the first set of variates:

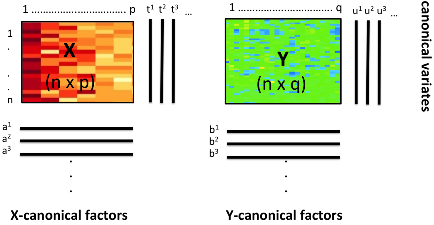
$$\arg\max_{a,\,b} \mathrm{cor}(X\boldsymbol{a}, Y\boldsymbol{b})$$

$$\text{subject to} \quad \mathrm{var}(X\boldsymbol{a}) = \mathrm{var}(Y\boldsymbol{b}) = 1,$$

where $\boldsymbol{X}$ is a matrix $(n \times p)$ and $\boldsymbol{Y}$ is a matrix $(n \times q)$, the pair of vectors $(\boldsymbol{t} = X\boldsymbol{a}, \boldsymbol{u} = Y\boldsymbol{b})$ are the **canonical variates**, and $(\boldsymbol{a}, \boldsymbol{b})$ are the associated **canonical factors**.

Other Canonical variates follow with the condition that they are orthogonal to each other.

Canonical Correlation Analysis



CCA is solution to the eigenvalues problem:

$$S_{XX}^{-1} S_{XY} S_{YY}^{-1} S_{YX} \boldsymbol{a} = \lambda^2 \boldsymbol{a},$$
$$S_{YY}^{-1} S_{YX} S_{XX}^{-1} S_{XY} \boldsymbol{b} = \lambda^2 \boldsymbol{b}.$$

- S_{XX} and S_{YY} are the sample correlation matrices of X and Y
- $S_{XY} = S'_{YX}$ are the sample cross-correlation matrix between X and Y
- $\rho = \sqrt{\lambda} = cor(\boldsymbol{a}, \boldsymbol{b})$ is the fist canonical correlation

Projection to Latent Structures

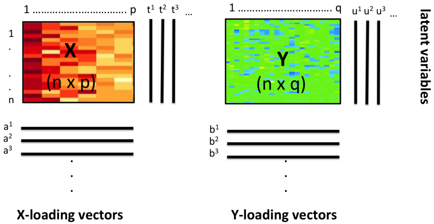
PLS objective function for the first set of variates:

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

where $\boldsymbol{X}$ is a matrix $(n \times p)$ and $\boldsymbol{Y}$ is a matrix $(n \times q)$, the pair of vectors $(\boldsymbol{t} = X\boldsymbol{a}, \boldsymbol{u} = Y\boldsymbol{b})$ are the **latent variables**, and $(\boldsymbol{a}, \boldsymbol{b})$ are the associated **loading vectors**.

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

Projection to Latent Structures



PLS can be solved via SVD:

$$X'Y = A\Lambda B'$$

- A $(p \times r)$ and B $(q \times r)$ contain the left and right singular vectors a^h and b^h (loading vectors), $h = 1, \ldots, H$, $H \leq r$, where r is the rank of the matrix $X'Y$.
- Latent variables (t, u) can be calculated as: $\boldsymbol{t} = X\boldsymbol{a}$ and $\boldsymbol{u} = Y\boldsymbol{b}$

PLS can also be solved iteratively via successive regressions of $\boldsymbol{t}$ on X and Y to maximise $cov(\boldsymbol{t}, \boldsymbol{u})$, see following slides.

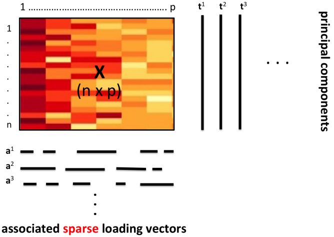
PLS-Discriminant Analysis: $\boldsymbol{Y}$ categorical response variable is coded as a dummy matrix.

Sparse multivariate analysis

High throughput biological experiments: too many variables, noisy or irrelevant.

→ clearer signal if some of the variable weights $\{a_1, \ldots, a_p\}$ were set to 0 for the 'irrelevant' variables (small weights) e.g. in PCA:

$$\boldsymbol{t} = 0 * \boldsymbol{x}^1 + a_2\boldsymbol{x}^2 + a_3\boldsymbol{x}^3 + \cdots + 0 * \boldsymbol{x}^p$$



Important weights = important contribution to define the PCs.

Null weights = those variables are not taken into account when calculating that PC.

Rank-l approximation matrix with PCA

Since PCA is solved through SVD ($X = U\Delta A^T$), the closest rank-l matrix approximation to X is:

$$X^{(l)} \equiv \sum_{h=1}^{l} \delta_h \boldsymbol{u}^h \boldsymbol{a}^{h'}.$$

Therefore, the best rank-1 approximation of X in terms of Frobenius norm* is:

$$\min_{\boldsymbol{t},\boldsymbol{a}} ||X - \boldsymbol{t}\boldsymbol{a}^{'}||_F^2$$

when $\boldsymbol{t} = \delta_1 \boldsymbol{u}^1$ and $\boldsymbol{a} = \boldsymbol{a}^1$.

*The Frobenius norm between X and $X^{(l)}$ is defined as:
$||X - X^{(l)}||_F^2 = \text{trace}\{(X - X^{(l)})(X - X^{(l)})^T\}$.

Solving sparse PCA

In PCA, $\boldsymbol{a}$ can also be solved via a least square regression of a fixed component t on X:

$$\boldsymbol{t} = X\boldsymbol{a} + \epsilon.$$

Therefore LASSO penalization λ can be introduced such that

$$\min_{\lambda} \sum_{i=1}^{n} (t_i - x_i \boldsymbol{a})^2 + \lambda \sum_{j=1}^{p} |a_j|.$$

The objective function of sPCA can be written as

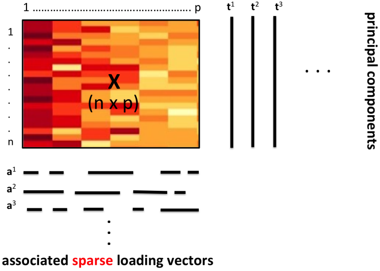
$$\min_{\boldsymbol{t},\boldsymbol{a}} ||X - \boldsymbol{t}\boldsymbol{a}^T||_F^2 + P_{pen}(\boldsymbol{a}), \quad s.t. \quad ||\boldsymbol{a}|| = 1.$$

In practice P_{pen} is a soft thresholding function that approximates the LASSO.

sparse loadings vectors in PCA

sPCA is solved iteratively via the algorithm Non Linear Iterative Partial Least Squares (NIPALS, Wold 1987):

- remove irrelevant variables when calculating the principal components,
- perform internal variable selection.



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*.

Regularized CCA

When $n << p$ and $n << q$ S_{XX} and S_{YY} are singular and ill-conditioned. → CCA leads to unreliable results.

Solution: regularization of the correlation matrices in CCA:

$$S_{XX}(\tau_1) = S_{XX} + \tau_1 \mathbb{1}_p$$
$$S_{YY}(\tau_2) = S_{YY} + \tau_2 \mathbb{1}_q \,,$$

where τ_1 and τ_2 are non-negative numbers, estimated with cross-validation[1] or shrinkage method[2].

[1] González I. et al., 2009. Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis.*Journal of Biological Systems* **17**(2).

[2] Schäfer and Strimmer (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *SAGMB*, **4**(1).

Rank-l approximation matrix with PLS

In the same vein as sPCA, PLS is solved through SVD ($X^TY = A\Lambda B^T$) and the best rank-1 approximation of X^TY is:

$$\min_{\boldsymbol{a},\boldsymbol{b}} ||X^TY - \boldsymbol{a}^T\boldsymbol{b}||_F^2$$

In PLS, the loading vectors $\boldsymbol{a}, \boldsymbol{b}$ can also be solved through successive least squares regressions of t on X and Y:

- Repeat until convergence of $\boldsymbol{u}$:
  1. $\boldsymbol{a} = X^T\boldsymbol{t}/\boldsymbol{t}^T\boldsymbol{t}$, norm $\boldsymbol{a}$
  2. $\boldsymbol{t} = X\boldsymbol{a}/\boldsymbol{a}^T\boldsymbol{a}$
  3. $\boldsymbol{b} = Y^T\boldsymbol{t}/\boldsymbol{t}^T\boldsymbol{t}$, norm $\boldsymbol{b}$
  4. $\boldsymbol{u} = Y\boldsymbol{b}/\boldsymbol{b}^T\boldsymbol{b}$

→ introduce LASSO penalisations on both $\boldsymbol{a}$ and $\boldsymbol{b}$!

Rank-l approximation matrix with PLS

The objective function of sPLS can be written as

$$\min_{a,b} ||X^T Y - a^T b||_F^2 + P_{pen}(a) + P_{pen}(b), \quad s.t. \quad ||a|| = 1, ||b|| = 1.$$

In practice P_{pen} is a soft thresholding function to approximate the LASSO penalisations:

- simultaneous sparse loadings a and b for each set of PLS components.
- selected variables from both data sets are correlated across samples.

Lê Cao et al (2008). A sparse PLS for variable selection when integrating omics data. *SAGMB* **7**.

Outline

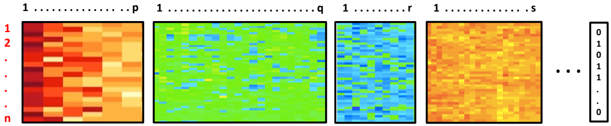
1 Introduction

2 Multivariate analysis for biological data

3 Integration for multiple data sets

4 Results

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 are sample matched
- Select relevant biological features that are correlated within across heterogeneous data sets
- Extend sPLS, sPLS-DA (new!) and rCCA

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

For J blocks of variables $\boldsymbol{X}_1, \dots, \boldsymbol{X}_J$ of size $(n \times p)$, $(n \times q)$, $\dots$, GCCA optimizes the problem:

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

with the constraints

for regularised GCCA: $\tau_j ||\boldsymbol{a}^j||_2 + (1 - \tau_j) \text{Var}(\boldsymbol{X}_j \boldsymbol{a}^j) = 1$

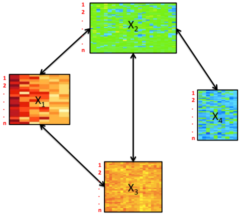
or

for sparse GCCA: $||\boldsymbol{a}^j||_2 = 1$ and $||\boldsymbol{a}^j||_1 \leq \lambda_j$

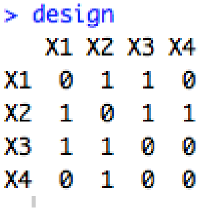
$\boldsymbol{C} = \{c_{j,k}\}$ is the design matrix
$\boldsymbol{a}^j$ are the loading vectors associated to each block j,
τ_j is the regularization parameter on each data set j
λ_j is the lasso parameter on each data set j, $j = 1, \dots, J$

The design matrix C in GCCA

The design to 'link' the datasets (link == covariance is maximised) has an impact:



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
```

Parameters to tune

How to best choose the GCCA parameters?

- The number of components H
- The design matrix C
- rGCCA: Regularization parameters τ_j for each covariance matrix from each data set→ shrinkage method
- sGCCA: Number of variables to select on each component of each data set (instead of Lasso parameters λ_j^h) → cross-validation

Prediction in GCC-DA

Let's go back one step with the simple PLS-DA model where Y is a categorical response vector coded as a dummy matrix.

The PLS-DA model is formulated as:

$$Y = X\beta + E,$$

where β is the matrix of the regression coefficients and E is the residual matrix.

The prediction of a new sample is then:

$$\hat{Y} = X_{new}\hat{\beta},$$

where $\hat{\beta}$ is directly obtained from the loading vectors $(\boldsymbol{a}^1, \boldsymbol{a}^2, \dots, \boldsymbol{a}^H)$, where H is the chosen PLS dimension and X_{new} data matrix of a new sample.

$\hat{Y}$ is a continuous numerical value (not a class number!)
→ we use distances to obtain the class prediction.

Prediction in GCC-DA

In GCCA we model each data set X_j as:

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

with the GCCA constraints and the maximisation of the covariance btw components of each data set.

The prediction of a new sample is then for each type of data:

$$\hat{Y}_1 = X_{new}\hat{\beta}_1, \quad \hat{Y}_2 = X_{new}\hat{\beta}_2, \dots, \hat{Y}_J = X_{new}\hat{\beta}_J$$

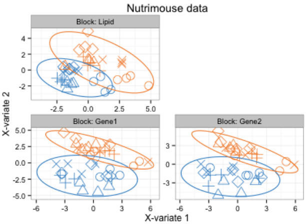
where each $\hat{\beta}_j$ are obtained from the set of loading vectors $(\boldsymbol{a}^1, \boldsymbol{a}^2, \dots, \boldsymbol{a}^H)$, with H the chosen GCCA dimension and X_{new} data matrix of a new sample.

To obtain the final prediction of a new sample:

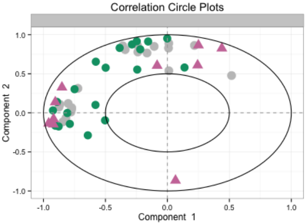
- we use distances on either the average of all $\hat{Y}_j$ or
- we take the majority vote of all predictions from all data sets

What is there for our fellow biologists?

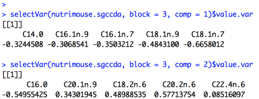
Visualisations to make sense of those large data sets.



Using components to project samples in their own subspace



$cor(X_j, X_j \boldsymbol{a}_j^h)$ projects the variables on each h component $\boldsymbol{t}^{j,h} = X_j \boldsymbol{a}_j^h$



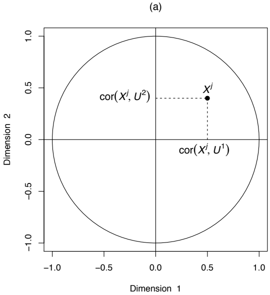
```
>
> selectVar(nutrimouse.sgccda, block = 3, comp = 1)$value.var
[[1]]
     C14.0   C16.1n.9   C16.1n.7   C18.1n.9   C18.1n.7
-0.3244508 -0.3068541 -0.3503212 -0.4843100 -0.6658012

> 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
```

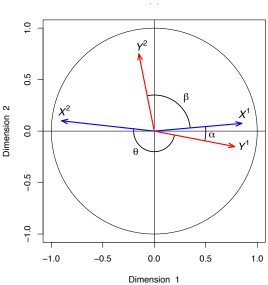
List of biomarkers of different molecular types

What **more** is there for our fellow biologists?

Correlation circle plots to understand the relationships between those large biological data sets



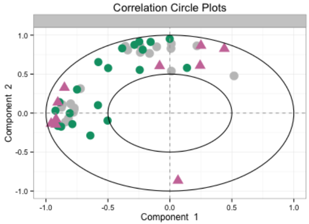
Project variables on the components $(\boldsymbol{t}^1, \boldsymbol{t}^2)$: $(cor(X, \boldsymbol{t}^1), cor(X, \boldsymbol{t}^2))$



Project X and Y variables on the components $(\boldsymbol{t}^{1,1}, \boldsymbol{t}^{1,2})$ and $(\boldsymbol{t}^{2,1}, \boldsymbol{t}^{2,2})$:
$(cor(X, \boldsymbol{t}^{1,1}), cor(X, \boldsymbol{t}^{1,2}))$ and
$(cor(Y, \boldsymbol{t}^{2,1}), cor(Y, \boldsymbol{t}^{2,2}))$

Correlation Circle plots when integrating different types of variables

Correlation circle plots generalised to more than 2 types of variables



Project X_j selected variables on their components $(X_j \boldsymbol{a}^{j,1}, X_j \boldsymbol{a}^{j,2})$ with coordinates $(cor(X_j, \boldsymbol{t}^{j,1}), cor(X_j, \boldsymbol{t}^{j,2}))$

- Different types of variables projected in comparable spaces*
- Enables visualisation of strong positive and negative correlations
- To put in relation with sample plots

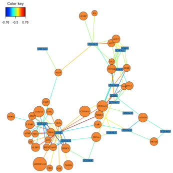
* assuming we have maximised the covariance between components

Bipartite relevance networks

Define similarity between different types of variables using components as intermediate steps:

$$sim(X_j^l, X_k^m) \simeq \sum_{h=1, j \neq k}^{H} \mathrm{cor}(X_j^l, \boldsymbol{t}^{j,h})\mathrm{cor}(X_k^m, \boldsymbol{t}^{k,h})$$

- Efficient to compute
- In rCCA and sPLS showed to unravel 'true' correlations in simulated data*
- Assumption: cov or cor btw components is maximal
- Similarity matrix is input into network visualisation



*González I., Lê Cao K.-A., et al (2012) Visualising association between paired 'omics' data sets. *J. Data Mining*.

Outline

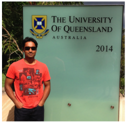
1 Introduction

2 Multivariate analysis for biological data

3 Integration for multiple data sets

4 Results

Context



PhD project of Amrit Singh (UBC Vancouver), who came for a 3-month scientific visit to UQDI in 2014 as part of his Ph.D project to integrate multiple ‘omics data sets.

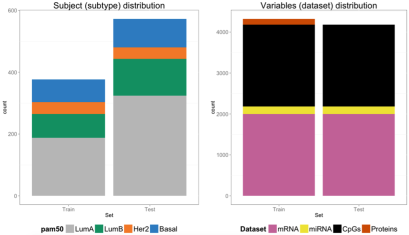
Breast cancer is a heterogeneous disease with respect to molecular alterations, cellular composition, and clinical outcome.

- challenge in developing tumor classifications that are clinically useful with respect to prognosis or prediction
- intrinsic classifier based on a signature of 50 genes (PAM50 classifier[1])

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

Multi ‘omics Breast cancer study from The Cancer Genome Atlas

- Four intrinsic subtypes luminal A, luminal B, HER2-enriched, basal-like
- training set $n = 377$, test set $n = 573$
- mRNA, miRNA, proteomics and methylation data with max 2,000 features (mRNA without the PAM50 genes!)



Comparisons with other methods

	Single 'omics	Multiple 'omics
Unsupervised	PCA	
Supervised	sPLS-DA[1] eNet[2]	Concatenation[3] + eNet/sPLS-DA Ensemble[4] + eNet/sPLS-DA sGCC-DA null design sGCC-DA full design

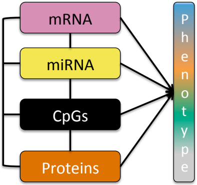
[2] elastic net: regularized regression method that linearly combines l_1 (lasso) and l_2 (ridge) penalties.
[3] concatenate all 'omics data sets;
[4] apply eNet/sPLS-DA classifier on each data set separately and combine the different lists of selected variables.

[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).

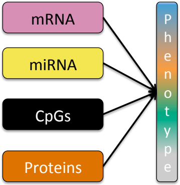
[2] Zou, Hastie (2005). Regularization and Variable Selection via the Elastic Net.

Comparisons with other methods

	Single ‘omics	Multiple ‘omics
Unsupervised	PCA	
Supervised	sPLS-DA[1] eNet[2]	Concatenation[3] + eNet/sPLS-DA Ensemble[4] + eNet/sPLS-DA sGCC-DA null design sGCC-DA full design



Full Design

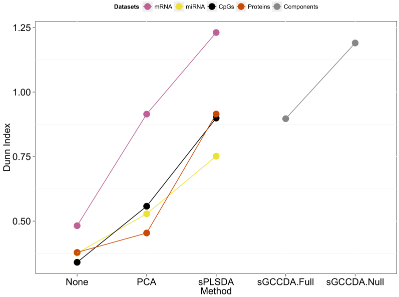


Null Design

Understanding the data: clustering

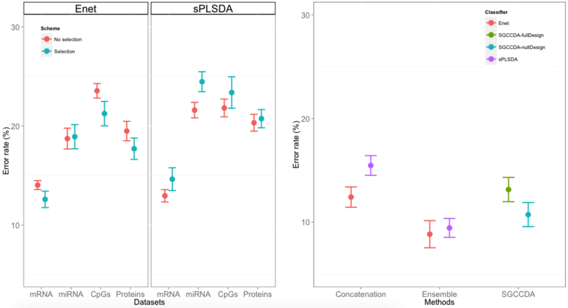
Dunn Index is a metric to evaluate clusterings - here based on the known tumour subtypes.

Calculated based on 3 components for each method with Euclidian distance.



The mRNA data set clusters tumour subtypes well.
sGCC-DA null design clusters as well as mRNA while integrating all 4 data sets.

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



Left: eNet generally performs better than sPLS-DA; variable selection overlap $\sim$ 10-30%

Right: Ensemble performs bettter than sGCC-DA; design matters in performance; variable selection overlap $\sim$ 20-50%

Performance of sGCC-DA on list of 60 features per ‘omic

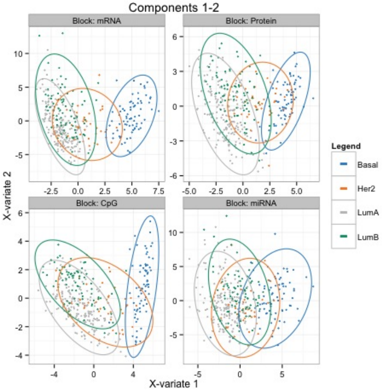
Mean classification error rate based on a sGCC-DA model with 3 components and a selection of 20 variables per component* (training: 50 x 5-fold cross-validation):

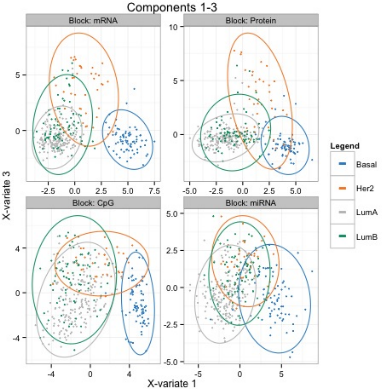
	Basal	Her2	LumA	LumB	Overal error rate
Training set	0.00 (0.00)	11.3 (2.17)	7.71(0.84)	49.09 (2.72)	15.01 (0.76)
Test set	3.23	13.51	8.64	58.82	18.50

- Similar error rates between training and test set.
- LumB subtype difficult to classify. May need to add extra components in sGCC-DA.

* Note: optimal tuning not performed yet

Samples projected in each 'omic subspace spanned by the components: integration is not an easy task!





Fun part omitted: representing the ellipse from the training set and the test samples as dots.

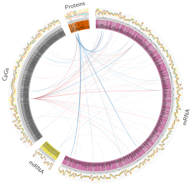
Integrative methods are more efficient at unravelling associations between variables of different types

	Concatenation	Ensemble	sGCC-DA null design	sGCCDA full design
associations	752	458	1,343	1,671

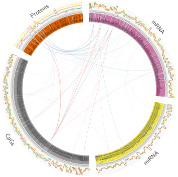
Number of associations are determined as the number of pair-wise correlation (Pearson) $|r| > 0.6$.

The total number of selected variables is the same in each method ($\sim$ 390 features).

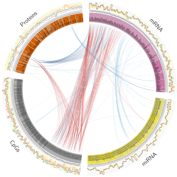
Relevance networks based on Pearson's correlation



Concatenation



Ensemble



sGCC-DA full design

- Network based on circos plot representing only inter correlations.
- Similarities based on components not calculated here (not feasible with the eNet approach for Concatenation and Ensemble).

Dr Michael Vacher, The University of Western Australia

Preliminary Gene Ontology analysis on selected features

Lists of 60 genes and 60 proteins selected on the training set appears the 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.

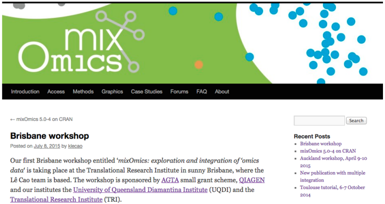
In addition,

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

Dr Casey Shannon, PROOF Centre of Excellence, Vancouver, Canada

It is all about mixOmics

`mixOmics` is not only an R package, it is also (finally!) part of a research program!



← mixOmics 5.0-4 on CRAN

Brisbane workshop

Posted on July 8, 2015 by klecao

Our first Brisbane workshop entitled '*mixOmics: exploration and integration of 'omics data*' is taking place at the Translational Research Institute in sunny Brisbane, where the Lê Cao team is based. The workshop is sponsored by AGTA small grant scheme, QIAGEN and our institutes the University of Queensland Diamantina Institute (UQDI) and the Translational Research Institute (TRI).

Recent Posts

- Brisbane workshop
- mixOmics 5.0-4 on CRAN
- Auckland workshop, April 9-10 2015
- New publication with multiple integration
- Toulouse tutorial, 6-7 October 2014

Website with tutorials
www.mixOmics.org

- Most GCCA approaches recoded, improved and implemented in the R package `mixOmics`
- More to come (visualisation, other super cool features)

To put it in a nutshell

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

- data exploration
- classification
- integration of multiple data sets
- variable selection

Coming up in `mixOmics`:

- 16S data analysis
- Integration of time course data
- Meta analysis / multi group analysis

Acknowledgements

mixOmics development	
Sébastien Déjean	Univ. Toulouse
Ignacio González	Univ. Toulouse
Francois Bartolo	Univ. Toulouse
Xin Yi Chua	QFAB
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

Questions, feedback?

Okay – wait. What were we studying, again?
Other
Don't know
Streeter
© Original Artist / Search ID: bstn474
Rights Available from CartoonStock.com
CARTOONSTOCK .COM
www.betsystreeter.com

mixomics@math.univ-toulouse.fr

Register to our newsletter for the latest updates
http://mixomics.org/a-propos/contact-us/

http://www.mixOmics.org