



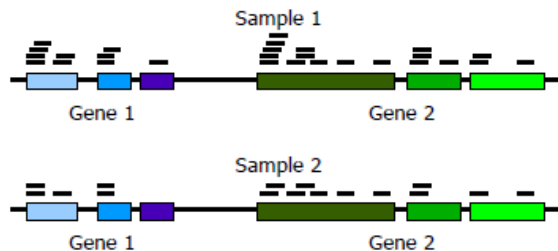
Model-based clustering of genes expression data with external annotations

Méline Gallopin, Gilles Celeux, Florence Jaffrézic, Andrea Rau

Université Paris Sud 11
INRA, Jouy-en-Josas





RNA-seq data



Gene	E1	E2	E3	E4	E5	E6
13CDNA73	4	0	6	1	0	5
A2BP1	19	18	20	7	1	8
A2M	2724	2209	13	49	193	548
A4GALT	0	0	48	0	0	0
AAAS	57	29	224	49	202	92
AACS	1904	129	4	507	3	5365
AADACL1	3	13	239	683	158	40
[...]						

Analysis of RNA-seq data

Methods developed and R packages

	Microarray (1995) <i>continuous data</i>	NGS (2008) <i>count data</i>
Differential analysis	 Gaussian : limma <i>(Smyth & al, 2005)</i>	
Network inference	Gaussian : SIMoNe <i>(Chiquet & al, 2010)</i>	
Model-based clustering	Gaussian : Rmixmod <i>(Biernacki & al, 2006)</i>	

Analysis of RNA-seq data

Methods developed and R packages



NGS (2008)
count data



Microarray (1995)
continuous data



Differential analysis

Gaussian : limma
(Smyth & al, 2005)

Negative Binomial : DESeq ; EdgeR
(Anders & al, 2010 ; Robinson & al, 2010),

Network inference

Gaussian : SIMoNe
(Chiquet & al, 2010)

Poisson ; Poisson log-normal
(Allen & al, 2012 ; Gallopín & al, 2013)

Model-based clustering

Gaussian : Rmixmod
(Biernacki & al, 2006)

Poisson : HTScluster
(Rau & al, submitted)

Model-based clustering of genes expression data , with external annotations

Analysis of RNA-seq data

Methods developed and R packages



NGS (2008)
count data

Microarray (1995)
continuous data

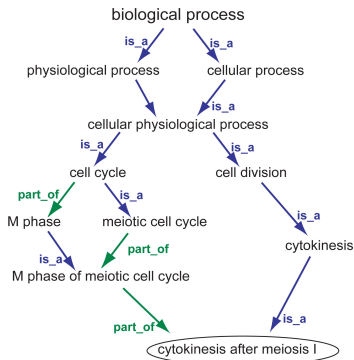


Differential analysis	<p>Gaussian : limma (Smyth & al, 2005)</p>	<p>Negative Binomial : DESeq ; EdgeR (Anders & al, 2010 ; Robinson & al, 2010), OR Gaussian on transformed data : limma + "voom" (Law & al, 2014)</p>
Network inference	<p>Gaussian : SIMoNe (Chiquet & al, 2010)</p>	<p>Poisson ; Poisson log-normal (Allen & al, 2012 ; Gallopin & al, 2013) OR Gaussian on transformed data ?</p>
Model-based clustering	<p>Gaussian : Rmixmod (Biernacki & al, 2006)</p>	<p>Poisson : HTSCluster (Rau & al, submitted) OR Gaussian on transformed data ?</p>

Model-based clustering of genes expression data , with external annotations

External information on genes functions

Gene Ontology terms (GO terms)



Example : GO annotations for the RPS6KA2 gene

- ▶ regulation of meiosis (GO :0040020)
- ▶ stress-activated MAPK cascade (GO :0051403)
- ▶ neurotrophin TRK receptor signaling pathway (GO :0048011)
- ▶ ...

Taking external knowledge into account in model-based clustering



- Option 1 :** Use external gene annotation only to validate clusters
⇒ *does not use external information in clusters estimation*
- Option 2 :** Incorporate external gene annotation into the mixture model
⇒ *use external information... to what extent ?*

Our solution :

Include external gene annotation in model selection only

Model-based clustering of genes

Microarray or RNA-seq gene expression data



- ▶ $\mathbf{y}_{(n \times d)}$ matrix of expression data

	sam. 1	sam.2	sam.3	...	sam.j	...
gene 1	478	425	718		.	
gene 2	15	0	86		.	
gene 3	678	875	767		.	
gene 4	3	0	0		.	
gene 5	13878	20078	19082		.	
...					.	
gene i	y_{ij}	.
					.	

- ▶ Mixture of K components : $f(\mathbf{y}_i; K, \Theta_K) = \sum_{k=1}^K p_k f_k(\mathbf{y}_i, \mathbf{a}_k)$
- ▶ Latent variable : \mathbf{z} matrix of size $n \times K$
 - $z_{ik} = 1$ if gene i comes from component k
 - $z_{ik} = 0$ otherwise.

Model selection for mixture model

We select the number of clusters $K = 1, \dots, K^{max}$ that maximizes one of the following criteria :

- Bayesian Information Criterion (BIC) : Schwarz, 1978

$$\text{BIC}(K) = \log f(\mathbf{y}|\hat{\theta}) - \frac{\nu_K}{2} \log(n)$$

- Integrated Completed Likelihood (ICL) : Biernacki et al., 2000

$$\text{ICL}(K) = \text{BIC}(K) + \sum_{i,k} \hat{z}_{ik} \log(\hat{t}_{ik})$$



- Taking into account an external classification : Baudry et al., 2014

$$\text{SICL}(K) = \text{ICL}(K) + \sum_{l=1}^U \sum_{k=1}^K n_{kl} \log\left(\frac{n_{kl}}{n_k}\right)$$

GO annotation variables



\mathbf{go}^m : vector summarizing the gene annotations for a given GO term m

	\mathbf{y}	\mathbf{go}^1	\mathbf{go}^2
gene 1	1	0
gene 2	1	0
gene 3	1	1
gene 4	0	0
gene 5	0	0

gene i	. . y_{ij} . .	u_i^1	u_i^2

GO annotations :

$$\mathbf{go}_i^m = \begin{cases} 1 & \text{if gene } i \text{ is annotated for the GO term } m \\ 0 & \text{if gene } i \text{ is not annotated for the GO term } m \\ 0 & \text{if we do not have any information (missing data)} \end{cases}$$

Taking GO annotations into account



Allocation of annotated genes \mathbf{a}^m : $n \times K$ matrix

$$a_{ik}^m = \begin{cases} 1 & \text{with probability } p_k^m \text{ if } go_i^m = 1 \\ 0 & \text{if } go_i^m = 0 \end{cases}$$

Our model selection criterion **ICaL** :

$$\log f(\mathbf{y}, \mathbf{z}, \mathbf{a}^m; K) = \log \int_{\theta_K} f(\mathbf{y}, \mathbf{z}, \mathbf{a}^m; K, \theta_K) \pi(\theta | K) d\theta$$

$$\simeq \text{ICL}(K) + \sum_{k=1}^K n_k^m \log\left(\frac{n_k^m}{n^m}\right)$$

$$n^m = \text{card}\{i : u_i^m = 1\}$$

$$n_k^m = \text{card}\{i : z_{ik} = 1 \text{ et } u_i^m = 1\}$$

Model selection criterion ICaL in practice

$$\text{ICaL}(K) = \text{ICL}(K) + \sum_{m=1}^M \sum_{k=1}^K n_k^m \log\left(\frac{n_k^m}{n^m}\right)$$



Example : ICaL and SICL penalty terms for 4 GO terms for a given clustering of 500 genes in 5 clusters

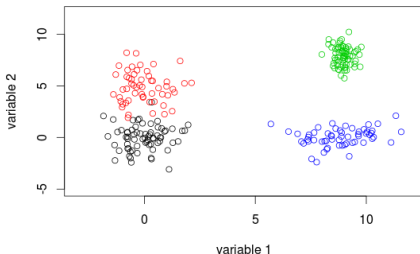
n_k^m contingency table	go ¹	go ²	go ³	go ⁴
cluster $k = 1$ (100 genes)	50	48	10	25
cluster $k = 2$ (100 genes)	0	2	10	25
cluster $k = 3$ (100 genes)	0	0	10	0
cluster $k = 4$ (100 genes)	0	0	10	0
cluster $k = 5$ (100 genes)	0	0	10	0
pen. _{ICaL} $\sum_{k=1}^K n_k^m \log\left(\frac{n_k^m}{n^m}\right)$	0	-8.4	-80.4	-34.7
pen. _{SICL} $\sum_{l=1}^U \sum_{k=1}^K n_{kl} \log\left(\frac{n_{kl}}{n_k}\right)$	-69.3	-235.5	-162.5	-112.5

Numerical example



Simulated dataset :

- ▶ 200 genes
- ▶ 2 samples
- ▶ 4 gaussian components



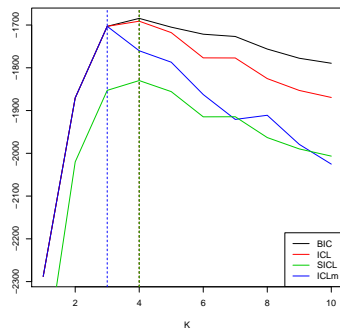
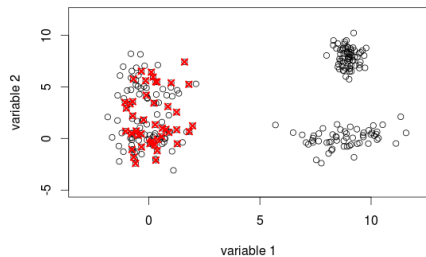
On 100 replicated datasets :

K	1	2	3	4	5	6	7	8
BIC	0	0	19	81	2	0	0	0
ICL	0	0	53	47	0	0	0	0

Model-based clustering of genes expression data , with external annotations

Relevant GO annotation

Shared by genes in components 1 and 2

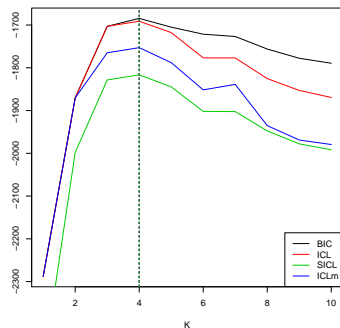
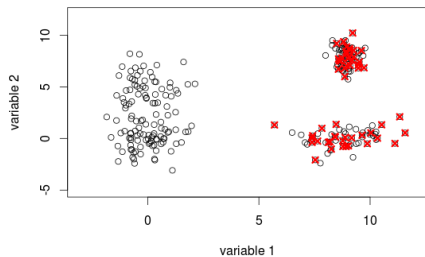


K	1	2	3	4	5	6	7	8
BIC	0	0	19	81	2	0	0	0
ICL	0	0	53	47	0	0	0	0
SICL	0	0	51	49	0	0	0	0
ICaL	0	0	100	0	0	0	0	0

Model-based clustering of genes expression data , with external annotations

Non-relevant GO annotation

Shared by genes in components 3 and 4



K	1	2	3	4	5	6	7	8
BIC	0	0	19	81	2	0	0	0
ICL	0	0	53	47	0	0	0	0
SICL	0	0	53	47	0	0	0	0
ICaL	0	0	53	47	0	0	0	0

Model-based clustering of genes expression data , with external annotations

Real datasets analysis

Expression Differences along Porcine Small Intestine Evidenced by Transcriptome Sequencing, Mach & al, 2014



- ▶ **Differential expression analysis** between the 3 tissues sequenced by RNA-seq in 4 healthy piglets (using R package EdgeR with an adjustment for any baseline differences between piglets and FDR control fixed to 0.05 %)
- ▶ **Clustering** of the 1844 differentially expressed genes using Gaussian mixture model (using R package Rmixmod with 10 initialisations)
- ▶ **Genes annotations** from the MSigDB database (*Liberzon, 2011*)
⇒ we selected 6 GO terms from the BP ontology : *carboxylic acid metabolic process, lipid metabolic process, organic acid metabolic process, nitrogen compound metabolic process, response to chemical stimulus and cellular lipid metabolic process*
- ▶ Model selection performed with BIC, ICL, SICL and ICaL

Criterion comparison

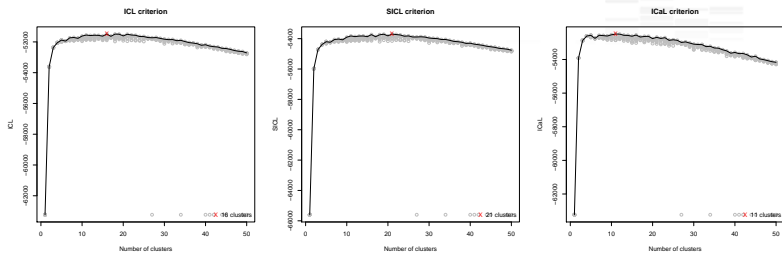


Figure: ICL, SICL et ICaL selected models

- ▶ ICL model : 16 clusters → 3 associations between GO and clusters
 - ▶ SICL model : 21 clusters → 5 associations between GO and clusters
 - ▶ ICaL model : 11 clusters → 6 associations between GO and clusters
- ⇒ cluster 11 in the ICaL model might be of particular interest.

Conclusion



On model selection for model-based gene clustering

- ▶ improve the existing model-based clustering methods with external annotations
- ▶ with the guarantee that those external annotations do not damage the clustering





General conclusion

- ▶ more work on finding the best transformation for RNA-seq for clustering and network inference
- ▶ more work on network inference on modules of genes



Thanks for your attention



-  Mach, N., Berri, M., Esquerré, D., Chevaleyre, C., Lemonnier, G., Billon, Y., Lepage, P., Oswald, I., Doré, J., Rogel-Gaillard, C., Estellé, J. (2014) *Extensive Expression Differences along Porcine Small Intestine Evidenced by Transcriptome Sequencing*. PLoS ONE 9(2) :e88515.
-  Liberzon A1, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. (2011) *Molecular signatures database (MSigDB) 3.0.*. Bioinformatics. 2011 Jun 15 ;27(12) :1739-40.
-  Biernacki, C., Celeux, G., Govaert, G. (2000) *Assessing a mixture model for clustering with the Integrated Classification Likelihood*. IEEE Transaction on PAMI, 22, 719-725.
-  Baudry, J-P., Cardoso, M., Celeux, G. , Amorim, M-J. , Sousa Ferreira, A. (2014) *Enhancing the selection of a model-based clustering with external qualitative variables*. Advances in Data Analysis and Classification (ADAC).