



**Comparaison et combinaisons de *méthode de sélection d'attributs*** (pour l'analyse du transcriptome)

**Antoine Cornuéjols<sup>1</sup>**

**J.-P. Comet<sup>3</sup>, M. Dutreix<sup>2</sup>, Ch. Froidevaux<sup>1</sup>**

**J. Mary<sup>1</sup>, G. Mercier<sup>2</sup>**

<sup>1</sup>LRI (Orsay) - <sup>2</sup>Institut Curie (Orsay) - <sup>3</sup>LAMI (Evry)

[antoine@lri.fr](mailto:antoine@lri.fr), <http://www.lri.fr/~antoine>



# Plan

- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion

- 1- Illustration**
- 2- Le problème de la sélection d'attributs**
- 3- L'approche classique**
- 4- Combiner des méthodes**
- 5- Comparaison**
- 6- Combinaison**
- 7- Conclusion**



## *Illustration : un pb d'analyse du transcriptome*

- Illustration

- Le pb de la sélection d'attributs

- Méthode standard

- Combiner des méthodes

- Comparaison

- Combinaison

- Conclusion

- Projet INRS, Bioingénierie 2001

- [2001-2004]

*Étude de l'effet des très faibles radiations*



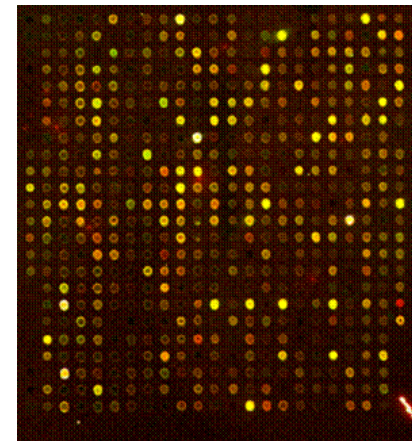


## Protocole expérimental

### • Illustration

- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion

- *S. Cerevisiae* en croissance exponentielle (séquencée complètement et eucaryote avec peu de gènes).
- **Six** cultures (Irradiées **I**) exposées pendant 20 heures entre 15 et 30 mGy/h
- **Douze** cultures non exposées (Non Irradiées **NI**)
- Mesure effectuées sur puce Corning où l'hybridation a été faite avec double marquage fluorescent (Cy3 pour les cADN contrôles et Cy5 pour les cADN étudiés).





# Questions des biologistes

## • Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

• Conclusion

- L'irradiation à de faibles doses est-elle **déTECTABLE** ?
- **Nombre de gènes** impliqués dans la réponse à une irradiation à faible dose ?
- **Groupes de gènes** impliqués dans la réponse à l'irradiation et de quelle manière ?
- Est-il possible de **deviner le traitement** subi par une levure en regardant l'expression de son génome ?
- Peut-on **généraliser cette approche** à d'autres types de traitements (pollutions, cancer, ...)



## « Précarité » des données

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

• Conclusion

- **Extrêmement peu de données / dimension**  
(12 - (non irradiées) & 6 + (irradiées) vs. 6135 gènes)
- **Données imparfaites**
  - Bruit expérimental
    - Irradiation
    - Puces à ADN
  - Prétraitement et normalisation
- **Pas idéales :**
  - Déséquilibre des classes + et -
  - Absence d'indépendance conditionnelle entre les gènes



## Le problème de la sélection d'attributs

- **Problème NP-difficile**
- **Mais a priori plus simple** que celui de la classification (apprentissage de la relation de dépendance)
- **E.g. Supposons 3 attributs binaires et fonctions booléennes**

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

• Conclusion

a1	a2	a3	XOR
0	0	0	-
0	0	1	+
0	1	0	+
0	1	1	-
1	0	0	-
1	0	1	+
1	1	0	+
1	1	1	-

$$2^{2^3} = 2^8 = 256$$

fonctions possibles

Mais seulement :  
10 tris possibles  
sur les attributs  
(e.g. (a1,a2,a3))





## *Le problème de la sélection d'attributs (2)*

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

• Conclusion

- Pourtant **il manque une théorie** fournissant des **garanties sur la qualité des classements** (analogue à théorie statistique de l'apprentissage)

- Pas d'équivalent du risque empirique
- Tâche non supervisée

➡ Méthodes (essentiellement) de nature **heuristique**



# Définitions de la « pertinence »

[Blum & Langley, 97], [Bell & Wang, 00]

## ■ Pas de définition unique car dépend du domaine

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

• Conclusion

### ■ Par rapport à la cible

- $d_i$  est pertinent si  $\exists$  une paire d'exemples ne différant qu'en  $d_i$  et de classes différentes

### ■ Idem par rapport à la distribution (ou à l'échantillon)

- Idem, sauf que la paire d'exemples peut être tirée avec une probabilité non nulle (ou appartient à l'échantillon)

### ■ Faible pertinence

- Si pertinent quand on retire un sous-ensemble des attributs

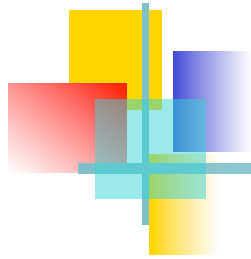
■ ...



*Pertinent si permet une meilleure classification*



*... si permet de comprendre mieux*



# Les approches

[Blum & Langley, 97]  
[Guyon & Elisseeff, 03]

## 1. Approche directe (« embedded »)

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

• Conclusion

## 2. « Wrapper methods »

■ Utilisent la performance en aval pour sélectionner les attributs

■ Deux stratégies

■ *Ascendante* (« forward selection »)

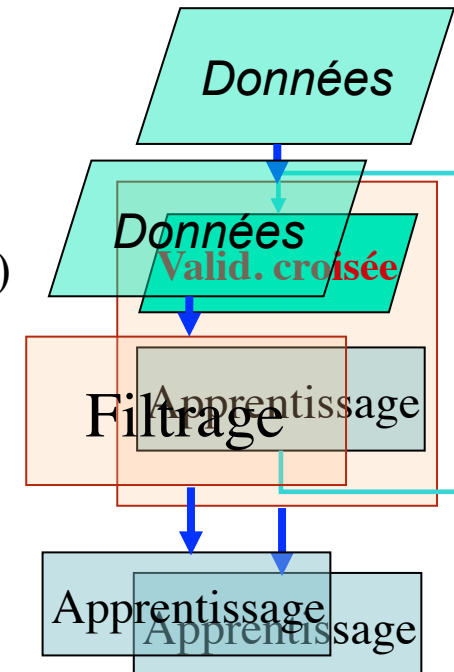
■ Par ajouts successifs d'attributs

■ *Descendante* (« backward selection »)

■ Par retraits successifs d'attributs

## 3. « Filter methods »

■ Indépendantes des traitements aval





# Comparaison

---

- Illustration

- Le pb de la sélection d'attributs

- Méthode standard

- Combiner des méthodes

- Comparaison

- Combinaison

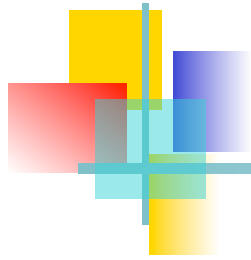
- Conclusion

- « **Filter methods** »

- Peu coûteuses

- « **Wrapper methods** »

- Coûteuses
- Plus précises ?



# Evaluation des attributs

[Guyon & Elisseeff, 03]

## ■ Hypothèse de linéarité

## ■ Critères de performance

### ■ Mesurer la **corrélation** entre un attribut et la classe

#### ■ Corrélation de Pearson (Critère de Fisher, T-test, ...)

- Détecte uniquement les dépendances linéaires entre variables

$$\mathcal{R}_i = \frac{\text{cov}(X_i, Y)}{\sqrt{\text{var}(X_i)\text{var}(Y)}}$$

#### ■ Puissance prédictive de l'attribut

#### ■ Marge liée à chaque attribut

### ■ Critères liés à la **théorie de l'information**

- E.g. évaluation empirique de l'estimation mutuelle entre variables

$$\mathcal{I}_i = \int_{x_i} \int_y p(x_i, y) \log \frac{p(x_i, y)}{p(x_i) \cdot p(y)} dx dy$$

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

• Conclusion



## Critères d'arrêt

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

• Conclusion

- **Evaluation en-dessous d'un certain seuil**

- **Méthode par « témoin »**

- Inclure des attributs aléatoires
- Ne pas retenir les attributs dont l'évaluation est en-dessous



# La sélection d'attributs en pratique

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

• Conclusion

- **Recours à des méthodes d'évaluation raisonnables**
  - Hypothèse d'indépendance des attributs (*linéarité*)
    - On peut les évaluer indépendamment
  - *Spectre large* de régularités détectables
- **Utilisation de connaissances *a priori***
  - E.g. : Groupement de gènes *a priori* (réseaux de régulation)
- **Méthode de filtrage (« filter »)**
  - E.g. : SAM, ANOVA, **RELIEF**
- **Estimation**
  - On ordonne les attributs en fonction d'un *critère de performance*
    - ➔ Quel seuil (choisi globalement) ?
    - ➔ Quelle confiance ?



# Critères de performance

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

• Conclusion

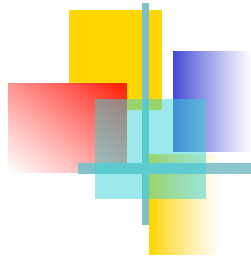
## ■ Hypothèse de distribution paramétrique $\mathcal{N}(\mu, \sigma)$

- Comparaison à hypothèse nulle locale : ANOVA
- Idem (mais différent) : SAM

## ■ Méthodes non paramétriques

- Critère heuristique : RELIEF





# Utilisation d'ANOVA

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

• Conclusion

- **Deux classes (Irradiée / Non Irradiée)**

- $\mathcal{N}(\mu_1, \sigma)$  et  $\mathcal{N}(\mu_2, \sigma)$

- **Comparaison**

- Variance intra-classe
- Variance inter-classes

- **Hypothèse nulle  $\mathcal{H}_0 : \mu_1 = \mu_2$**

- **Rejet si**

$$\frac{V_{\text{inter}}/k-1}{V_{\text{intra}}/n-k}$$

significativement trop grand par rapport aux quantiles de la loi  $\mathcal{F}(k-1, n-k)$



# SAM (Significance Analysis of Microarrays)

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

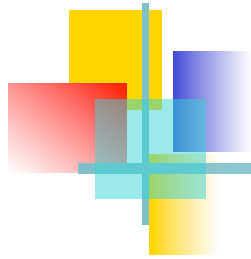
• Conclusion

- Pour chaque gène :

$$d(i) = \frac{x_I(i) - x_{NI}(i)}{s(i) + s_0}$$

déviatiion standard      Constante > 0

- **Gènes potentiellement significatifs** : gènes dont le score  $d(g)$  est supérieur au score moyen du gène obtenu après permutations des classes, de plus d'un certain seuil  $\Delta$
- Calcul du nombre de gènes **faussement significatifs** : nombre moyen de gènes faussement significatifs pour chaque permutation
- **Taux de fausse découverte** (FDR)



## RELIEF (1)

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

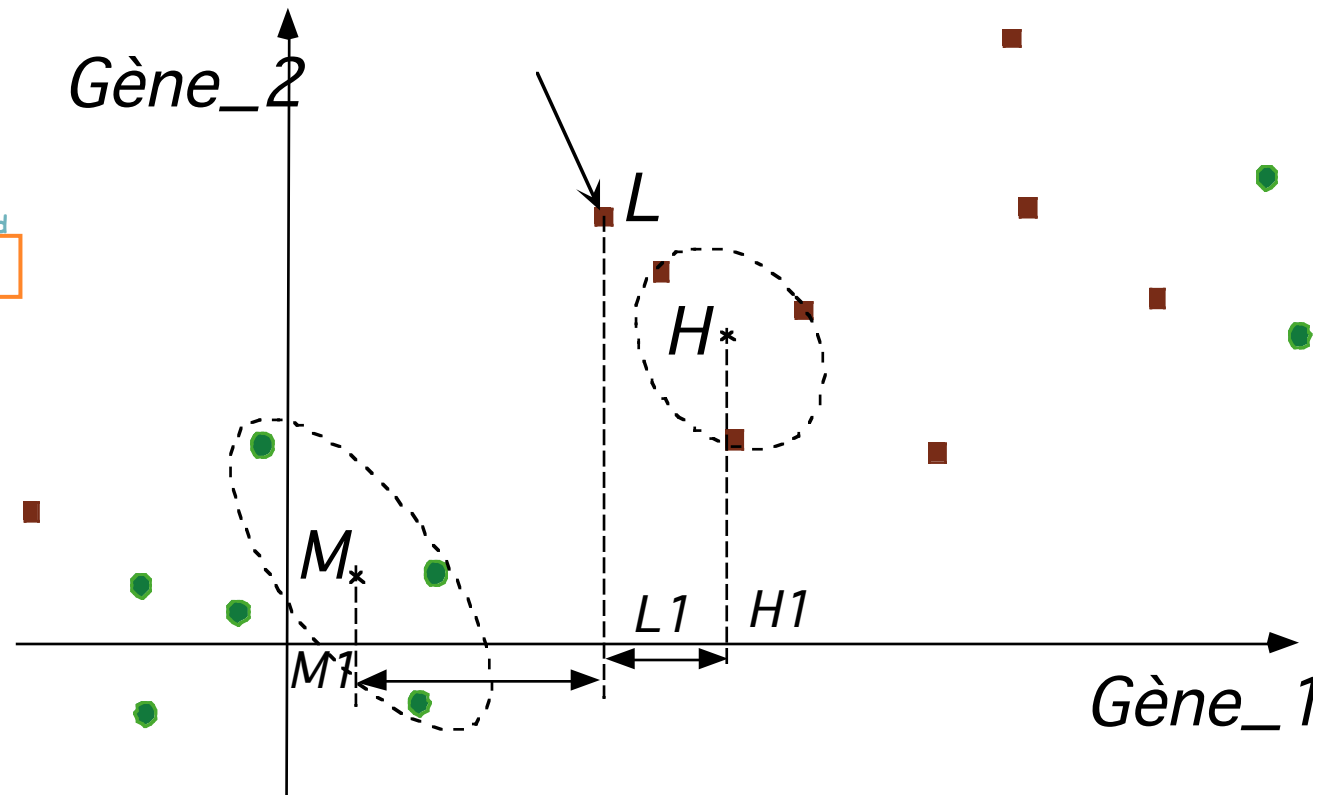
• Conclusion

- [Kira & Rendell,92], [Kononenko,94]
- **Les attributs les plus pertinents sont ceux qui varient plus lorsque l'exemple (lame) considéré change de classe que lorsqu'il ne change pas**
  - Complexité faible
  - Grande résistance au bruit



## RELIEF (2)

- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion





## RELIEF (3)

- Une lame  $L$  est vue comme un point dans un espace à  $p = 6157$  dimensions

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

• Conclusion

- On cherche ses  $k$  plus proches voisins dans la même classe et on note  $H$  (nearest Hit) leur *barycentre*.
- On calcule ses  $k$  plus proches voisins dans l'autre classe et on note  $M$  (nearest Miss) leur *barycentre*.

$$\text{poids}_{\text{gène}} = \frac{1}{m} \sum_{L=1}^m \left\{ \left[ \text{expr}_{\text{gène}}(L) - \text{expr}_{\text{gène}}(M) \right] - \left[ \text{expr}_{\text{gène}}(L) - \text{expr}_{\text{gène}}(H) \right] \right\}$$

où  $\text{expr}_{\text{gène}}(x)$  est la projection selon *gène* du point  $x$ , et  $m$  est le nombre total de lames.

- Le poids calculé pour chaque gène *gène* est ainsi une approximation de la différence de deux probabilités comme suit :

$$\begin{aligned} \text{Poids}(\text{gène}) = & \text{P}(\text{gène a une valeur différente} / k \text{ plus proches voisins dans une classe différente}) \\ & - \text{P}(\text{gène a une valeur différente} / k \text{ plus proches voisins dans la même classe}) \end{aligned}$$

- **Algorithme polynomial** :  $\Theta(pm^2)$
- **Rôle de  $k$**  : prise en compte du bruit



## Sélection des attributs

---

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

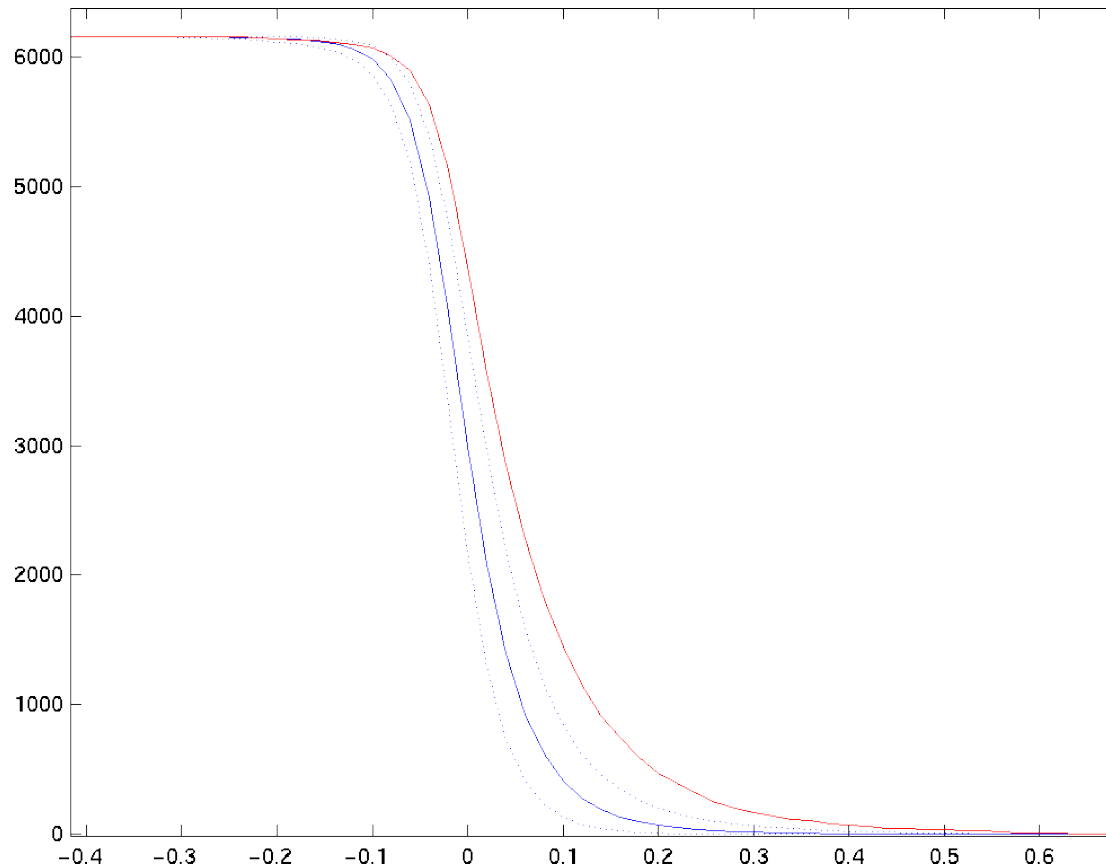
• Conclusion

- **Y a-t-il vraiment de l'information dans les données ?**
- **Quels gènes retenir ?**
- **Avec quelle confiance ?**



# Hypothèse nulle globale

- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion



Nombre de gènes dont le poids dépasse la valeur repérée en abscisse

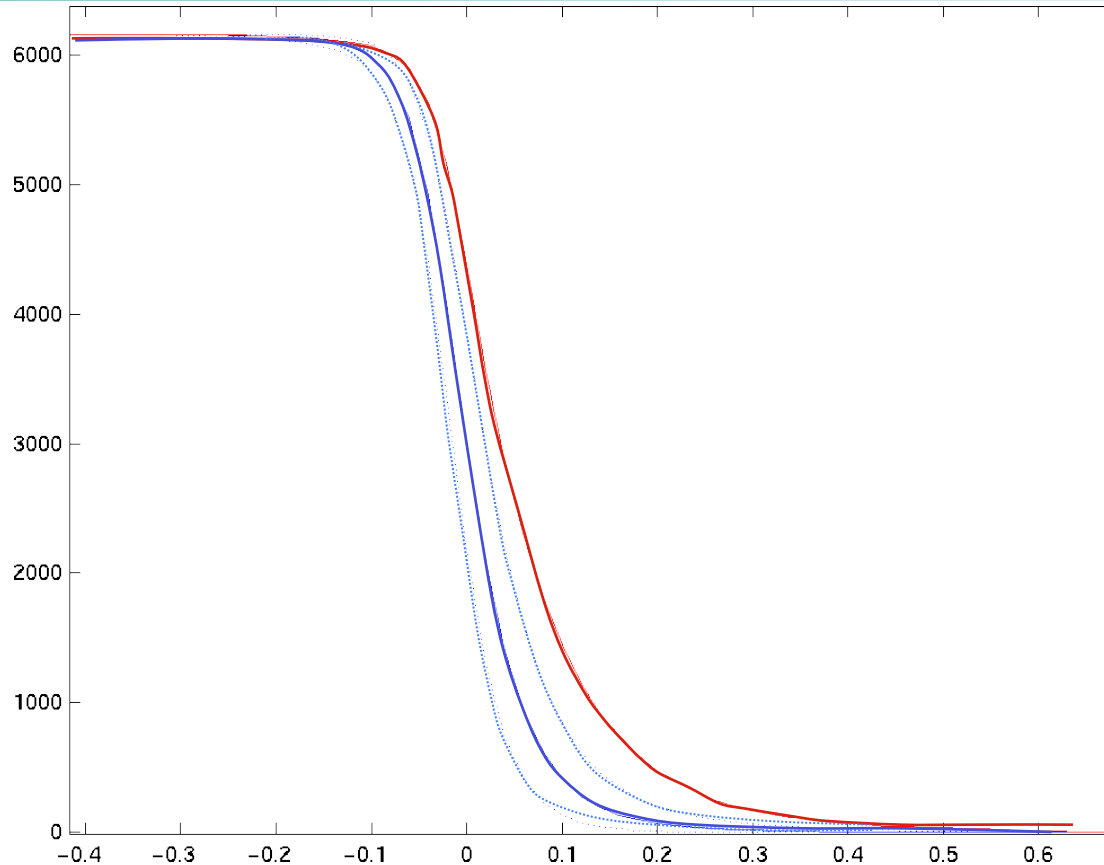
rouge : Avec les classes réelles ;

bleu : Courbe moyenne obtenue avec des classes aléatoires



## Hypothèse nulle globale

- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion



Nombre de gènes dont le poids dépasse la valeur repérée en abscisse

rouge : Avec les classes réelles ;

bleu : Courbe moyenne obtenue avec des classes aléatoires





# Précision ou rappel : choix d'un seuil

## Il faut choisir entre :

- Une liste contenant **presque tous les gènes impliqués mais comportant des faux-positifs**
- Une liste de **gènes impliqués de manière quasi-certaine** dans la réponse à l'Irradiation (quitte à ne pas avoir tous les gènes impliqués)

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

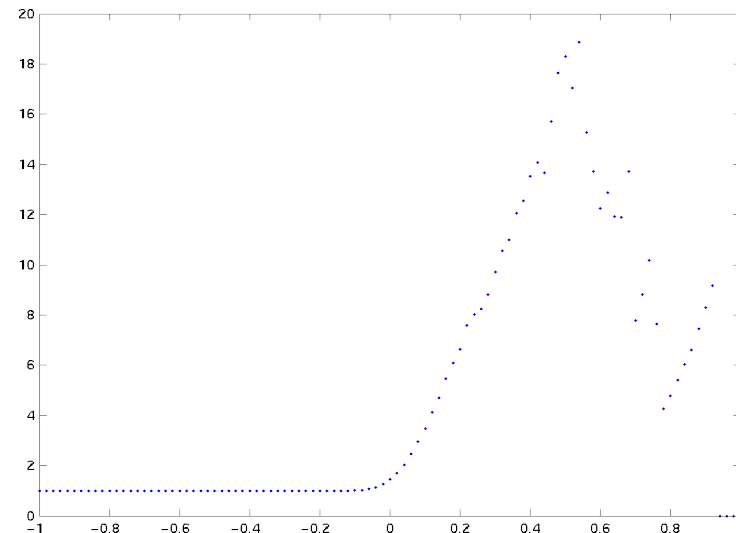
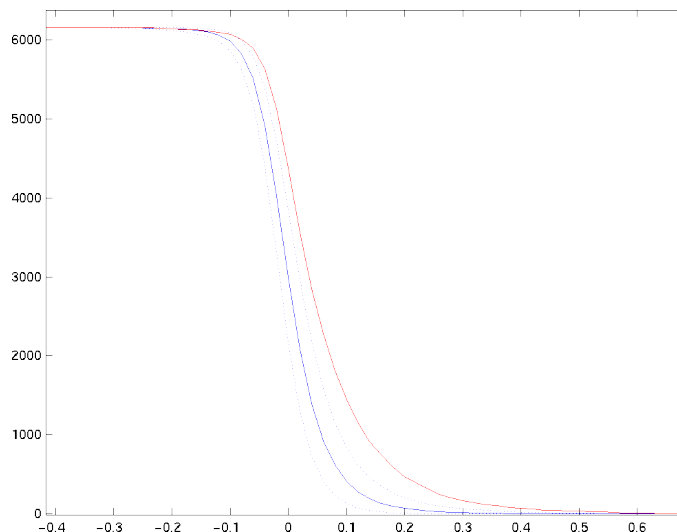
• Comparaison

• Combinaison

• Conclusion



## Problème du seuil





## Combinaison de méthodes ?

- Illustration

- Le pb de la sélection d'attributs

- Méthode standard

- Combiner des méthodes

- Comparaison

- Combinaison

- Conclusion

### ■ *Peut-on faire mieux avec deux méthodes ?*

- Est-ce mieux de prendre l'intersection de leurs sélections ?

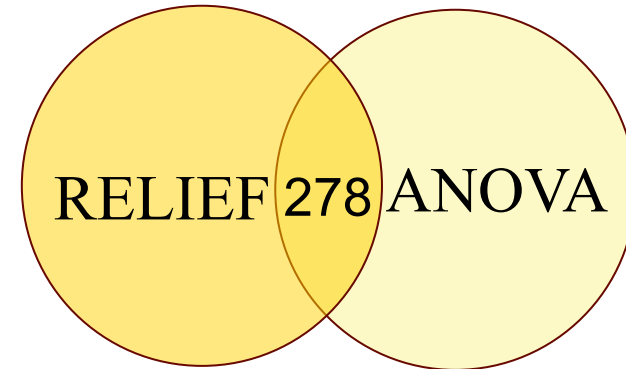
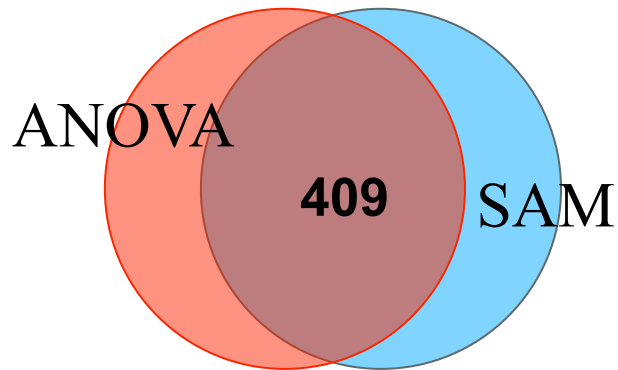
- Doit-on avoir plus de confiance dans la valeur du résultat ainsi obtenu ?



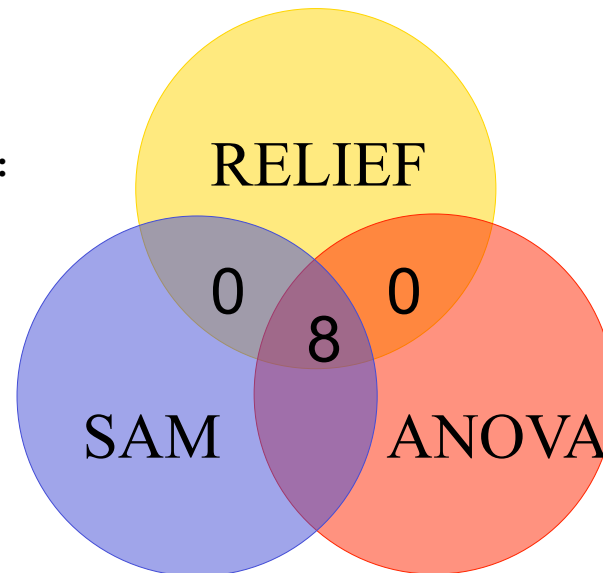
# Intersections (1)

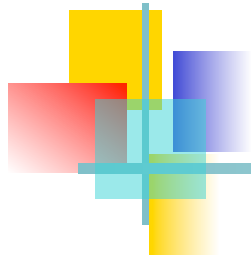
Pour les **500** meilleurs gènes de chaque technique (poids 0.2) :

- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- **Comparaison**
- Combinaison
- Conclusion



Pour les **35** meilleurs (poids 0.5) :





## Intersections (2)

*Est-ce que ces intersections sont significatives ?*

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

• Conclusion

### ■ Problème :

Étant données 2 méthodes sélectionnant au hasard chacune  $n$  gènes parmi  $N$  gènes, quelle est la probabilité que ces deux paquets de  $n$  gènes aient une intersection de cardinal supérieur ou égal à  $k$  ?

$\Rightarrow$  *loi hypergéométrique*  $H(n, N-n, k)$

avec  $N = 6157$  :

- $n = 500$  :  $P(\text{taille intersection} \geq 257) = 10^{-169}$
- $n = 35$  :  $P(\text{taille intersection} \geq 8) = 10^{-12}$

➡ Le biologiste est satisfait !



## Répartition des meilleurs gènes

- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion

function of 91 induced genes/171	number of ORFs	% in this list	% total ORFS (6158)	pp
unknown	38	41,8	50,4	0,8
oxidative stress response	4	4,4	0,3	14,3
oxidative phosphorylation	9	9,9	0,3	30,5
transport	4	4,4	2,2	2,0
gluconeogenesis	1	1,1	0,1	16,9
protein processing & synthesis	3	3,3	2,0	1,6
ATP synthesis	7	7,7	0,4	20,6
glucose repression	1	1,1	0,2	4,8
respiration	2	2,2	0,1	22,0
function of 80 repressed genes/171	number of ORFs	% in this list	% total ORFS	sur-rep
unknown	45	56,3	50,4	1,1
stress response (putative)	1	1,3	0,2	7,0
glycerol metabolism	2	2,5	0,1	30,8
protein processing & synthesis	3	3,8	2,0	1,9
secretion	2	2,5	2,0	1,3
transport	4	5,0	2,2	2,3
glycolysis	2	2,5	1,0	2,5



# Interprétation biologique

- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion

## Cytochrome bc1

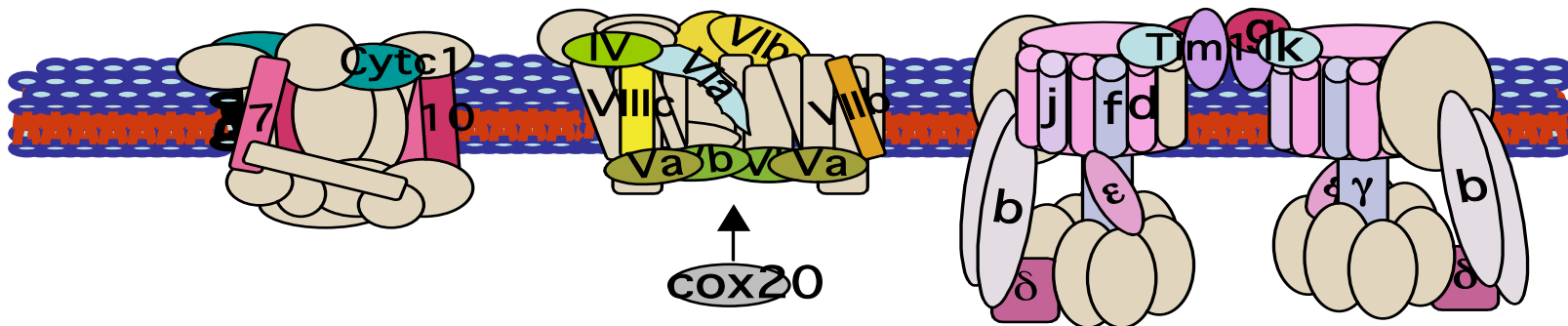
Cyt1  
QCR7  
QCR10

## Cytochrome oxidase

COX5A  
COX6  
COX4  
COX 13  
COX12  
COX7  
COX8  
COX20

## ATP synthase

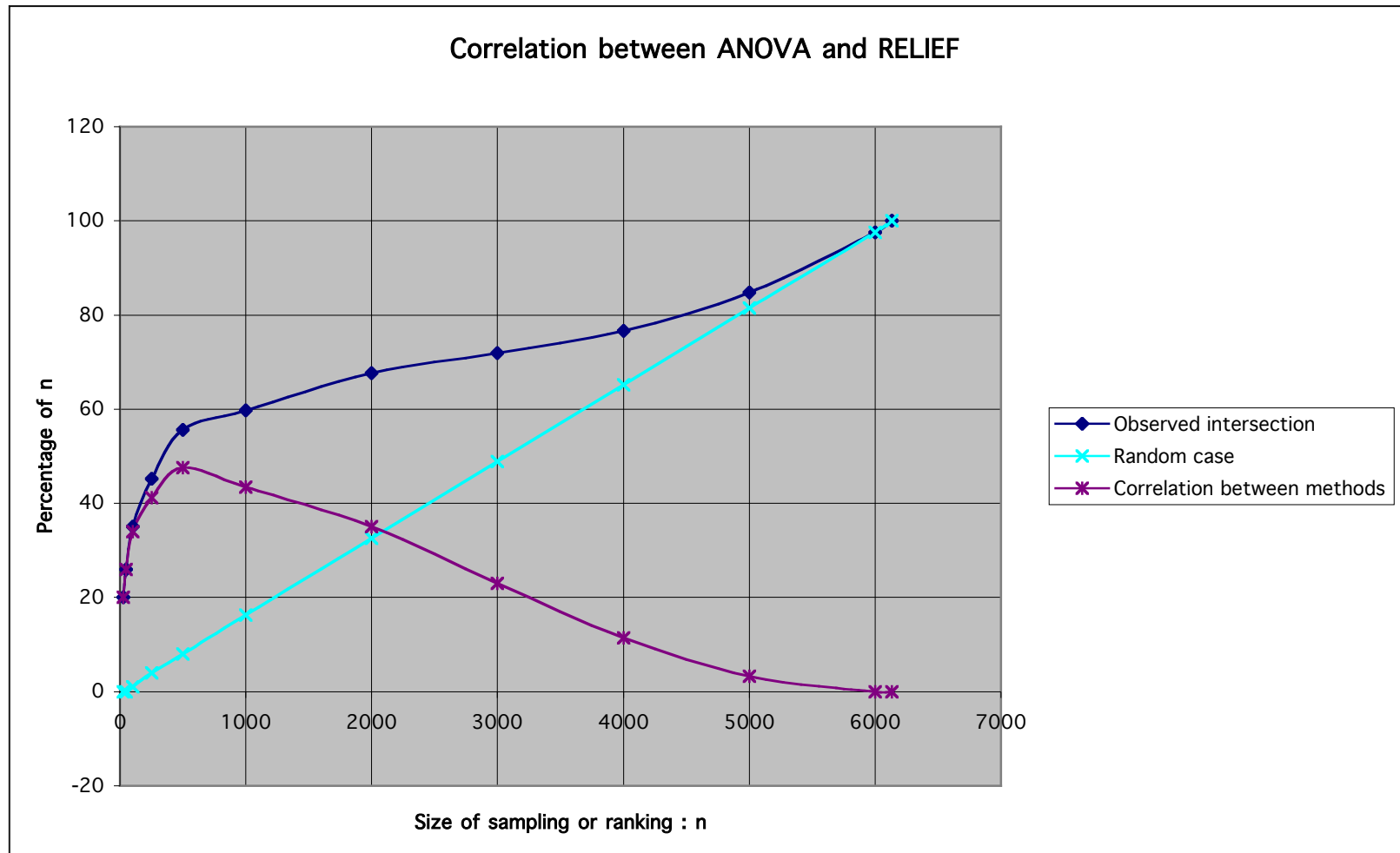
ATP3  
ATP5  
ATP16  
ATP15  
ATP7  
ATP17  
ATP18  
ATP19  
ATP20  
TIM11





# Comparaison de méthodes de sélection

- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- **Combinaison**
- Conclusion





## Comparaison de méthodes de sélection (2)

- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- **Combinaison**
- Conclusion

### ■ Causes possibles de l'intersection :

- **Information dans les données** que les deux méthodes parviennent à détecter
- **Corrélation *a priori*** des méthodes

### ■ Exemple

- **278 gènes dans  $(RELIEF \cap ANOVA)_{500}$**
- 40 attendus par simple chance (loi hypergéométrique)
- 238 ?
  - Information ?
  - Corrélation *a priori* ?





## Mesure de la corrélation a priori

### ■ Nouvelle hypothèse nulle

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

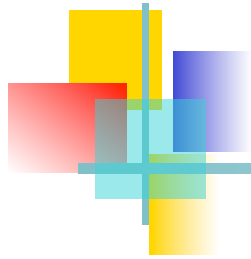
• Comparaison

• Combinaison

• Conclusion

- Pour toutes les permutations de 6 + & 12 - sur les données
- Calculer :  $(\text{RELIEF} \cap \text{ANOVA})_{500}$
- Faire la moyenne

⇒ Intersection due à la corrélation a priori des méthodes



## Comment l'interpréter ?

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• **Combinaison**

• Conclusion

■ Si  $(\text{RELIEF} \cap \text{ANOVA})_{500} = \dots$

■ 0 : ?

■ 40 : ?

■ 278 : ?

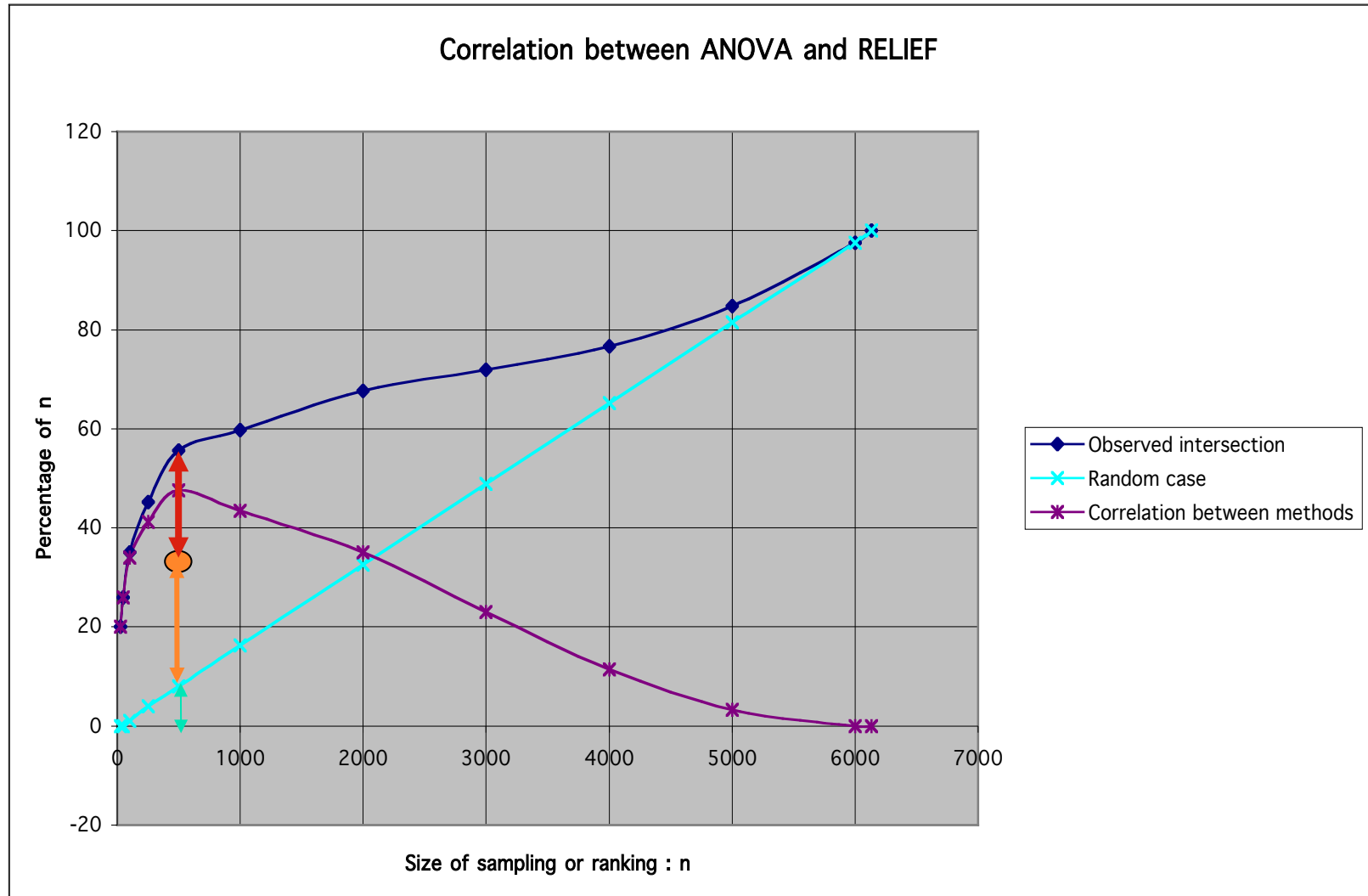
■ 500 : ?

Ici : **170**



# Mesure de corrélation

- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- **Combinaison**
- Conclusion





## Combinaison de méthodes

### *Peut-on tirer de l'information de la combinaison de deux méthodes ?*

- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion

- On dispose de la loi empirique de

$$k = (\text{RELIEF} \cap \text{ANOVA})_n$$

en fonction de  $n$  (intersection des « top\_n »)

➔ Peut-on la comparer à une **courbe théorique** paramétrée et trouver les paramètres maximisant la vraisemblance ?



# Combinaison de méthodes

## 1. Construction d'un « modèle génératif »

- On suppose deux méthodes  $M_1$  et  $M_2$  d'évaluation d'attributs telles que:
  - $(M_1 \cap M_2)_n = k$
  - $M_1$  retourne  $p_1$  attributs pertinents dans  $n$
  - $M_2$  retourne  $p_2$  attributs pertinents dans  $n$
- On suppose  $p$  vrais attributs pertinents sur  $d$  attributs en tout
- On calcule la loi :

$$k = \text{fct}(d, n, k_{\text{corr}}, p, p_1, p_2)$$

## 2. Principe de maximum de vraisemblance

- On retient  $(p, p_1, p_2)$  maximisant la vraisemblance par rapport à la courbe observée

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

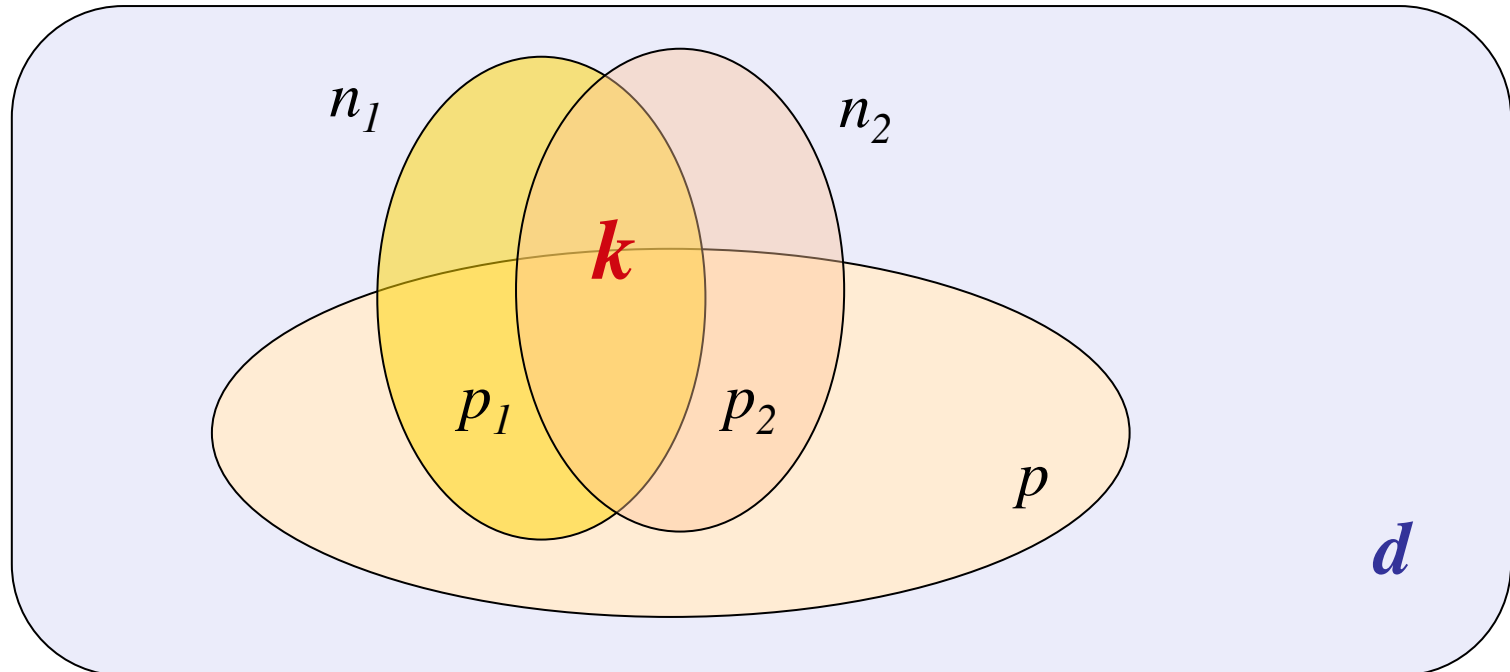
• Combinaison

• Conclusion



## Combinaison de méthodes

- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion

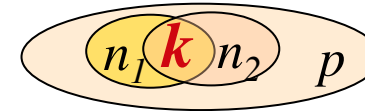


***Combien parmi les  $k$  sont positifs ?***



# Formules

- $n \leq p_1 \leq p_2 \leq p$



- Illustration

- Le pb de la sélection d'attributs

- Méthode standard

- Combiner des méthodes

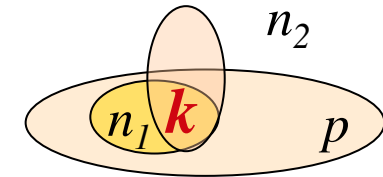
- Comparaison

- Combinaison

- Conclusion

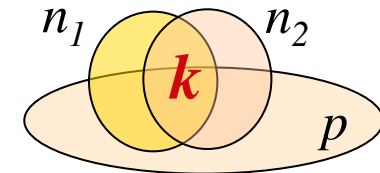
$$p(\cap = k | d, p, n_1 = n_2 = n, k_C) = \binom{n}{k} \binom{p-n}{n-k} / \sum_{i=k_C}^n \binom{n}{i} \binom{p-n}{n-i}$$

- $p_1 \leq n \leq p_2 \leq p$



$$p(\cap = k | d, p, n_1 = n_2 = n, k_C) = \frac{\binom{d-n}{n-k} \binom{n}{k} \binom{p}{n}}{\binom{d}{n}^2} / \sum_{j=k_C}^n \frac{\binom{d-n}{n-j} \binom{n}{j} \binom{p}{n}}{\binom{d}{n}^2}$$

- $p_1 \leq p_2 \leq n$



$$p(\cap = k | d, p, n_1 = n_2 = n, k_C) = \frac{\binom{p}{p_1} \binom{d-p}{n-p_1} \sum_{k^+=\max(0, p_1+p_2-p)}^{p_1} \binom{p_1}{k^+} \binom{p-p_1}{p_2-k^+} \binom{n-p_1}{k-k^+} \binom{d-n-(p-p_1)}{n-p_2-(k-k^+)}}{\binom{d}{n}^2} / \sum_{i=k_C}^n \dots$$



## Et alors ...

- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion

### ■ Pour

- $d : 6135$
- $n : 500$
- $k_{corr} : 170$

### ■ Le maximum de vraisemblance est obtenu pour :

- $p \approx 410$
- $p_1 = p_2 \approx 330$





## Conclusion

- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion

*On peut tirer de l'information de l'utilisation de plusieurs méthodes*

- Pas de travaux connus dans ce domaine

- ***Propositions***

- Méthode de **mesure de corrélation *a priori*** des méthodes
- Méthode de ***maximum de vraisemblance*** pour suggérer le nombre d'attributs pertinents à partir de deux méthodes



## Références

• Illustration

• Le pb de la  
sélection  
d'attributs

• Méthode standard

• Combiner des  
méthodes

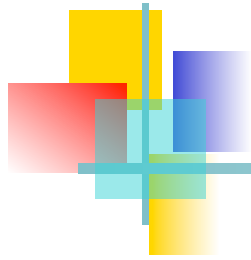
• Comparaison

• Combinaison

• Conclusion



- Bell, D., & Wang, H. (2000). A formalism for relevance and its application in feature subset selection. *Machine Learning Journal*, 41, 175-195.
- Blum, A., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence journal*(97), 245-271.
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157-1182.
- Liu, H., & Motoda, H. (1998). *Feature selection for knowledge discovery & data mining*. Kluwer Academic Publisher.
- Mercier, G., Berthault, N., Mary, J., Antoniadis, A., Comet, J.-P., Cornuéjols, A., Froidevaux, C., & Dutreix, M. (2004). Biological detection of low radiation by combining results of two analysis methods. *Nucleic Acids Research (NAR)*, 32(1), 1-8.



## Résultats

- Illustration
  - Les données reflètent-elles la présence de l' irradiation ? **oui**
- Le pb de la sélection d' attributs
- Méthode standard
  - Combien de gènes sont-ils impliqués ? **Plus de 100**
- Combiner des méthodes
- Comparaison
- Combinaison
  - Y a-t-il des groupes de gènes impliqués et lesquels ?
- Conclusion
  - Oui : ATP synthesis, oxidative phosphorylation et oxidative stress response**
  - Est-il possible de déterminer si une levure est irradiée en regardant son transcriptome ?  
**Oui et il suffit de ne regarder qu' un petit nombre de gènes**



# Tâche de classification

## ➤ Plusieurs techniques ont été utilisées

- Illustration ➤ Vote « d'experts »
- Le pb de la sélection d'attributs ➤ Technique du maximum de vraisemblance
- Méthode standard ➤  $K$  plus proches voisins
- Combiner des méthodes

## • Comparaison ➤ Essai de **classification en aveugle** sur six nouvelles lames :

• Combinaison

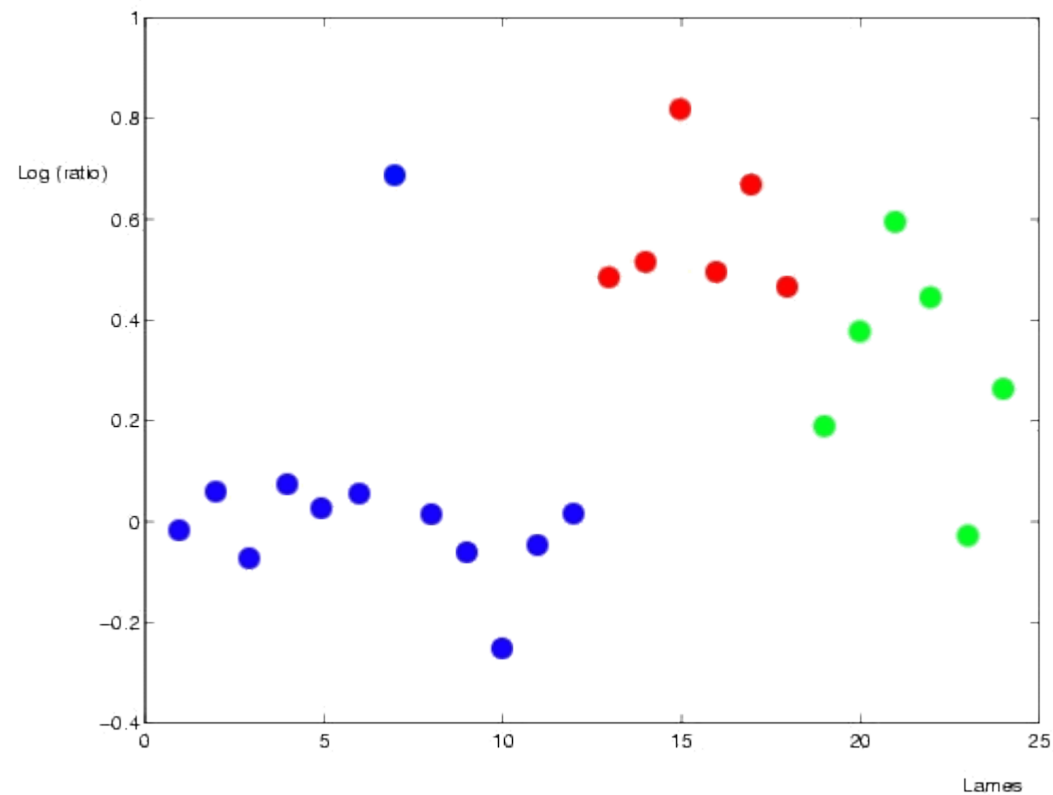
• Conclusion

Traitement	Dose	Avec sélection d'un seul gène (1575)		Avec les gènes sélectionnés par ANOVA		Avec les gènes sélectionnés par REL	
		Sain	Irradié	Sain	Irradié	Sain	Irradié
Irradiation	0.003 mGy/h	0,95	0,04	0,53	0,47	1	0
Irradiation	0.007 mGy/h	0,35	0,65	0,46	0,54	0,01	0,9
Irradiation	0.1 mGy/h	0,02	0,97	0,5	0,5	0	1
Irradiation	1.1 mGy/h	0,15	0,84	0,47	0,53	0	1
Formaldehyde	0.07 mM	1	0	0,65	0,35	1	0
aucun	0	0,82	0,17	0,55	0,44	1	0



# Le gène 1575

- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion





## Travaux en cours

- **Publication** des résultats biologiques obtenus
  - Illustration
  - Le pb de la sélection d'attributs
  - Étude sur d' autres données (Cancer de la vessie avec Curie, Paris)
  - Méthode standard
  - Combiner des méthodes
  - Comparaison
  - Mise au point d' une méthode de **classification** avec peu de gènes
  - Combinaison
  - Conclusion
- Étude du critère de **RELIEF**
  - Quelles propriétés ?
- Exploitation de multiples méthodes de sélection d' attributs



# Normalisation des données

➤ La normalisation a été réalisée par LOWESS ( LOcally WEighted Scatterplot Smoothing ), Julie PEYRE & Anestis ANTONIADIS (IMAG)

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

• Combinaison

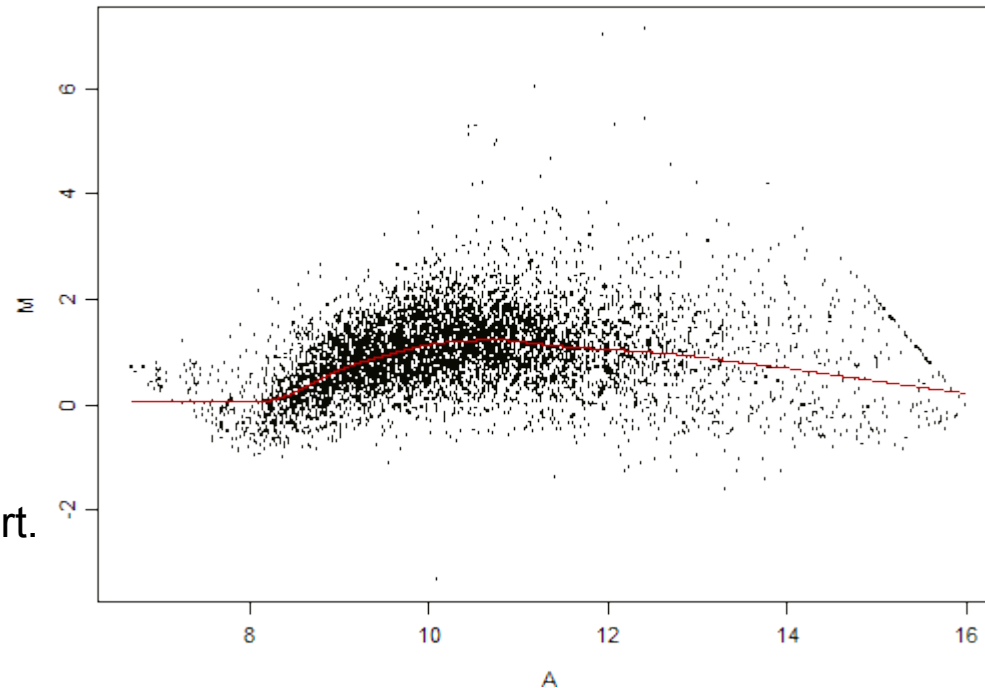
• Conclusion

$$A = \frac{1}{2} \log_2(R * G)$$

$$M = \log_2\left(\frac{R}{G}\right)$$

Où R et G sont les niveaux d'intensité de Rouge et de Vert.

Normalisation par lowess.





## Les sources de problèmes

### ➤ Présence de bruit dans les données à deux niveaux :

#### • Illustration

- Le pb de la sélection d'attributs → Imprécision de la mesure : **bruit classique** supposé gaussien, bruit qui est très élevé pour certains gènes (cf doubles mesures)

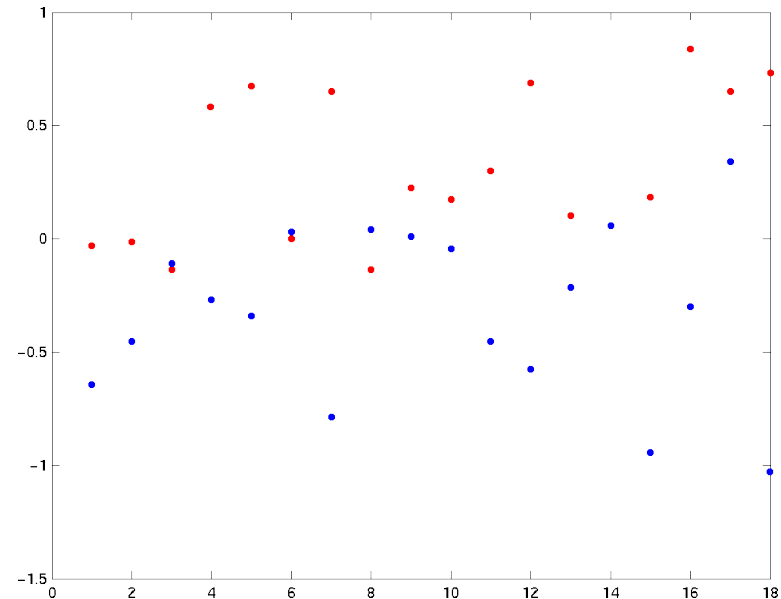
#### • Méthode standard

- Combiner des méthodes → Présence de **valeurs aberrantes** dues à un problème lors de l'hybridation

#### • Comparaison

#### • Combinaison

#### • Conclusion



### ➤ Données déjà **normalisées**

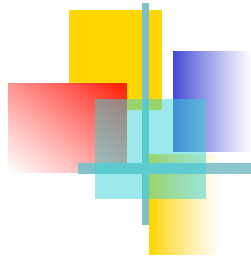
### ➤ Nombreux attributs : **6157 gènes**

### ➤ Très faible nombre d'instances : **12 cultures non-traitées, 6 irradiées**

### ➤ Classes **déséquilibrées** (elles ne contiennent pas le même nombre d'éléments)

### ➤ **Absence d'indépendance conditionnelle** probabiliste entre les gènes





# Sélection

• Illustration

• Le pb de la sélection d'attributs

• Méthode standard

• Combiner des méthodes

• Comparaison

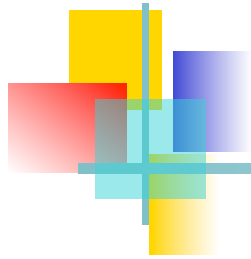
• Combinaison

• Conclusion

- **Trop peu de garantie sur chaque corrélation détectée (attribut)**

➔ **Comparaison à *hypothèse nulle globale***

➔ **Interprétation / confirmation par *les biologistes***



## Utilisation d'ANOVA (suite)

- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion

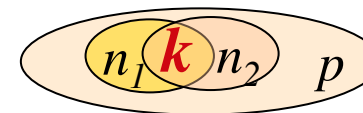
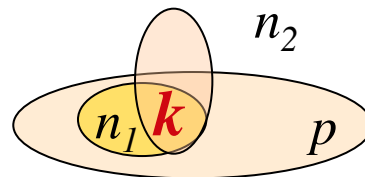
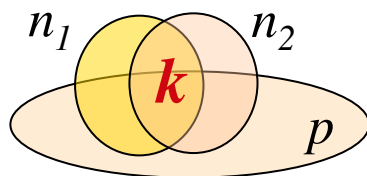
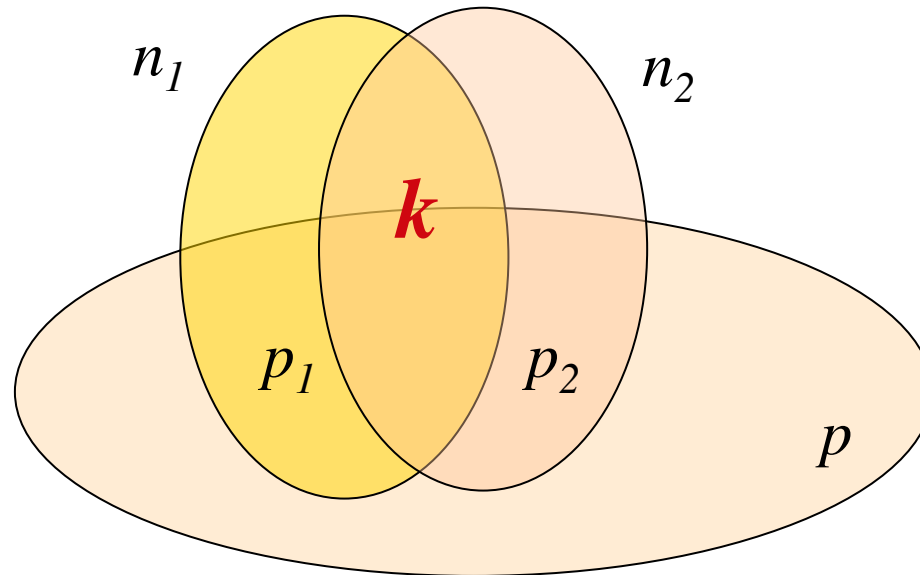
On peut aussi calculer la *p-value* pour chaque gène et ordonner les gènes

**Probabilité que le test rejette l'hypothèse  $\mathcal{H}_0$  à tort**

$$p(t) = \min\{F_0(t), 1 - F_0(t)\}$$



- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion





- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion

