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

Antoine Cornuéjols[1]

J-P. Comet[3], M. Dutreix[2], Ch. Froidevaux[1]

J. Mary[1], G. Mercier[2]

[1] *LRI (Orsay)* - [2] *Institut Curie (Orsay)* - [3] *LAMI (Evry)*

antoine@lri.fr, http://www.lri.fr/~antoine

Plan

Illustration : un pb d'analyse du transcriptome

- Projet INRS, Bioingéniérie 2001
- [2001-2004]

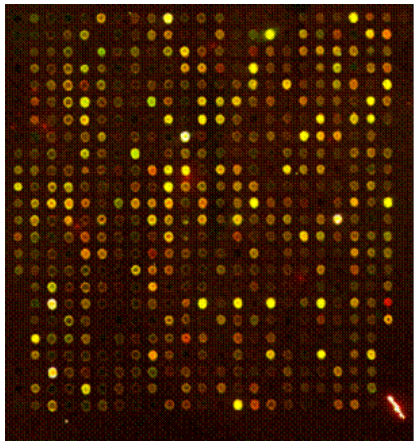
Étude de l'effet des très faibles radiations

Etude des radiations

- **Danger indiscutable dans certains cas. En particulier pour les fortes doses d'irradiation.**
- **Quel impact des faibles doses ?**
- **Biologiquement aucun détecté**
- **Y a-t-il des effets au niveau des gènes ?**

Protocole expérimental

- ***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

- **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

- **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

- **A priori plus simple que celui de la classification** (apprentissage de la relation de dépendance)
- **E.g. Supposons 3 attributs binaires et fonction booléennes**

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)

- **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

La sélection d'attributs en pratique

- **Recours à des méthodes raisonnables**
  - Hypothèse d'indépendance des attributs (linéarité)
    - On peut les évaluer indépendamment
  - Spectre large de régularités détectables

- ***Sélection***
  - Chaque attribut passe un test

- ***Estimation***
  - On ordonne les attributs en fonction d'un ***critère de performance***
    - ➥ Quel seuil (choisi globalement) ?
    - ➥ Quelle confiance ?

La sélection d'attributs en pratique

- **Sélection d'attributs**
  - Approche directe
  - Approche « wrapper »
  - Approche par **filtrage**
- **Réduction de dimensionnalité**
  - Groupement de gènes *a priori* (réseaux de régularisation)

- **Exemples de méthodes d'estimation d'attributs par filtrage**
  - SAM
  - ANOVA
  - **RELIEF**
  - …

Critères de performance

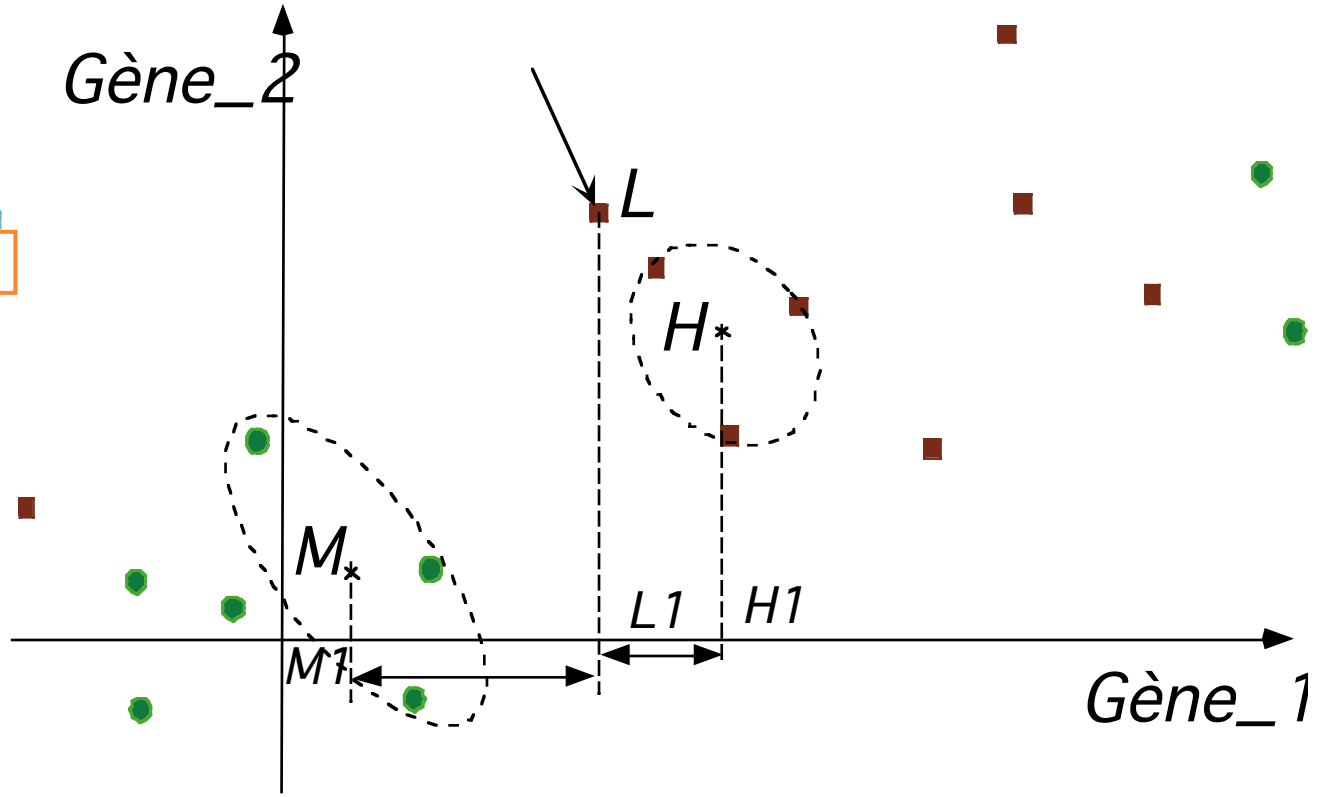
- **Hypothèse de distribution paramétrique $\mathcal{N}(\mu,\sigma)$**
  - Comparaison à hypothèse nulle locale : ANOVA
  - Idem (mais différent) : SAM
- **Méthodes non paramétriques**
  - Critère heuristique : RELIEF

RELIEF (1)

- [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**
  - 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***.

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

où $expr_{gène}(x)$ **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 :**

Poids(*gène*) = P (*gène* a une valeur différente / *k* plus proches voisins dans une classe différente)
- P (*gène* a une valeur différente / *k* plus proches voisins dans la même classe)

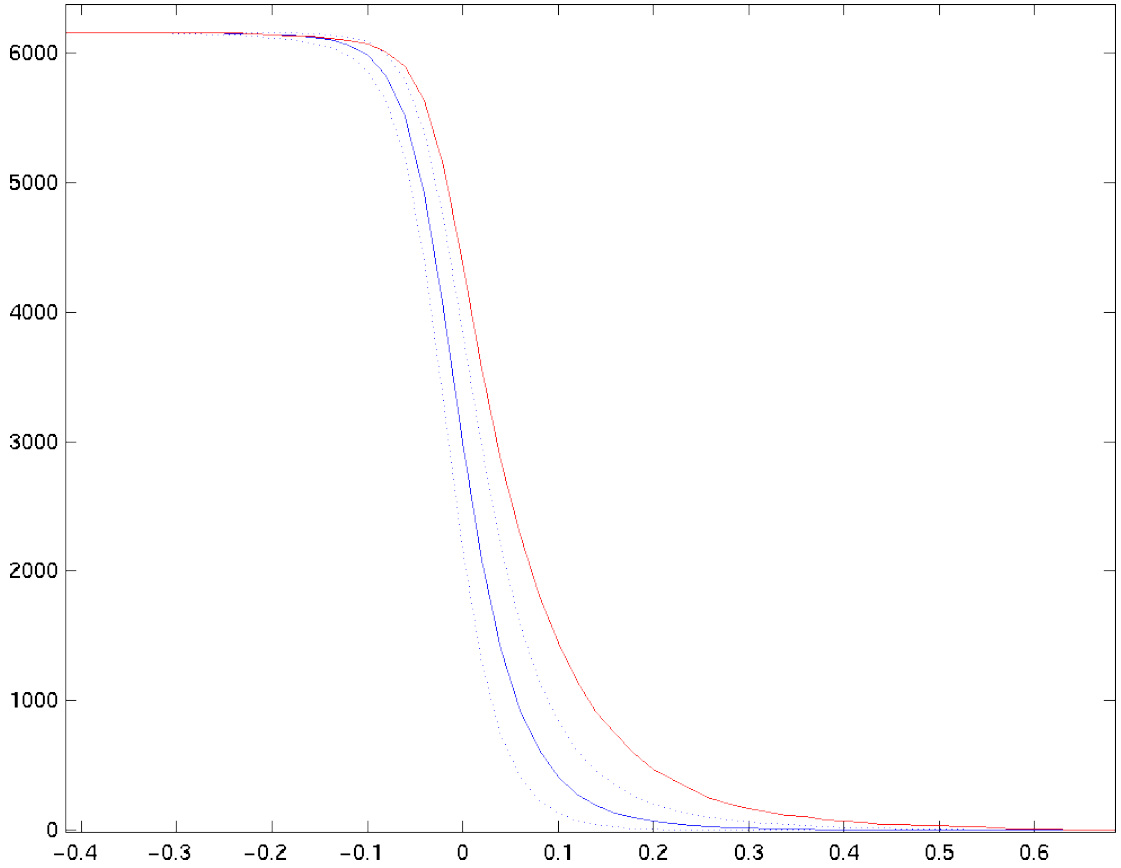
  - **Algorithme polynomial : $\Theta(pm^2)$**
  - **Rôle de *k* : prise en compte du bruit**

Sélection des attributs

- **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



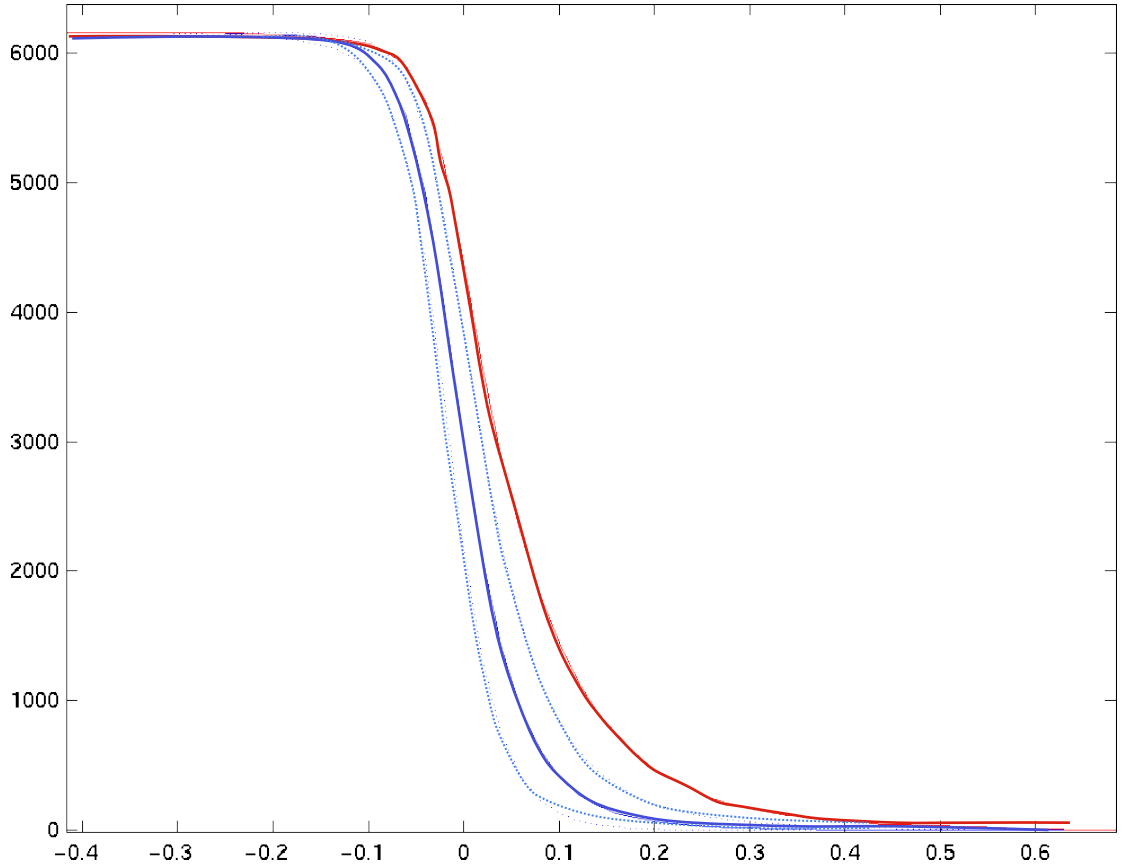
Nombre de gènes dont le poids dépasse la valeur repérée en abscisse

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 ;

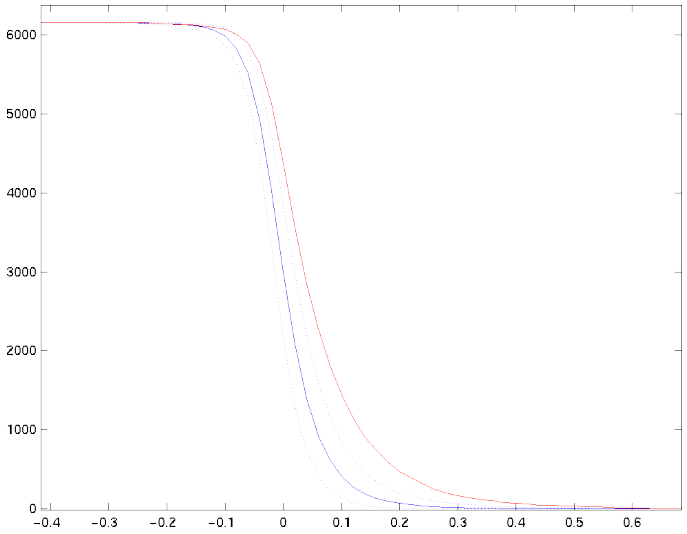
bleu : Courbe moyenne obtenue avec des classes aléatoires

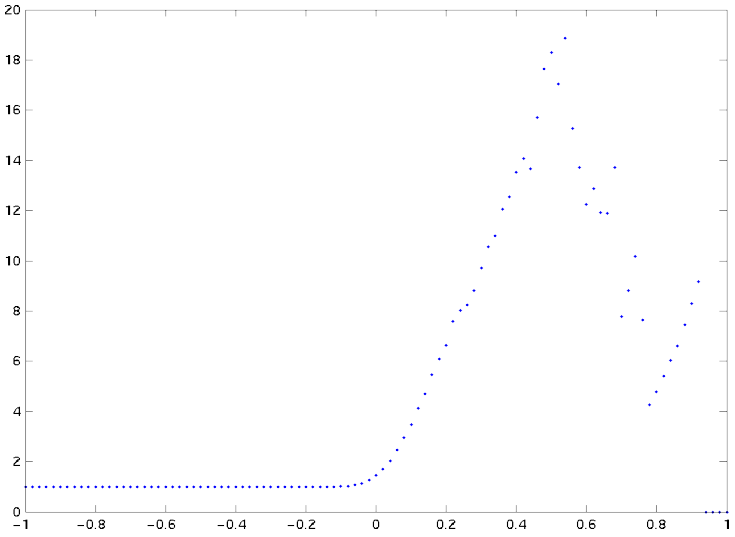
Précision ou rappel : choix d'un seuil

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

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

➠ Problème du seuil





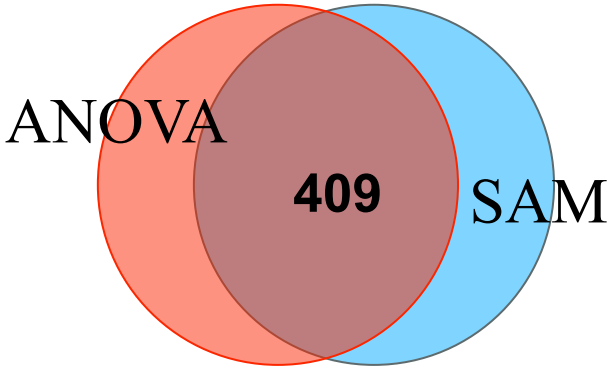
Combinaison de méthodes ?

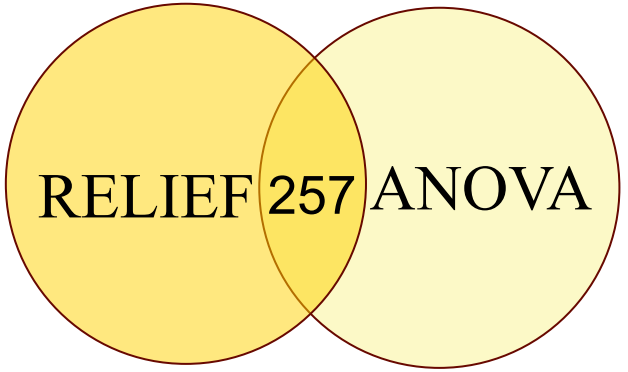
- ***Peut-on faire mieux avec deux méthodes ?***
  - Est-ce mieux de prendre l'intersection de leurs sélections ?
  - Doit-on avoir plus de confiance dans la valeur du résultat ainsi obtenu ?

Intersections (1)

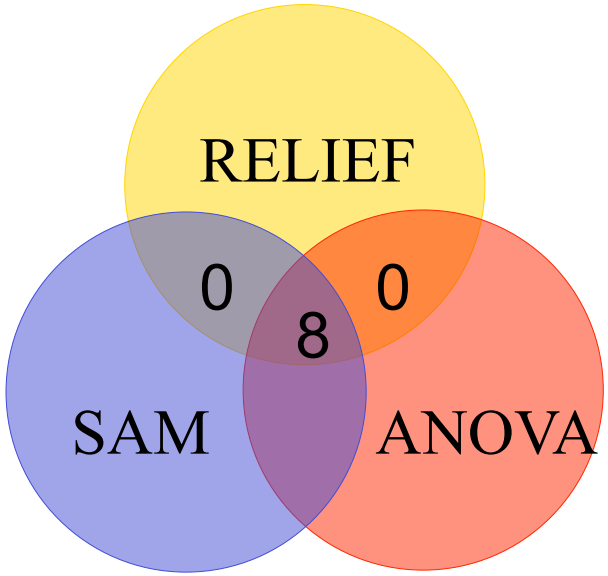
- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion

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





Pour les *35* meilleurs (poids 0.5) :



Intersections (2)

Est-ce que ces intersections sont significatives ?

- Problème :

  Étant donné 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 ?

  = = > ***loi hypergéométrique*** $H(n, N-n, k)$

  avec $N = 6157$:

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

➠ Le biologiste est satisfait !

Répartition des meilleurs gènes

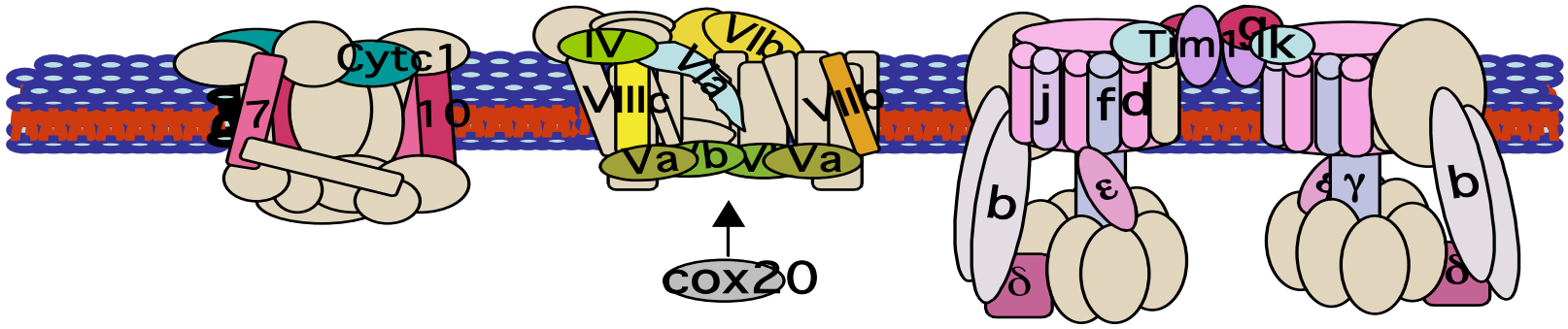
- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion

function of 91 induced genes/171	number of ORFs	% in this list	% total ORFS (6158)	sur-rep
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

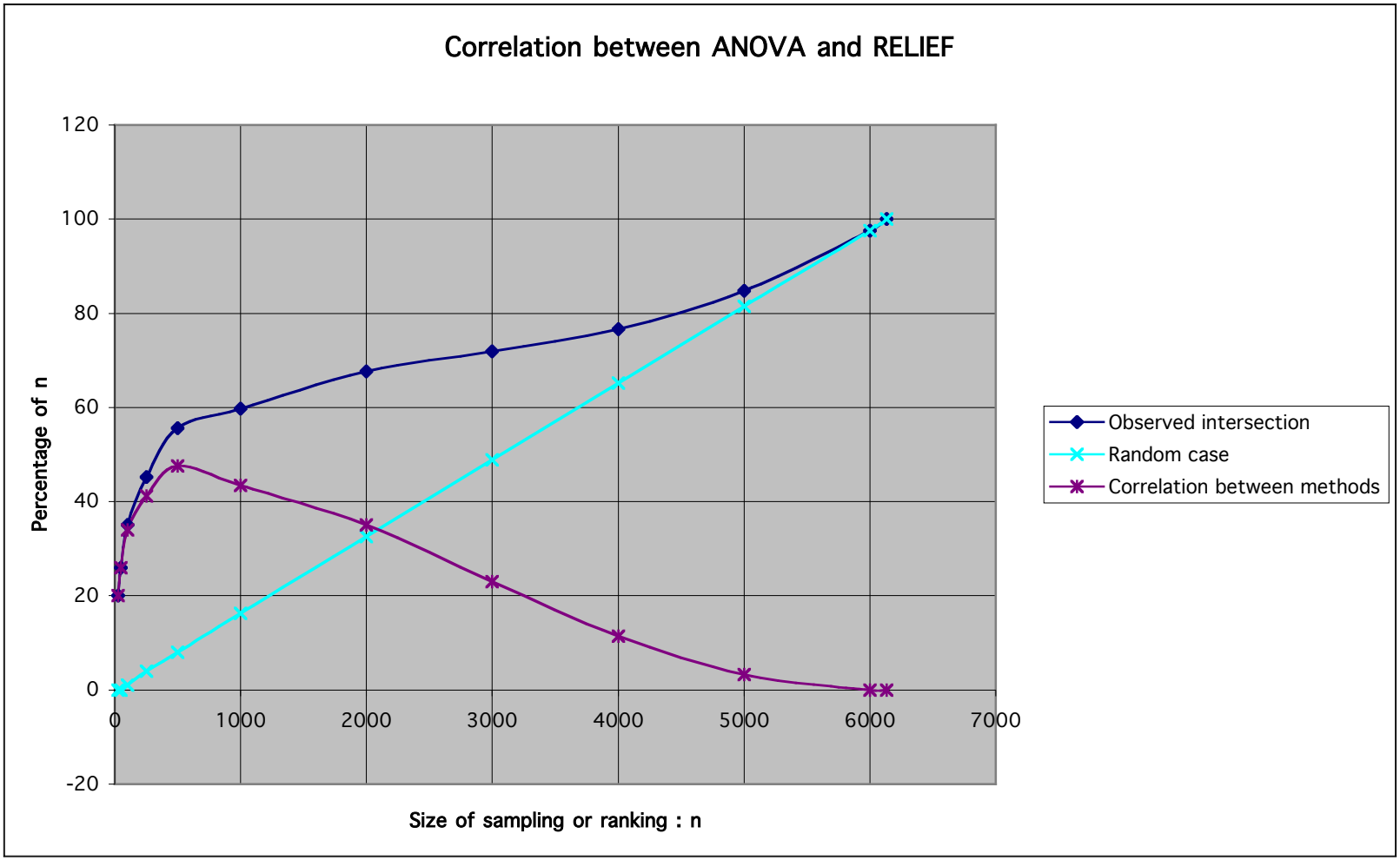
- Illustration
- Le pb de la sélection d'attributs
- Méthode standard
- Combiner des méthodes
- Comparaison
- Combinaison
- Conclusion

Cytochrome bc1	Cytochrome oxidase	ATP synthase
Cyt1	COX5A	ATP3
QCR7	COX6	ATP5
QCR10	COX4	ATP16
	COX 13	ATP15
	COX12	ATP7
	COX7	ATP17
	COX8	ATP18
	COX20	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)

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

Mesure de la corrélation a priori

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

- **Nouvelle hypothèse nulle**
  - Pour toutes les permutations de 6 + & 12 - sur les données
  - Calculer : **(RELIEF ∩ ANOVA)$_{500}$**
  - Faire la moyenne

➠ Intersection due à la corrélation a priori des méthodes

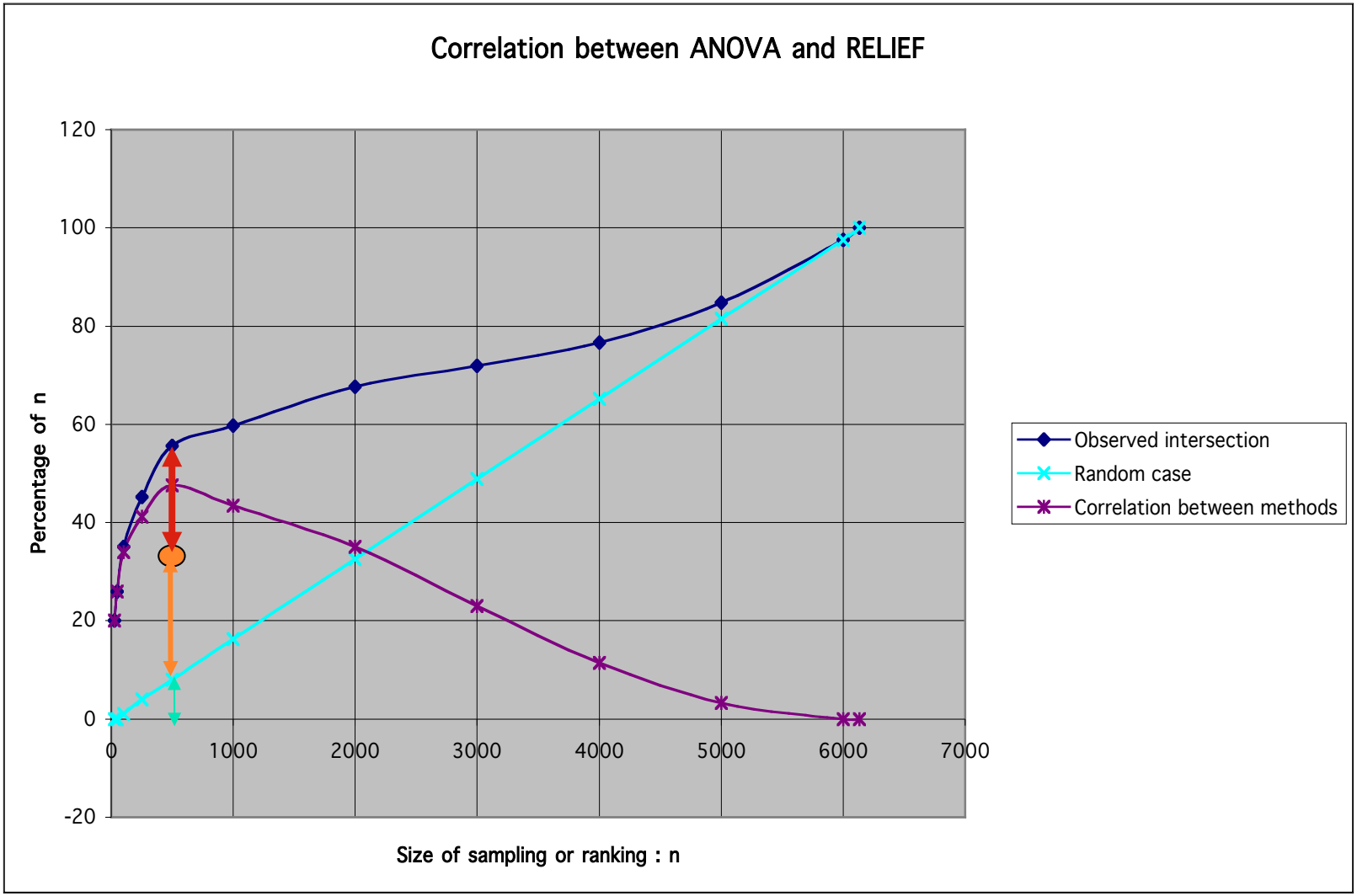
Comment l'interpréter ?

- **Si** $(\text{RELIEF} \cap \text{ANOVA})_{500} = \ldots\ldots$
  - 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 ?

- 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

- On suppose deux méthodes M_1 et M_2 d'évaluation d'attributs telles que :
  - On considère $(M_1 \cap M_2)_n = k$
  - 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 :**

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

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

Formules

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

- $n \leq n_1 \leq n_2 \leq p$

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

- $n_1 \leq n \leq n_2 \leq p$

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

- $n_1 \leq n_2 \leq n$

$$p(\cap = k|d,p,n_1 = n_2 = n,k_C) = \sum_{l=n_1}^{\min(p,n)} \frac{\binom{p}{l}\binom{d-p}{n-l}}{\binom{d}{n}} \frac{\sum_{i=0}^{l}\binom{l}{i}\binom{n-l}{k-i}\sum_{l_2=n_2}^{\min(p,n)}\binom{p-l}{l_2-i}\binom{d-n-(p-l)}{n-l_2-k+i}}{\binom{d}{n}} / (\ldots)$$

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

Résultats

- Les données reflètent-elles la présence de l'irradiation ? **oui**
- Combien de gènes sont-ils impliqués ? **Plus de 100**
- Y a-t-il des groupes de gènes impliqués et lesquels ?

  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

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

Plusieurs techniques ont été utilisées

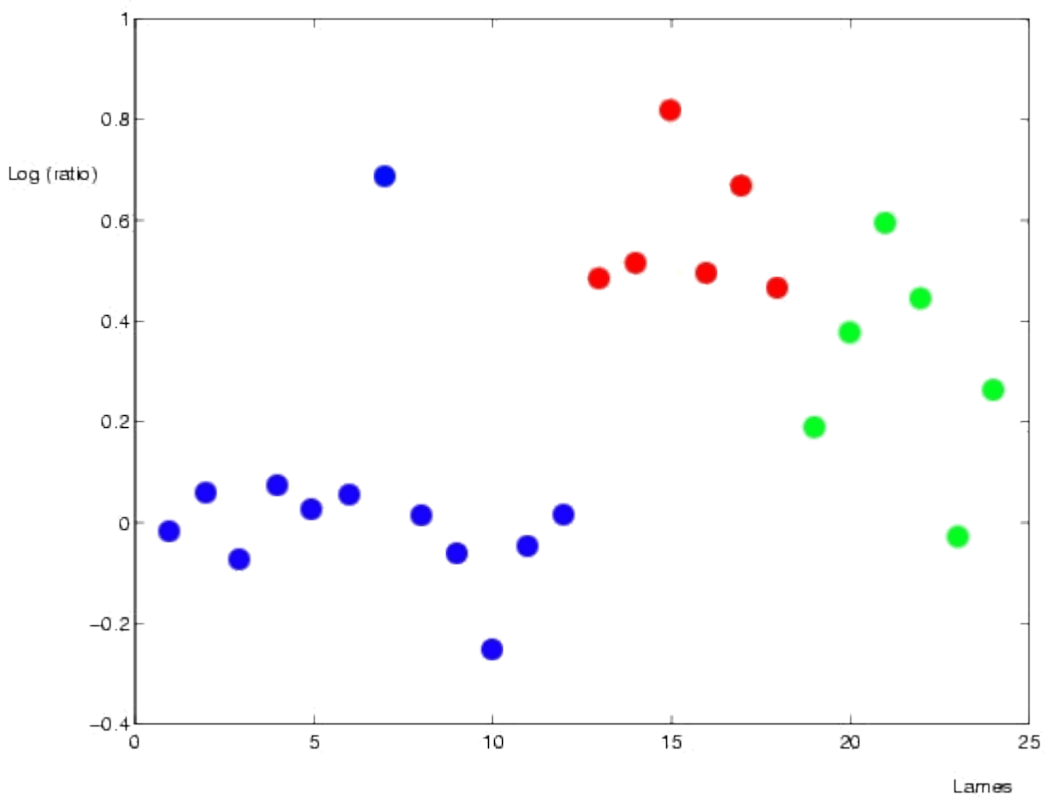
- Vote « d'experts »
- Technique du maximum de vraisemblance
- *K* plus proches voisins

Essai de classification en aveugle sur six nouvelles lames :

		Avec sélection d'un seul gène (1575)		Avec les gènes sélectionnés par ANOVA		Avec les gènes sélectionnés par REL	
Traitement	Dose	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

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

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

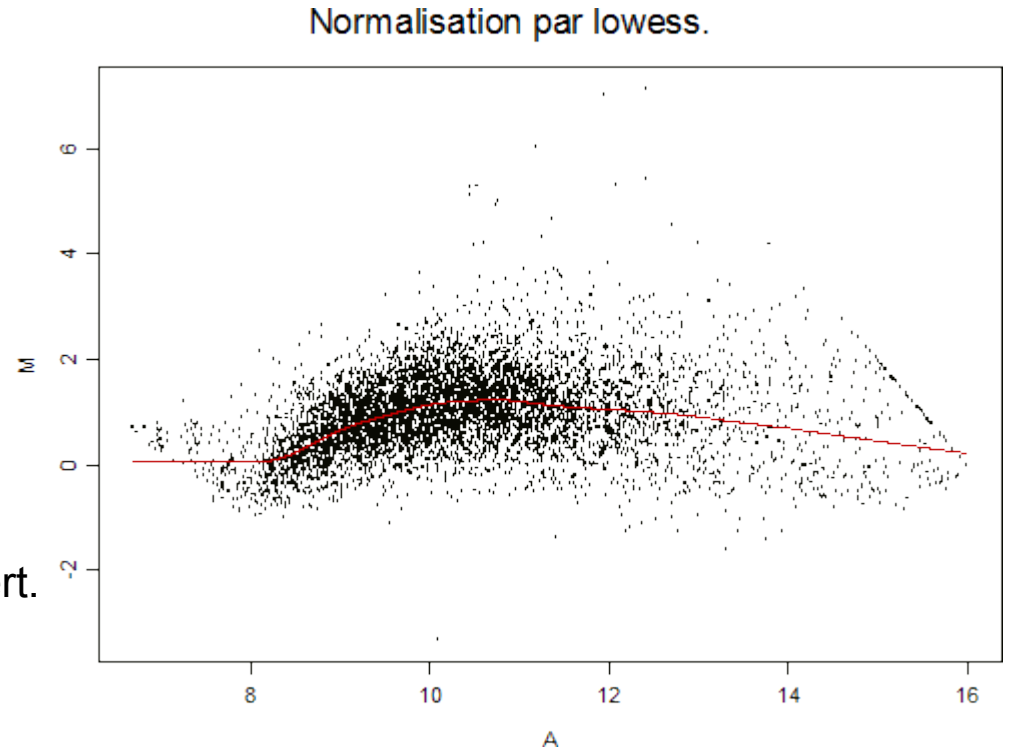
Normalisation des données

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

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

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

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



Les sources de problèmes

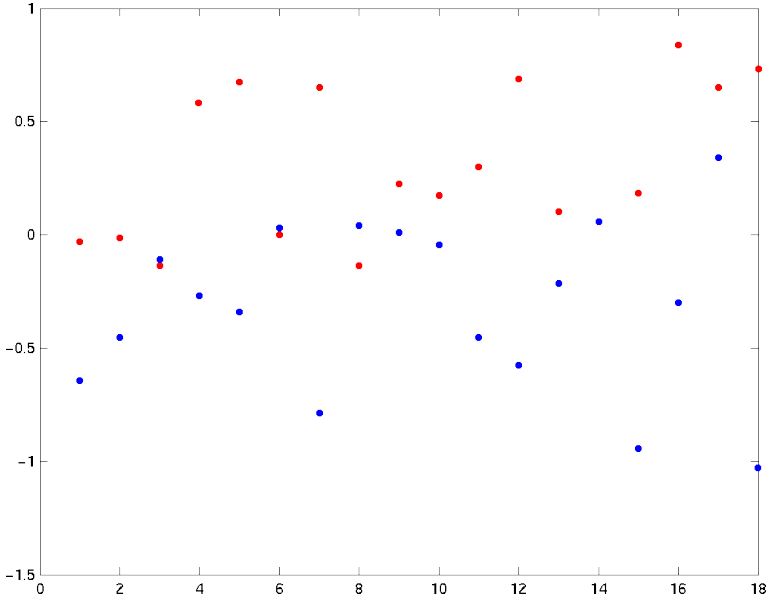
- **Présence de** bruit **dans les données à deux niveaux :**

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

Imprécision de la mesure : bruit classique

supposé gaussien, bruit qui est très élevé pour certains gènes (cf doubles mesures)

Présence de valeurs aberrantes dues à un problème lors de l'hybridation



- **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**

Démarche

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

- **Méthode directe de discrimination : illusoire**
  - Trop de « solutions »
  - Aucune garantie sur chacune d'elles

Prétraitement :

- **Sélection d'attributs**
  - Approche directe
  - Approche « wrapper »
  - Approche par **filtrage**
- **Réduction de dimensionnalité**
  - Groupement de gènes *a priori* (réseaux de régularisation)

Sélection

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

➥ **Comparaison à *hypothèse nulle globale***

➥ **Interprétation / confirmation par *les biologistes***

Utilisation d'ANOVA

- **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**

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

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

Utilisation d'ANOVA (suite)

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

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

SAM (Significance Analysis of Microarrays)

- **Pour chaque gène :**

$$d(i) = \frac{x_l(i) - x_{Nl}(i)}{s(i) + s_0}$$

déviation 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 Δ
- Calcul du nombre de gènes ***faussement significatifs*** : nombre moyen de gènes faussement significatifs pour chaque permutation
- ***Taux de fausse découverte*** (FDR)