Comparaison et combinaisons de

méthodes de sélection d'attributs

(pour l'analyse du transcriptome)

Antoine Cornuéjols , Romaric Gaudel

(*J-P. Comet, M. Dutreix, Ch. Froidevaux J. Mary1, G. Mercier*)

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

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

Plan

- Contexte
- Le pb de la sélection d'attributs
- Approche standard
- Combiner des méthodes
- Corrélation
- Combinaison
- Conclusion

Contexte : un pb d'analyse du transcriptome

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

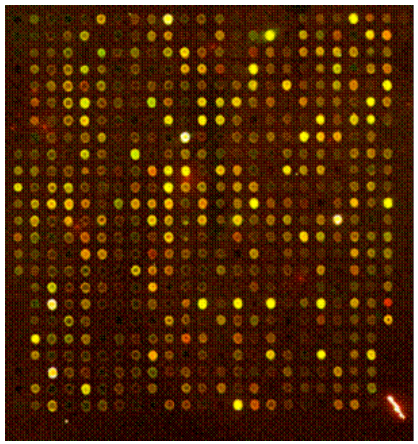
Étude de l'effet des très faibles radiations sur le génome

Etude des radiations

- **Danger indiscutable dans certains cas. En particulier pour les fortes doses d'irradiation.**
- **Quel impact des faibles doses ?**
- **Aucun détecté biologiquement**
- **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

- **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**

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

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

Définitions de la « pertinence »

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

- *Pas de définition unique car dépend du domaine*
  - Par rapport à la cible
    - d_i est pertinent si ∃ 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***

Objectifs de l'évaluation des attributs

- **Sélection**
  - D'un sous-ensemble d'attributs
- **Classement**
  - Calcule un score pour chaque attribut

Propriétés des attributs

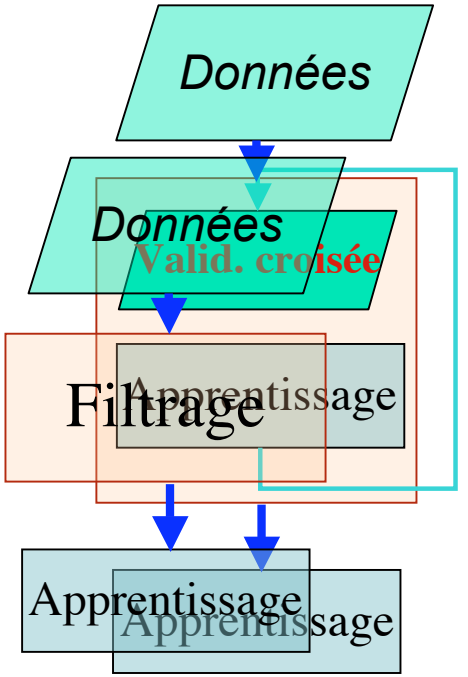
- **Bruités**
  - Imperfection des mesures
    - E.g. problèmes de normalisation
  - Contrôle imparfait des conditions expérimentales
- **Corrélés entre eux**
  - E.g. gènes codant deux bouts d'une même molécule
- **Décorrélés avec la classe**
  - E.g. gènes en rapport avec la taille lors d'une étude sur le cancer

Les approches

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

1. **Approche directe** (« embedded »)

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

- « **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

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

    - 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

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

Critères d'arrêt

- **Évaluation passant en-dessous d'un certain seuil**

- **Méthode par « témoins »**
  - Inclure des attributs aléatoires
  - Ne pas retenir les attributs dont l'évaluation est en-dessous

La sélection d'attributs en pratique

- **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

- **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

- **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-1,n-k)

SAM (Significance Analysis of Microarrays)

- **Pour chaque gène :**

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

déviation standard **Constante > 0**

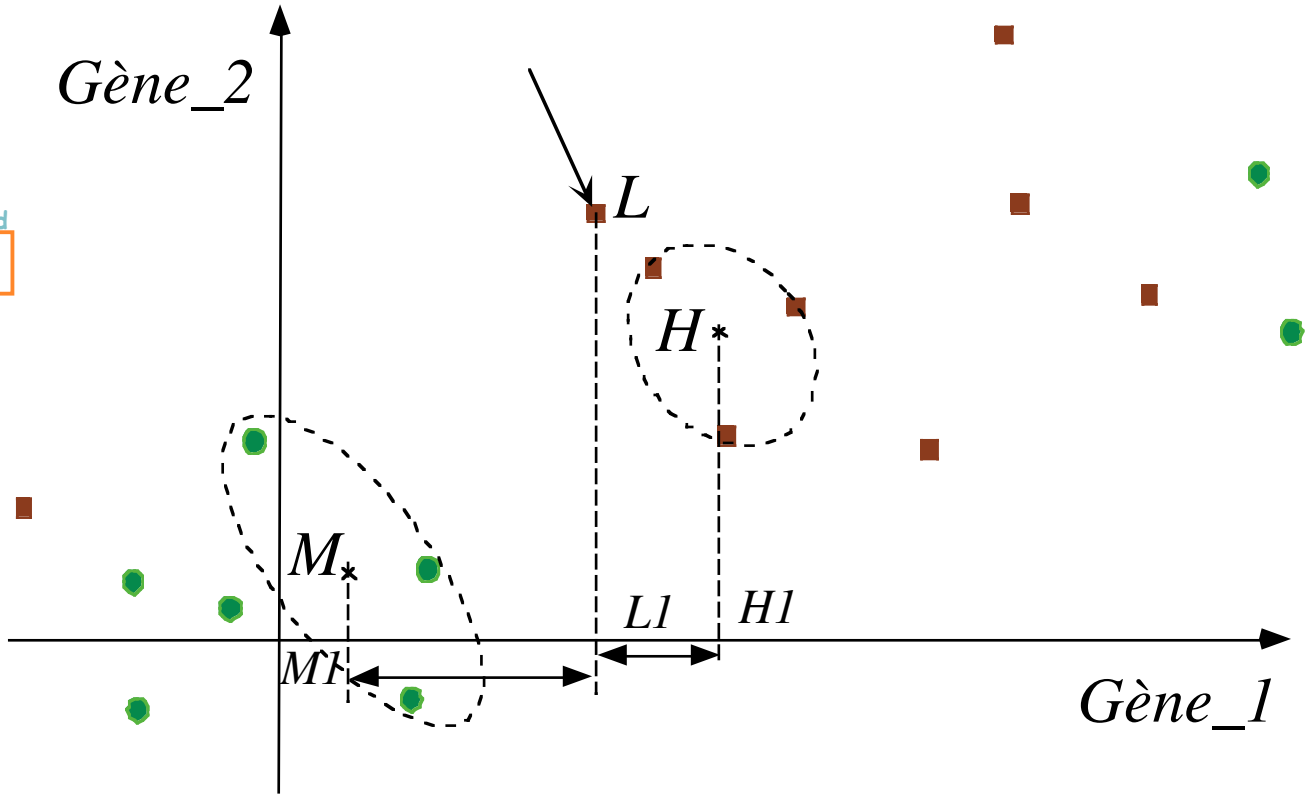
- ***Gènes potentiellement significatifs*** : gènes dont le score $d(g)$ est supérieur au score moyen du gène obtenu après permutations des classes, de plus d'un certain seuil Δ
- Calcul du nombre de gènes ***faussement significatifs*** : nombre moyen de gènes faussement significatifs pour chaque permutation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

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)

- Contexte
- Le pb de la sélection d'attributs
- Approche standard
- Combiner des méthodes
- Corrélation
- 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***.

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

où $\exp r_{gène}(x)$ **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)

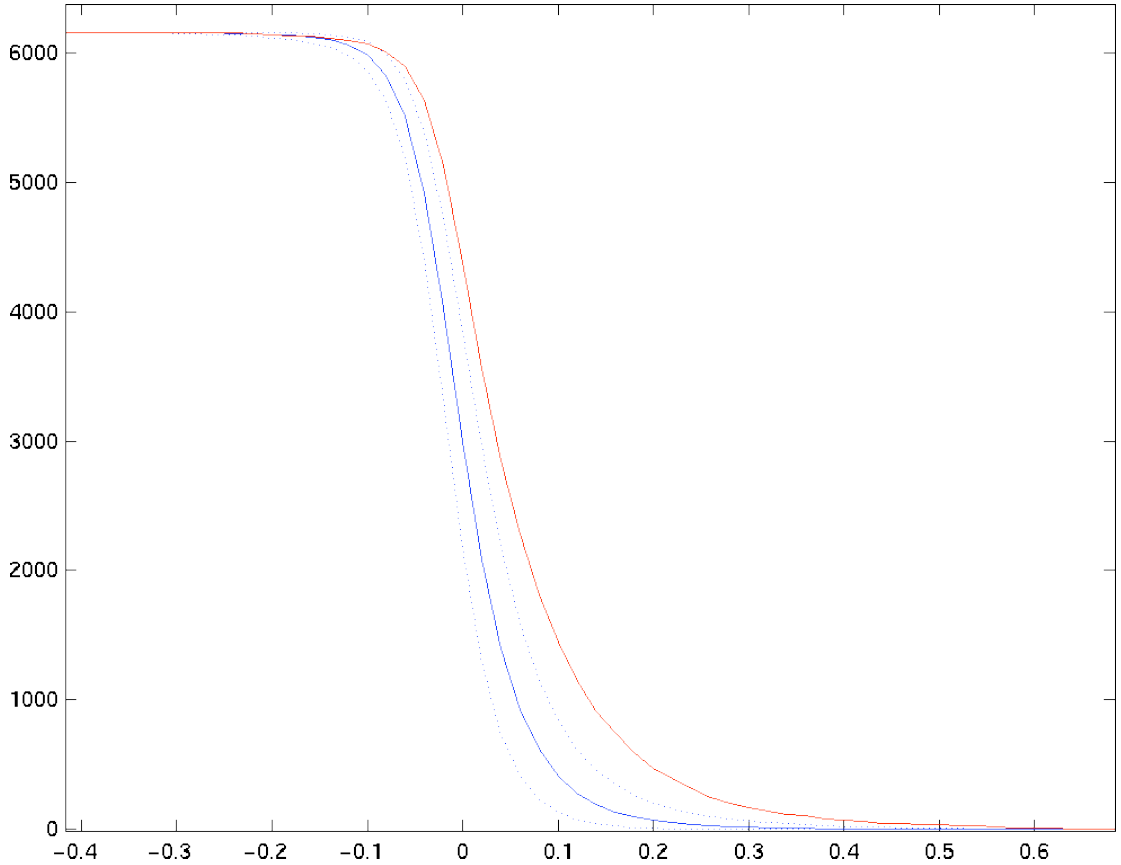
  - **Algorithme polynomial :** $\Theta(pm^2)$
  - **Rôle de k :** **prise en compte du bruit**

Sélection des attributs

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

Hypothèse nulle globale

- Contexte
- Le pb de la sélection d'attributs
- Approche standard
- Combiner des méthodes
- Corrélation
- Combinaison
- Conclusion



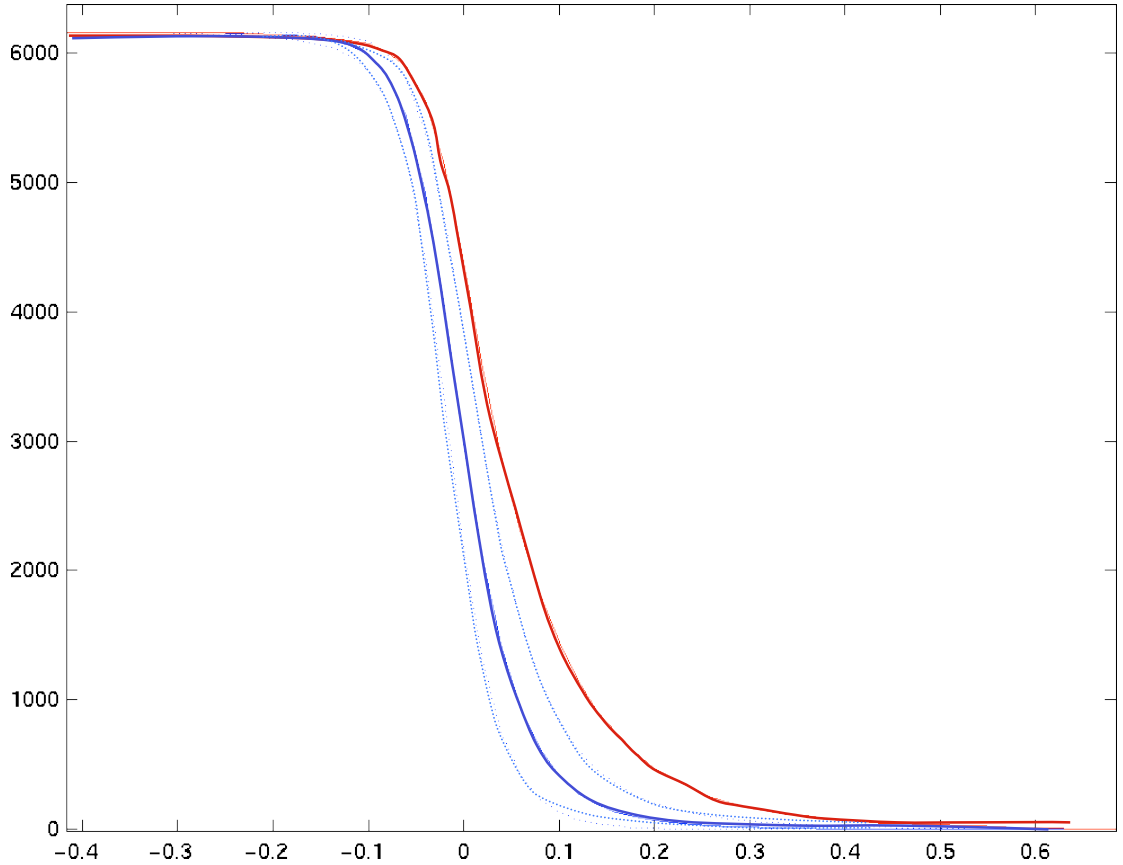
Nombre de gènes dont le poids dépasse la valeur repérée en abscisse

rouge : Avec les classes réelles ;

bleu : Courbe moyenne obtenue avec des classes aléatoires

Hypothèse nulle globale

- Contexte
- Le pb de la sélection d'attributs
- Approche standard
- Combiner des méthodes
- Corrélation
- Combinaison
- Conclusion



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

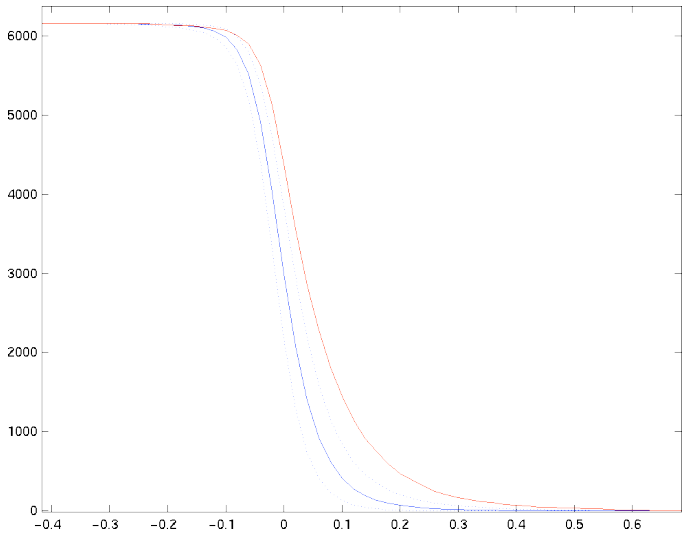
rouge : Avec les classes réelles ;

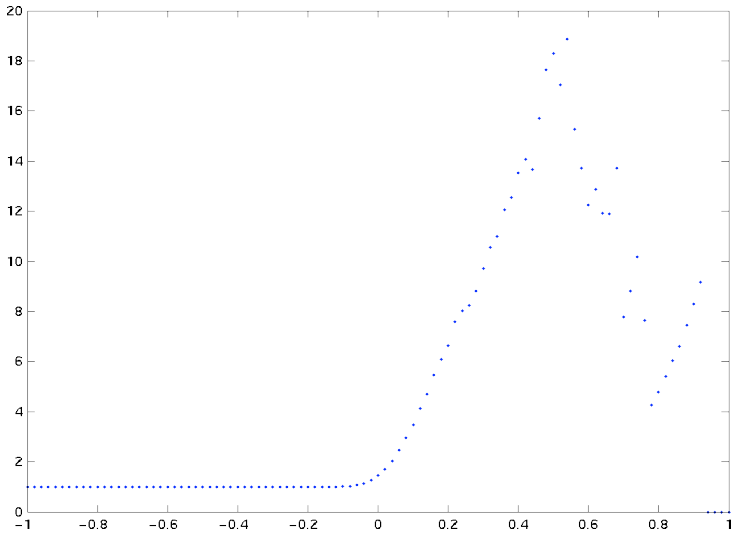
bleu : Courbe moyenne obtenue avec des classes aléatoires

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 impliquées de manière quasi-certaine dans la réponse à l'Irradiation (quitte à ne pas avoir tous les gènes impliqués)

➠ Problème du seuil

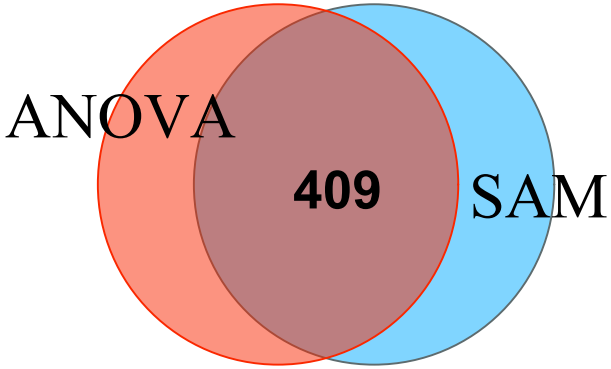


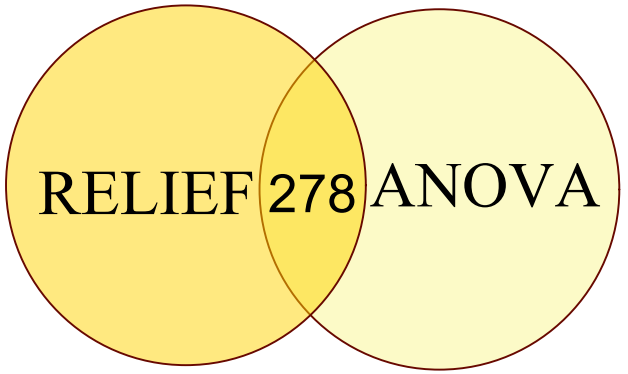


Intersections (1)

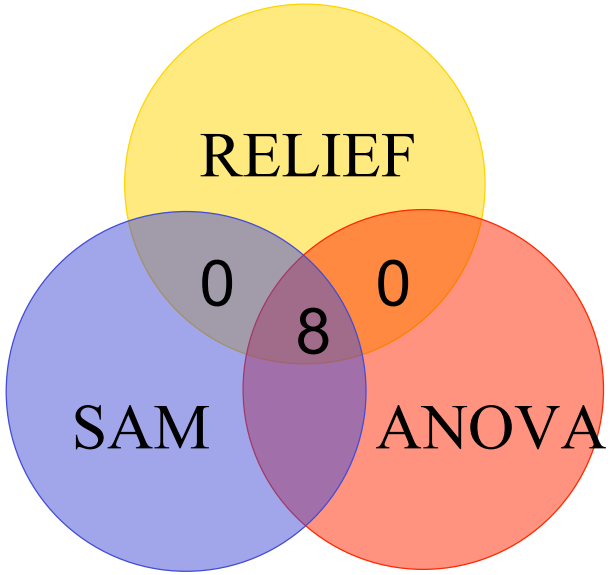
- Contexte
- Le pb de la sélection d'attributs
- Approche standard
- Combiner des méthodes
- Corrélation
- Combinaison
- Conclusion

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 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 ?

  = = > ***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

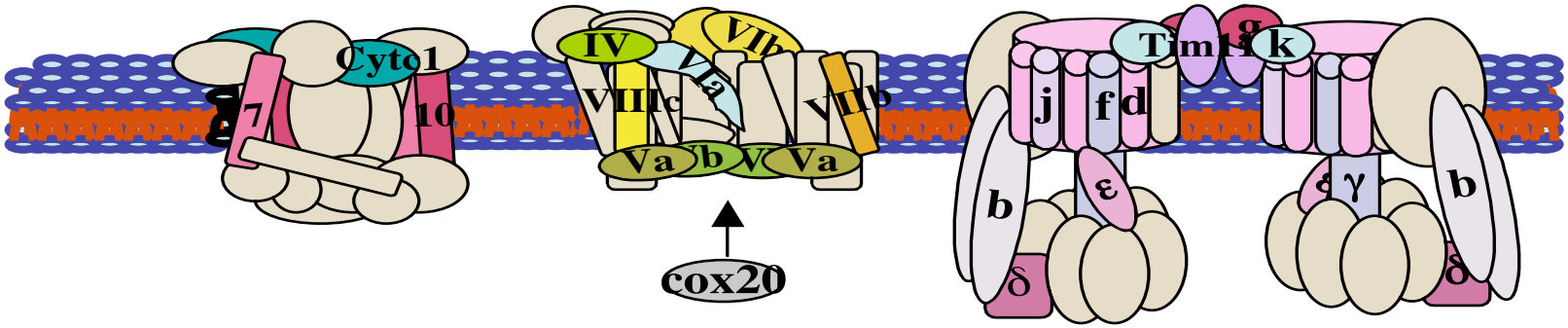
- Contexte
- Le pb de la sélection d'attributs
- Approche standard
- Combiner des méthodes
- Corrélation
- Combinaison
- Conclusion

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

- Contexte
- Le pb de la sélection d'attributs
- Approche standard
- Combiner des méthodes
- Corrélation
- Combinaison
- Conclusion

Cytochrome bc1	Cytochrome c 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



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 ?

Combinaison de méthodes ?

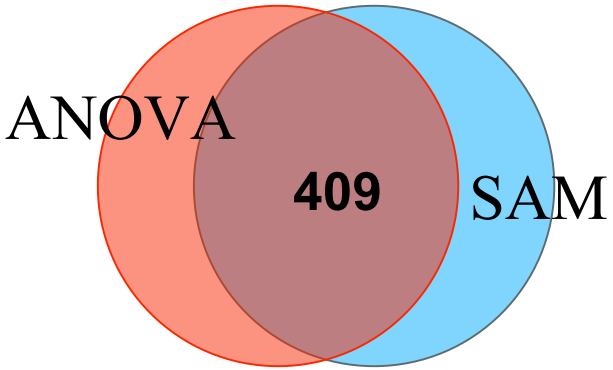
- **Intuition :**

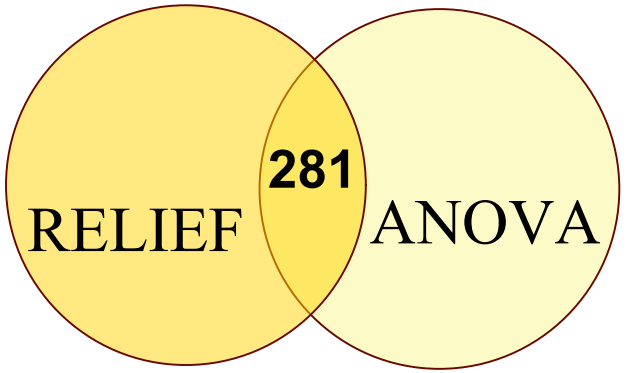
L'intersection des top-n suit

une certaine loi tant qu'il y a des attributs pertinents,

et une autre après

Mesure de corrélation entre méthodes





Trois causes possibles pour l'intersection :

1. Le **hasard**. Tirage aléatoire de deux sous-ensembles de n (e.g. 500) éléments parmi d (e.g. 6135)

$$P(k|n,d) = \frac{\binom{n}{k} \cdot \binom{d-n}{n-k}}{\binom{d}{n}}$$

2. La **corrélation des méthodes *a priori***.
3. Les **régularités dans les données** sur lesquelles les méthodes sont d'accord, au-delà de leur corrélation *a priori*.

Mesure de corrélation entre méthodes (2)

- **Exemple**
  - **281 gènes dans (RELIEF ∩ ANOVA)$_{500}$**
  - 40 attendus par simple chance (loi hypergéométrique)
  - + 241 ?
    - Information ?
    - Corrélation *a priori* ?

Mesure de la corrélation a priori

- Contexte
- Le pb de la sélection d'attributs
- Approche standard
- Combiner des méthodes
- Corrélation
- Combinaison
- Conclusion

Nouvelle hypothèse nulle

- Pour toutes les permutations de 6 + & 12 - sur les données
- Calculer : $(\mathbf{RELIEF} \cap \mathbf{ANOVA})_{500}$
- 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 : ?
  - 281 : ?
  - 500 : ?

Ici : **180** ± 40

Application aux données

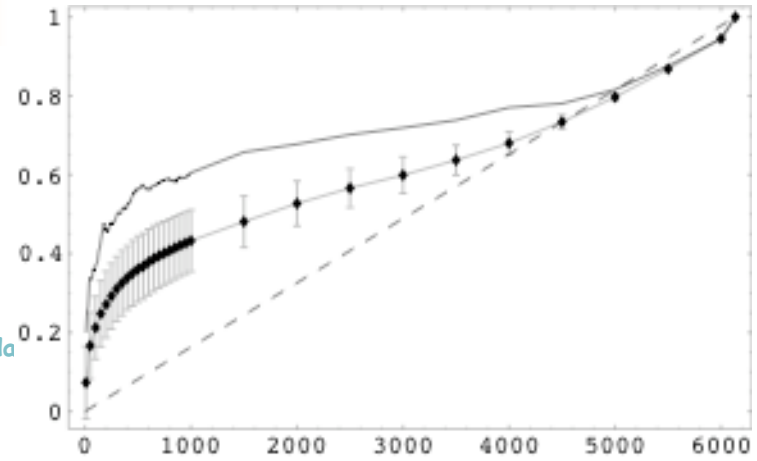
k : taille de l'intersection des top-n pour ANOVA et RELIEF sur les données pour plusieurs valeurs de n.

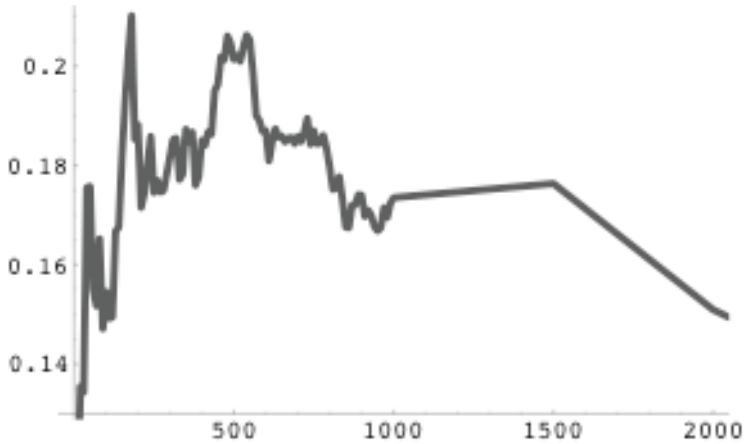
$\mu_{\mathcal{H}0}$: intersection mesurée pour les données étiquetées aléatoirement (mesure de corrélation *a priori*)

$\sigma_{\mathcal{H}0}$: écart-type.

n	100	200	300	400	500	600	700	800	900	1000
$\mu_{\mathcal{H}_0}$	**21.2**	**54.2**	**93.2**	**135.4**	**180.3**	**226.9**	**276.3**	**326.2**	**378.9**	**432.5**
$\sigma_{\mathcal{H}_0}$	8.0	16.9	24.5	32.3	41.8	50.3	57.7	64.1	71.3	78.0
k	**37**	**93**	**149**	**210**	**281**	**339**	**406**	**470**	**535**	**605**

Types d'information dans les données





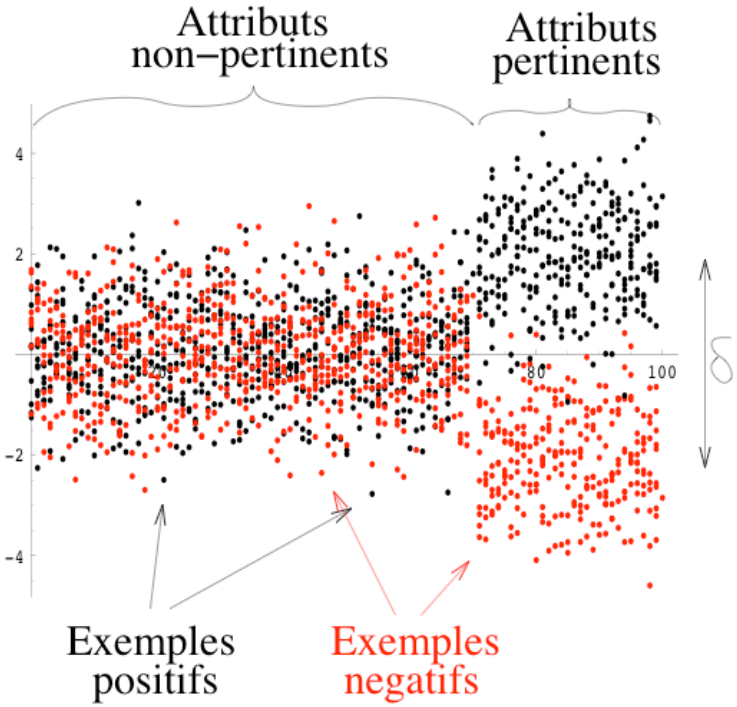
(À gauche) Valeur de n en abscisse et taille k de l'intersection en proportion de n en ordonnée (e.g. $k = 0.6\ n$). Courbe du haut : k, du milieu : $\mu_{\mathcal{H}0}$ (corrélation *a priori*), du bas : taille de l'intersection due au hasard.

(À droite) Différence relative entre k et $\mu_{\mathcal{H}0}$.

L'information apportée par les données
est donc maximale pour $n \approx 180$ et $n \approx 540$.

Sur des données artificielles

- 20 exemples (10 positifs, 10 négatifs)
- **d** : nombre d'attributs par exemple ($d = 1000$)
- **p** : nombre d'attributs pertinents ($p \in [|50, 400|]$)
- Attributs suivent une loi gaussienne
- σ : variance des attributs ($\sigma \in [1, 5]$)
- δ : écart entre la moyenne des attributs pertinents positifs et la moyenne des attributs pertinents négatifs ($\delta \in [0.1, 5]$)



Effet de p sur k (l'intersection des top-n)

- Courbes différentes suivant le nombre p d'attributs pertinents
- $\Rightarrow$ On peut déterminer p à partir de l'étude de la corrélation entre deux algorithmes

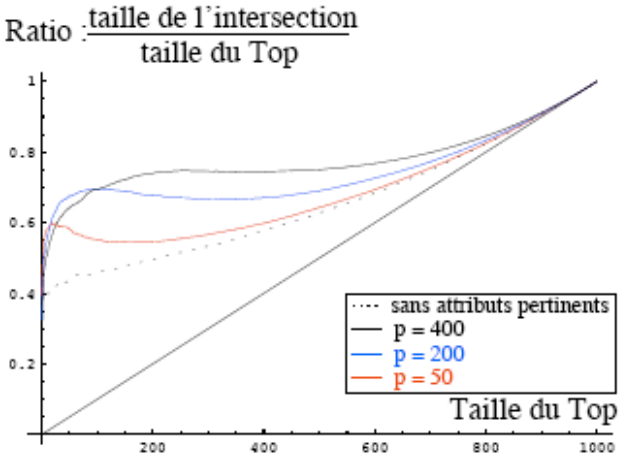


Fig.: Variation de l'intersection en fonction du nombre d'attributs pertinents ($\delta = 1$)

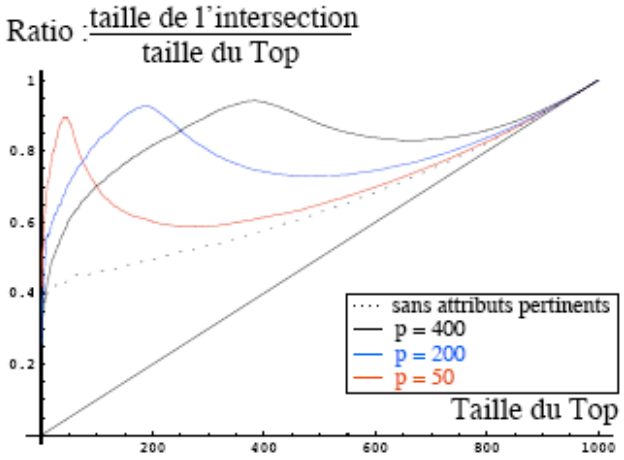


Fig.: Variation de l'intersection en fonction du nombre d'attributs pertinents ($\delta = 2$)

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

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 :

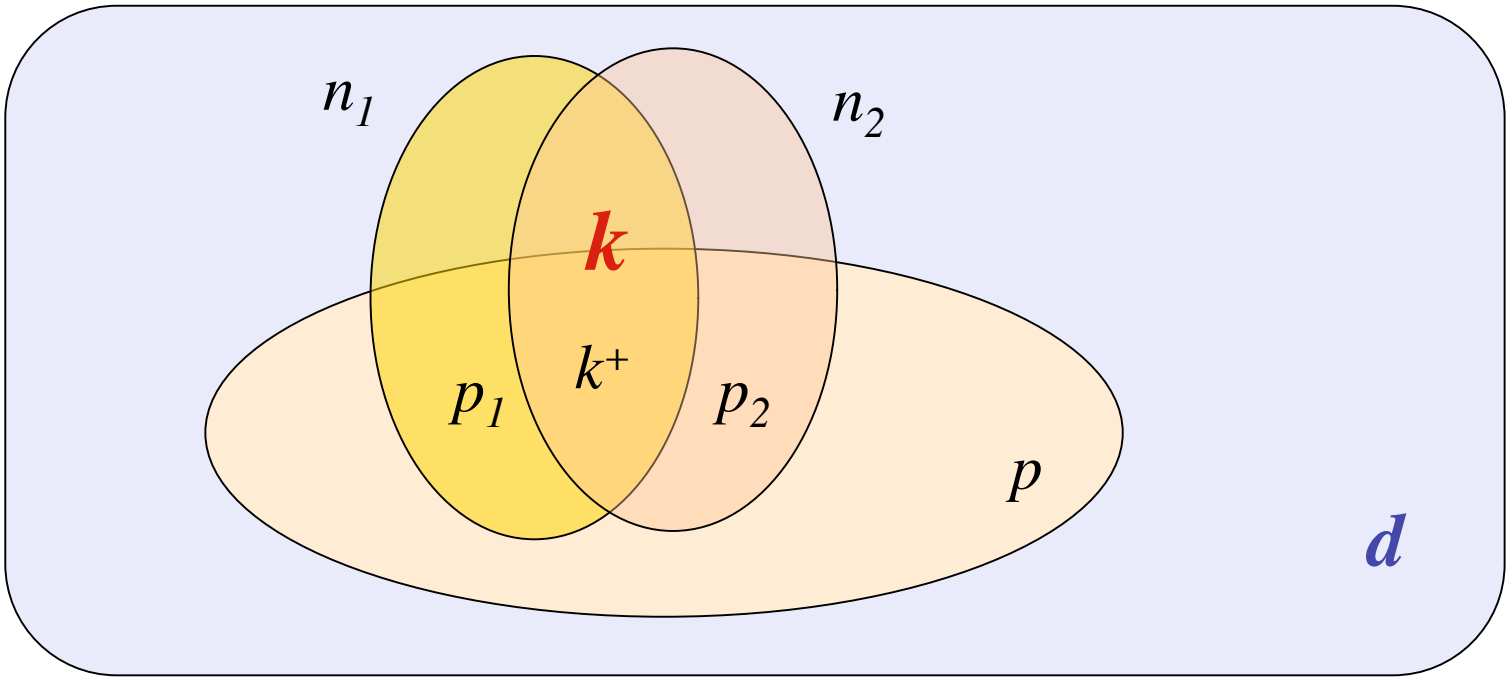
$$k = \mathbf{fct}(d, n, k_{corr}, p, p_1, p_2)$$

2. *Principe de maximum de vraisemblance*

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

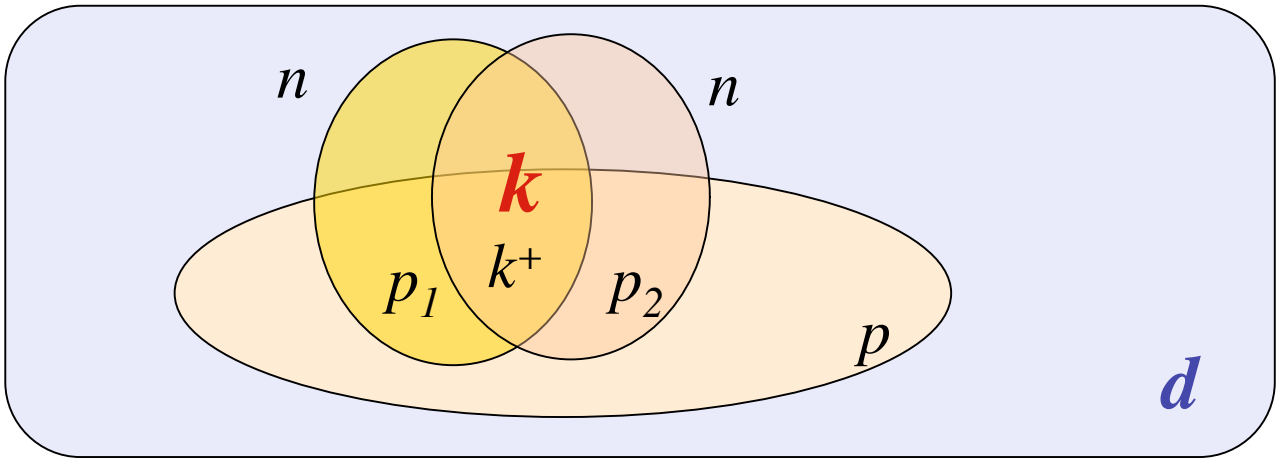
Combinaison de méthodes

- Contexte
- Le pb de la sélection d'attributs
- Approche standard
- Combiner des méthodes
- Corrélation
- Combinaison
- Conclusion



Combien parmi les k sont positifs ?

Combinaison de méthodes



de sac-1 : $$N_1 = \frac{\binom{p}{p_1} \cdot \binom{d-p_1}{n-p_1}}{\binom{d}{n}}$$

de sac-2 : $$N_2 = \frac{\sum_{k^+=0}^{k} \binom{p_1}{k^+}\binom{p-p_1}{p_2-k^+}\binom{n-p_1}{k-k^+}\binom{d-n-(p-p_2)}{n-p_1-(k-k^+)}}{\binom{d}{n}}$$

Combinaison de méthodes

$$P(k|d,n,p,p_1) = \begin{cases} 0 & \text{si } k < \mathbf{k_{corr}(n)} \\ \frac{N_1 N_2}{\text{terme de normalisation}} & \text{sinon} \end{cases}$$

À estimer

connus

mesuré

- **Comment calculer ce terme de normalisation ?**
  - Fonction de la corrélation *a priori*

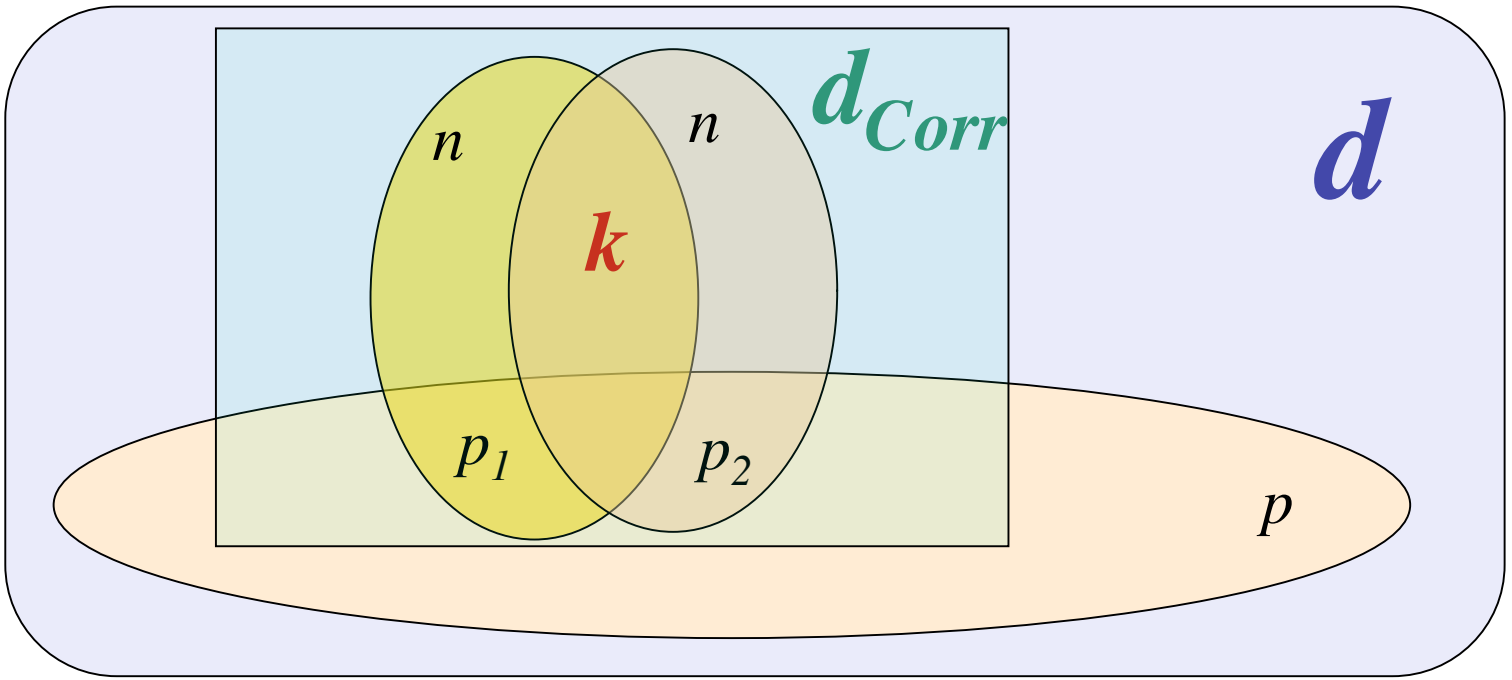
Terme de corrélation a priori

- Contexte
- Le pb de la sélection d'attributs
- Approche standard
- Combiner des méthodes
- Corrélation
- Combinaison
- Conclusion

$$\frac{\binom{p}{p_1} \cdot \binom{d-p_1}{n-p_1}}{\binom{d}{n}} \cdot \frac{\sum_{k^+=\mathbf{k_{Corr}}}^{k} \binom{p_1}{k^+}\binom{p-p_1}{p_2-k^+}\binom{n-p_1}{k-k^+}\binom{d-n-(p-p_2)}{n-p_1-(k-k^+)}}{\binom{d}{n}}$$

Pas d'intersection $k < k_{Corr}$

Terme de corrélation a priori



➠ Permet aussi de rendre compte d'une corrélation négative

$$P(k|d, n, p, p_1, d_{Corr})$$

Tests sur données artificielles

- Objectif des tests :
  - Vérifier l'égalité des courbes d'intersection :
    - mesurées pour des données artificielles
    - prédites par le modèle
- Paramètres des tests :
  - Données artificielles générées comme pour la première étude
  - $p \in [|50, 400|]$
  - $\sigma = 1$
  - $\delta \in [0.1, 5]$
  - m mesuré sur les données

Tests sur données artificielles

- m est la proportion d'attributs pertinents dans le Top_n
- Donc m est mesuré sur les données artificielles
- Problème : m diffère suivant l'algorithme

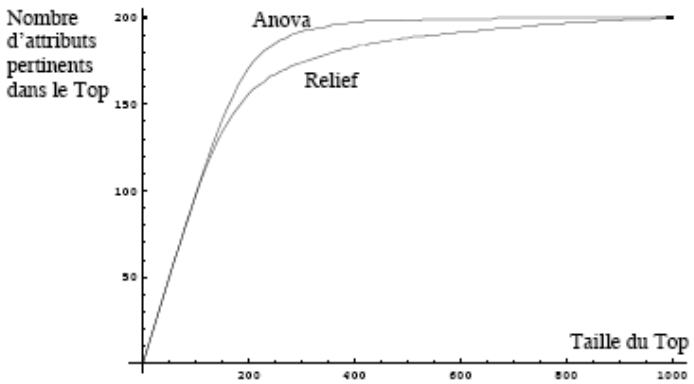


Fig.: Valeurs moyennes de m mesurées pour 1000 tirage ($p = 200$, $\sigma = 1$, $\delta = 1.5$)

- $\Rightarrow m = \frac{m_{Relief} + m_{Anova}}{2}$

Tests sur données artificielles

- Modèle imparfait pour les petites tailles d'intersection
- Modèle bon à partir du moment où m est proche de p

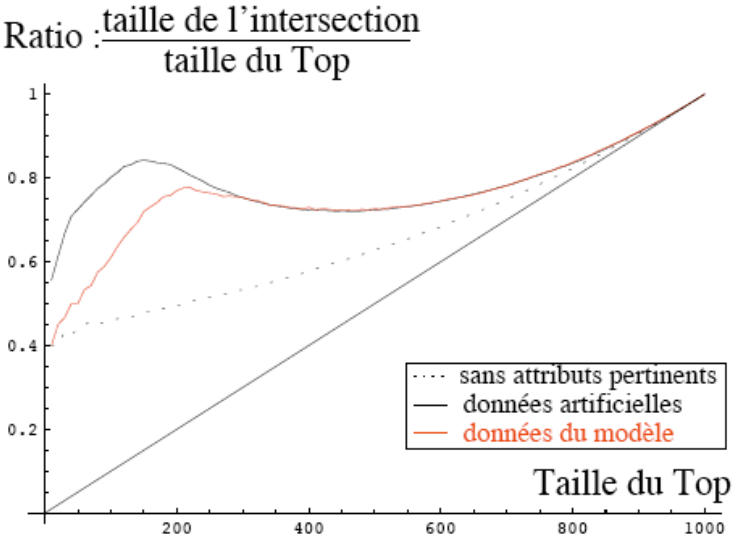


Fig.: Comparaison entre le modèle et les données artificielles

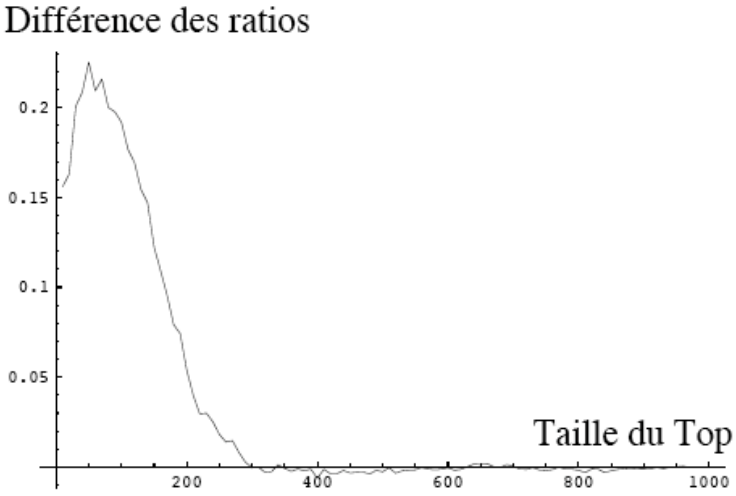


Fig.: Valeur sur les données artificielles moins valeur prédite par le modèle

Tests : Conclusions

- Modèle ***imparfait*** pour les petites tailles d'intersection
- Modèle ***correct*** pour $p_1 \geq p$

➠

  - L'intersection entre attributs non pertinents est bien modélisée
  - L'intersection entre attributs pertinents doit être modélisée différemment

    ➠ Revoir la corrélation *a priori*

Conclusions / Perspectives

- *Fixer le seuil de pour la sélection d'attributs est difficile*

Approche originale : combinaison de méthodes

Conclusions

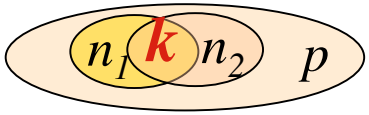
- La corrélation (taille intersection) dépend de p (# attributs pertinents) ***dépendance exploitable***
- Le modèle génératif est encore imparfait

Perspectives

- Tester sur des données réelles
- Étendre à corrélations à > 2 algorithmes
  - (cf. rank boosting [ECML-04])

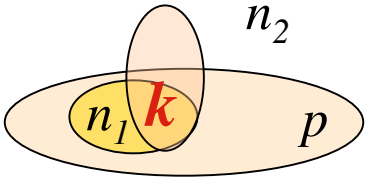
Formules

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



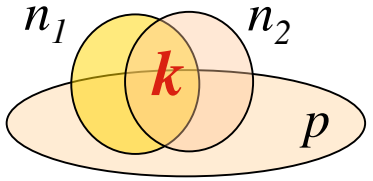
$$p(\cap = k|d, p, n_1 = n_2 = n, k_C) = \binom{n}{k}\binom{p-n}{n-k} / \sum_{i=k_C}^{n} \binom{n}{i}\binom{p-n}{n-i}$$

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



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

- $p_1 \leq p_2 \leq n$



$$p(\cap = k|d, p, n_1 = n_2 = n, k_C) = \frac{\binom{p}{p1}\binom{d-p}{n-p_1}\sum_{k^+=\max(0,p_1+p_2-p)}^{p_1}\binom{p_1}{k^+}\binom{p-p_1}{p_2-k^+}\binom{n-p_1}{k-k^+}\binom{d-n-(p-p_1)}{n-p_2-(k-k^+)}}{\binom{d}{n}^2} \Big/ \sum_{i=k_C}^{n} \ldots$$

Et alors …

- **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

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

- 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

- 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

- Contexte
- Le pb de la sélection d'attributs
- Approche standard
- Combiner des méthodes
- Corrélation
- Combinaison
- Conclusion

Plusieurs techniques ont été utilisées

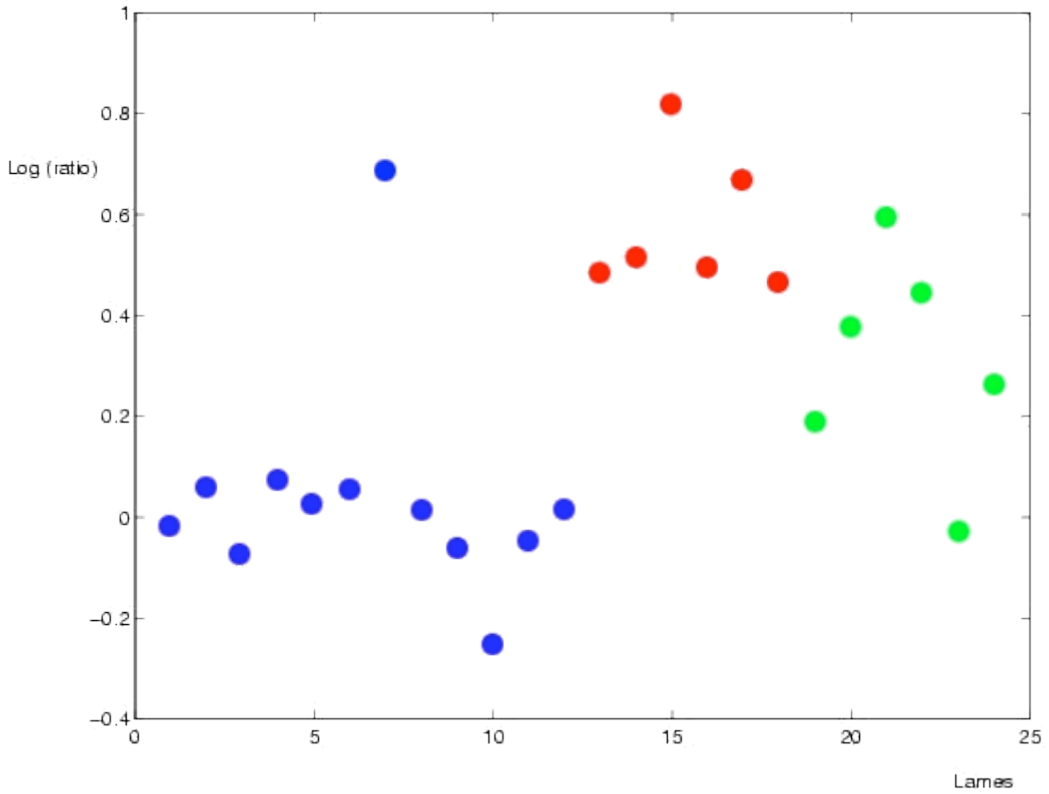
- Vote « d'experts »
- Technique du maximum de vraisemblance
- *K* plus proches voisins

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

- Contexte
- Le pb de la sélection d'attributs
- Approche standard
- Combiner des méthodes
- Corrélation
- Combinaison
- Conclusion



Travaux en cours

- **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

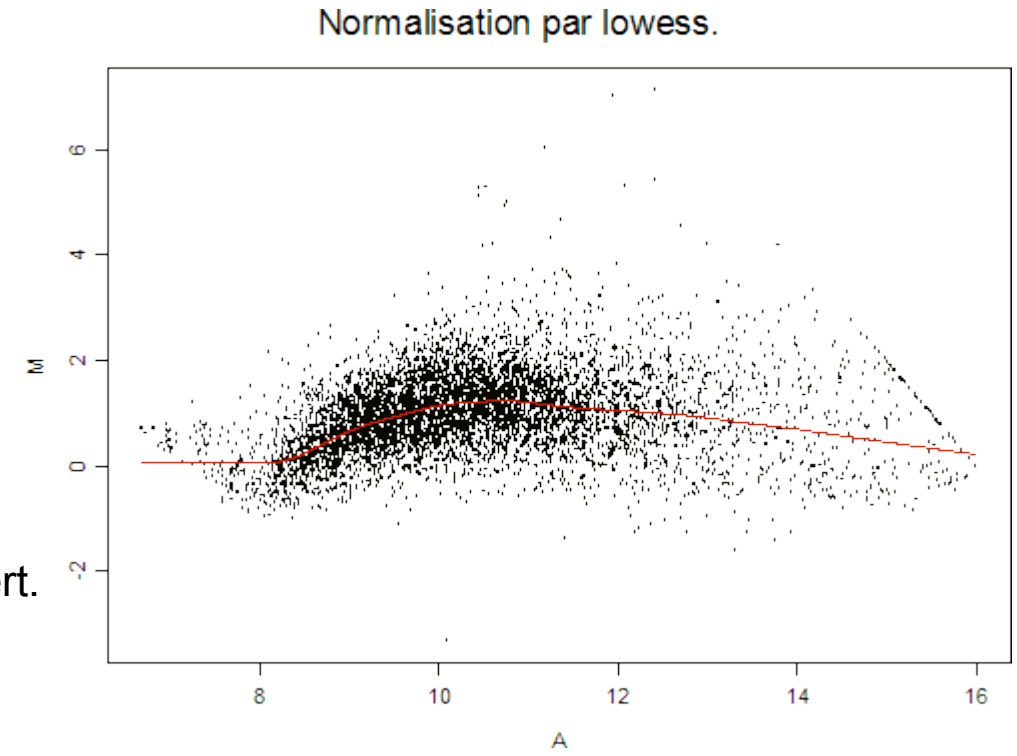
- Contexte
- Le pb de la sélection d'attributs
- Approche standard
- Combiner des méthodes
- Corrélation
- Combinaison
- Conclusion

- **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.



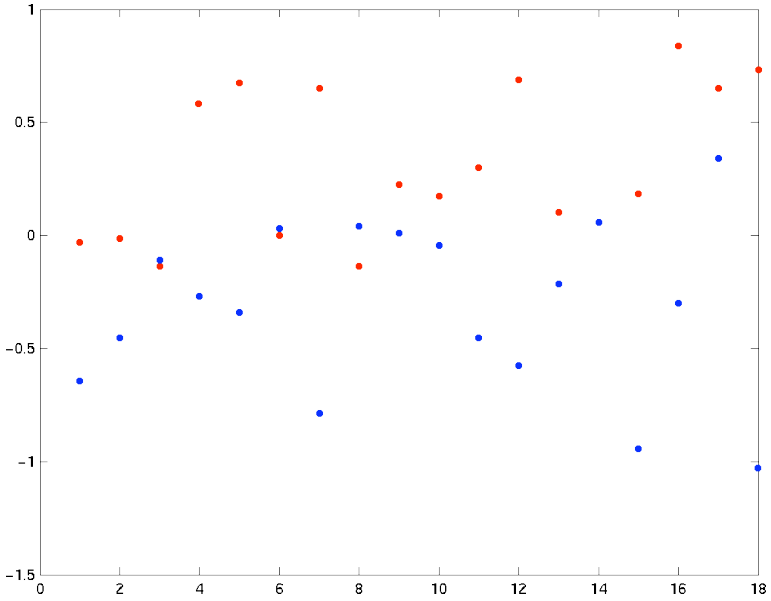
Les sources de problèmes

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

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**

Sélection

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

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

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

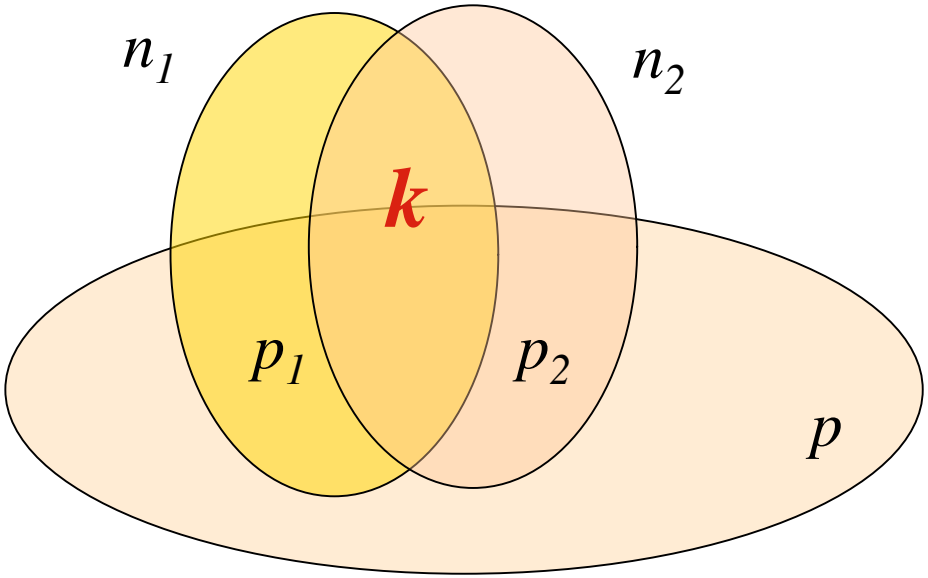
Utilisation d'ANOVA (suite)

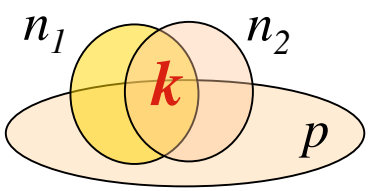
On peut aussi calculer la *p-value* pour chaque gène et ordonner les gènes

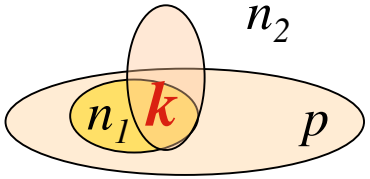
Probabilité que le test rejette l'hypothèse $\mathcal{H}_0$ à tort

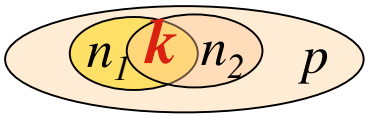
$$p(t) \;=\; \min\left\{F_0(t),\; 1 - F_0(t)\right\}$$

- Contexte
- Le pb de la sélection d'attributs
- Approche standard
- Combiner des méthodes
- Corrélation
- Combinaison
- Conclusion









- Contexte
- Le pb de la sélection d'attributs
- Approche standard
- Combiner des méthodes
- Corrélation
- Combinaison
- Conclusion