

AXE BIOINFORMATIQUE

Recherche d'attributs pertinents pour l'apprentissage de régularités en génomique

JÉRÉMIE MARY

Stage de DEA sous la direction d'Antoine CORNUÉJOLS et de Christine FROIDEVAUX

5 mars 2002 - 30 juin 2002

Plan de l'exposé

1

- Description du problème
- Mise en oeuvre de techniques classiques
- Introduction d'hypothèses à tester sur les données et contre le hasard
- Résultats et validation

Cadre du problème

2

But : analyser l'effet de très faibles doses de radiations.

On dispose de **mesures d'expression** du génome de levures :

- **saines**
- **irradiées**

On cherche à analyser le *transcriptome*, plus particulièrement, on est face à un problème **d'apprentissage supervisé** dont le but est la **classification** .

Questions des biologistes (Institut Curie)

3

- Les mesures effectuées sont-elles *utiles* pour détecter l'irradiation des levures?
- *Nombre* de gènes impliqués dans la réponse à une irradiation?
- *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?
- Quelles *erreurs* peuvent être commises lors de la prédiction et que faut-il faire pour les réduire?
- *Combien* d'expériences faudrait-il réaliser pour avoir confiance dans les résultats?

Caractéristiques des données

4

- Présence de **bruit** dans les données à deux niveaux :
 - Imprécision de la mesure (bruit classique supposé gaussien), ce bruit est très élevé pour certains gènes (cf doubles mesures);
 - Présence de valeurs aberrantes dues à un problème lors de l'hybridation ;
- **Nombreux attributs** : 6157 gènes ;
- Très **faible nombre d'exemples** : 12 cultures non-traitées, 6 irradiées.

⇒ Impossible de traiter l'apprentissage de manière générique. On va travailler sur **l'instanciation** du problème.

Différents algorithmes

5

- Outils **statistiques** : Significance Analysis for Microarrays (SAM), Analysis of Variance (ANOVA).
 - Peu de souplesse dans la mise en oeuvre ;
 - Fortement perturbés par les valeurs aberrantes.
- Algorithmes **d'apprentissage** : clustering, “wrapper methods”...
 - On est en dehors du domaine d'application habituel de ces méthodes: (rapport nb attributs / nb exemples trop grand) ;
 - Les résultats obtenus par les différentes techniques ne sont pas les mêmes et on ne sait pas comment conclure ;
 - De nombreux attributs sont non-pertinents: il semble donc utile d'appliquer un filtre sur les attributs.

- Application d'un **“cut-off”** sur les niveaux d'expression des gènes.
 - Systématiquement utilisé par les biologistes ;
 - Choix du seuil empirique.
- **FOCUS** : totalement inadapté car très sensible au bruit et impossible à mettre en œuvre pour plus d'une trentaine d'attributs ;
- **RELIEF** : intéressant car résistant au bruit et l'application n'est pas trop gênée par le grand nombre d'attributs.

Présentation de RELIEF (Kira & Rendel 92)

7

On estime pour chaque attribut A , un poids :

$$\text{Poids}(A) = P(\text{Valeur différente de } A \text{ pour l'instance la plus proche de la classe opposée à } A) - P(\text{Valeur différente de } A \text{ pour l'instance la plus proche de la même classe que } A).$$



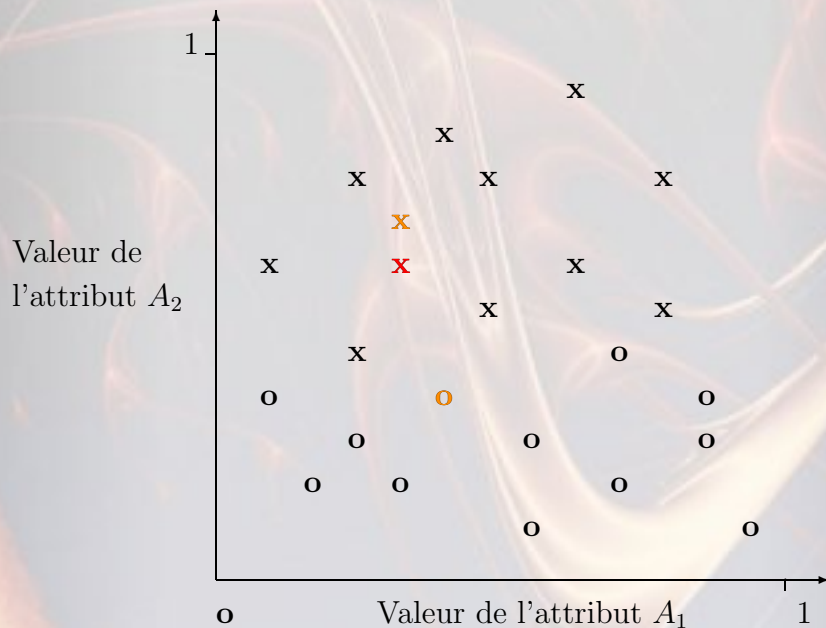
x : exemple positif ; o : exemple négatif

Présentation de RELIEF (Kira & Rendel 92)

7

On estime pour chaque attribut A , un poids :

$$\text{Poids}(A) = P(\text{Valeur différente de } A \text{ pour l'instance la plus proche de la classe opposée à } A) - P(\text{Valeur différente de } A \text{ pour l'instance la plus proche de la même classe que } A).$$



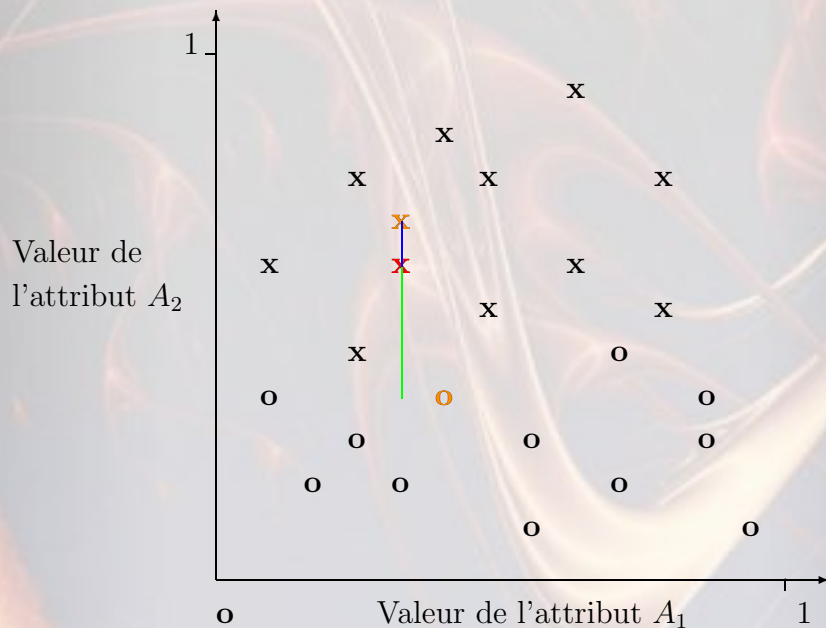
x : exemple positif ; o : exemple négatif

Présentation de RELIEF (Kira & Rendel 92)

7

On estime pour chaque attribut A , un poids :

$$\text{Poids}(A) = P(\text{Valeur différente de } A \text{ pour l'instance la plus proche de la classe opposée à } A) - P(\text{Valeur différente de } A \text{ pour l'instance la plus proche de la même classe que } A).$$



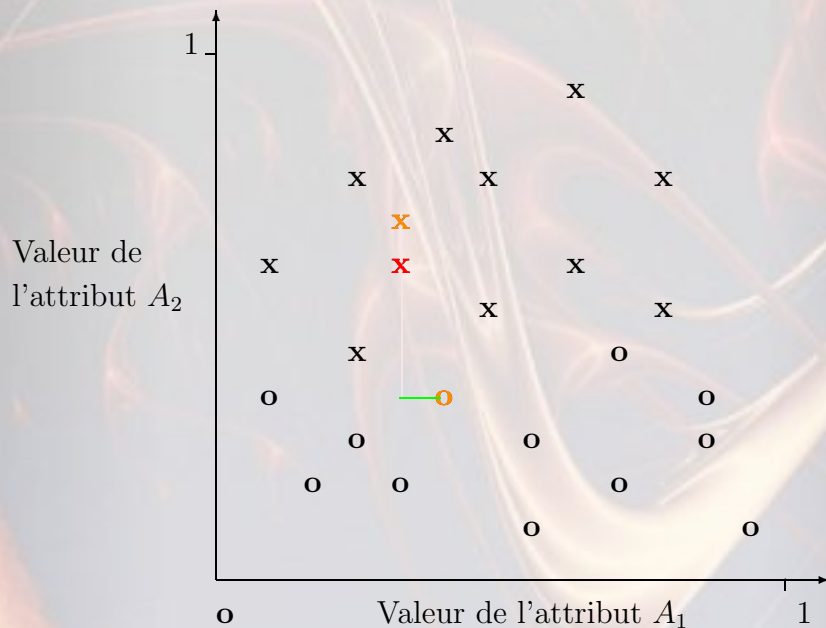
x : exemple positif ; o : exemple négatif

Présentation de RELIEF (Kira & Rendel 92)

7

On estime pour chaque attribut A , un poids :

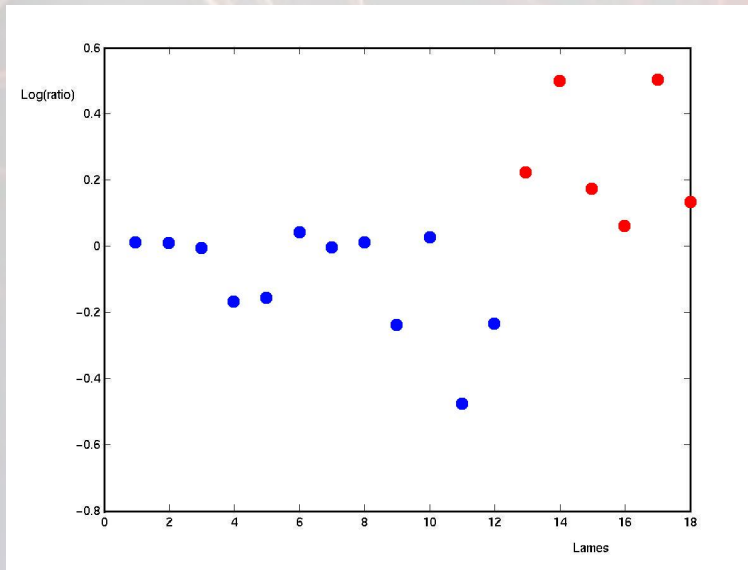
$$Poids(A) = P(\text{Valeur différente de } A \text{ pour l'instance la plus proche de la classe opposée à } A) - P(\text{Valeur différente de } A \text{ pour l'instance la plus proche de la même classe que } A).$$



x : exemple positif ; o : exemple négatif

Un premier exemple

Hypothèse : pour certains gènes, il existe un seuil permettant de classer parfaitement les données ; ex: expression du gène n^o 509.



– On trouve 23 gènes qui ont cette propriété.

Est-ce significatif?

Repérons par **x** les cultures saines et par **o** les levures irradiées.
Si on réorganise les expériences selon l'ordre donné par le niveau d'expression d'un gène on obtient par exemple :

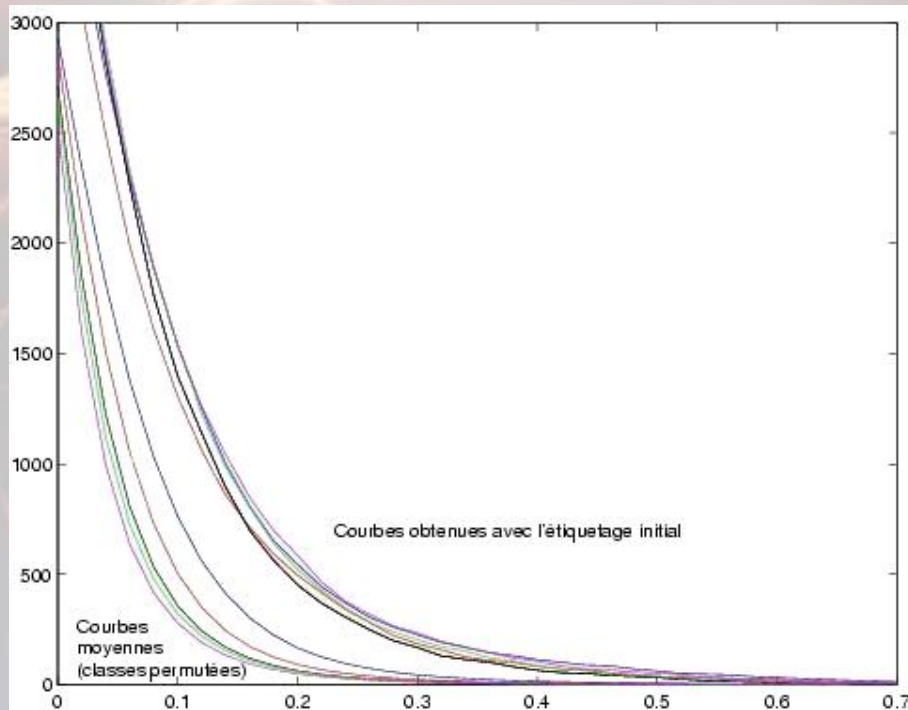
x x x o x o o x x x x o x o x x x o

Il y C_{18}^6 répartitions possibles.

Si de plus *les gènes sont indépendants* et qu'il n'y a *pas d'influence de la classe sur leur expression* (hypothèse nulle) alors toutes les répartitions sont *équiprobables*.

Donc la probabilité pour un gène de classer parfaitement les expériences est

$$\frac{2}{C_{18}^6}$$

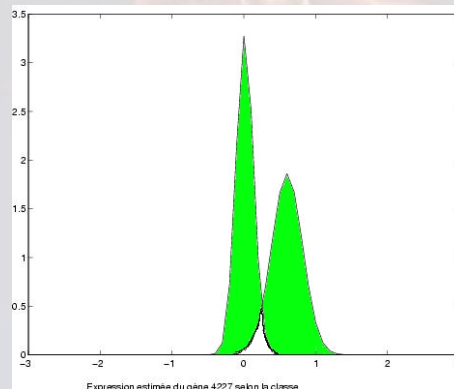


- Nb de gènes G tq $Poids(G) > seuil$
- Courbes moyennes : les classes sont permutées 1000 fois
- Nombre de voisins utilisés : 1(bleu), 2(rouge), 3(noir), 4(vert), 5(magenta).

$$\text{Pertinence}(G) = \frac{1}{2} \int_{-\infty}^{+\infty} f_S(x) \mathbb{I}_{f_S(x) > f_I(x)} + f_I(x) \mathbb{I}_{f_S(x) < f_I(x)} dx$$

avec

- f_S densité de l'expression du gène G pour les levures saines : $\mathcal{N}(m_S, \sigma_S)$.
- f_I densité de l'expression du gène G pour les levures irradiées : $\mathcal{N}(m_I, \sigma_I)$.



| Classe | Expression des gènes utilisés |
|--------|---|
| S | $\mathcal{N}(m_1, \sigma_1) \dots \mathcal{N}(m_k, \sigma_k)$ |
| I | $\mathcal{N}(m_1^*, \sigma_1^*) \dots \mathcal{N}(m_k^*, \sigma_k^*)$ |

La probabilité E de dire qu'une culture $(x_1 \dots x_k)$ est saine alors qu'elle est en fait irradiée est :

$$\begin{aligned}
 E &= P\left(\prod_{i=1}^k f(x_i) > \prod_{i=1}^k f^*(x_i) / \forall i, x_i \rightsquigarrow \mathcal{N}(m_i^*, \sigma_i^*)\right) \\
 &= \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \mathbb{I}_{\sum_{i=1}^k \frac{(x_i - m_i)^2}{2\sigma_i^2} - \frac{(x_i - m_i^*)^2}{2\sigma_i^{*2}} < \sum_{i=1}^k \ln\left(\frac{\sigma_i^*}{\sigma_i}\right)} f_1^*(x_1) \dots f_k^*(x_k) dx_1 \dots dx_k
 \end{aligned}$$

Sur le transcriptome

- Les données reflètent-elles la présence de l'irradiation?

Sur le transcriptome

- Les données reflètent-elles la présence de l'irradiation? **oui**

Sur le transcriptome

- Les données reflètent-elles la présence de l'irradiation? **oui**
- **Entre 500 et 1000 gènes** sont impliqués dans la réponse à l'irradiation.

Sur le transcriptome

- Les données reflètent-elles la présence de l'irradiation? **oui**
- **Entre 500 et 1000 gènes** sont impliqués dans la réponse à l'irradiation.
- Il est cependant **possible de déterminer si une levure a été irradiée ou non** en ne regardant qu'un petit nombre de gènes (cf. expériences de test).

Sur le transcriptome

- Les données reflètent-elles la présence de l'irradiation? **oui**
- **Entre 500 et 1000 gènes** sont impliqués dans la réponse à l'irradiation.
- Il est cependant **possible de déterminer si une levure a été irradiée ou non** en ne regardant qu'un petit nombre de gènes (cf. expériences de test).
- Les prédictions sont **fiables** malgré la présence de bruit dans les mesures et la faible pertinence des gènes. Cependant la présence de **mesures aberrantes reste un problème** pour la prédiction.

Et plus généralement ...

- **Test d'hypothèses contre le hasard** pour évaluer l'importance d'une régularité observée dans les données ;
- Mise en évidence du **rôle des attributs faiblement discriminants** en apprentissage. En effet, s'il sont nombreux :
 - Certains d'entre eux peuvent apparaître comme fortement discriminants ;
 - L'utilisation de plusieurs de ces attributs permet d'obtenir une bonne fiabilité.

- Continuer **faire des expérimentations** pour affiner le modèle et détecter le seuil minimal d'irradiation ;
- Chercher des **différences intra-classes** pour mettre en évidence des types de comportements différents ;
- Essayer de **détecter les valeurs aberrantes** en pré-traitement.
- établir au travers de plusieurs expériences des **degrés de fiabilité** de la mesure chaque gène ;
- Développer notre connaissance du **réseau transcriptionnel** de la levure au travers de plusieurs expériences, puis l'utiliser pour modéliser les dépendances entre les gènes.
- Continuer avec **l'homme et le cancer** : environ 25000 gènes et des interactions complexes entre les gènes.

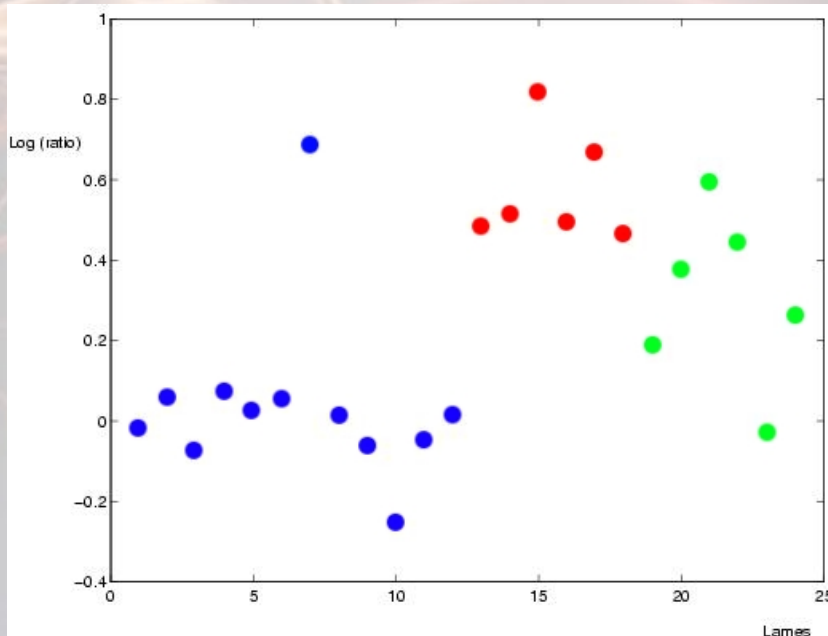
Les cultures à identifier

16

| Lame | Traitement | Dose |
|-------------|-------------------|--------------|
| 1 | Irradiation (I) | 0.03 mGy/h |
| 2 | Irradiation (I) | 0.0075 mGy/h |
| 3 | Irradiation (I) | 0.1 mGy/h |
| 4 | Irradiation (I) | 1.1 mGy/h |
| 5 | Formaldehyde (F) | 0.07 nM |
| 6 | Aucun (S) | 0 |

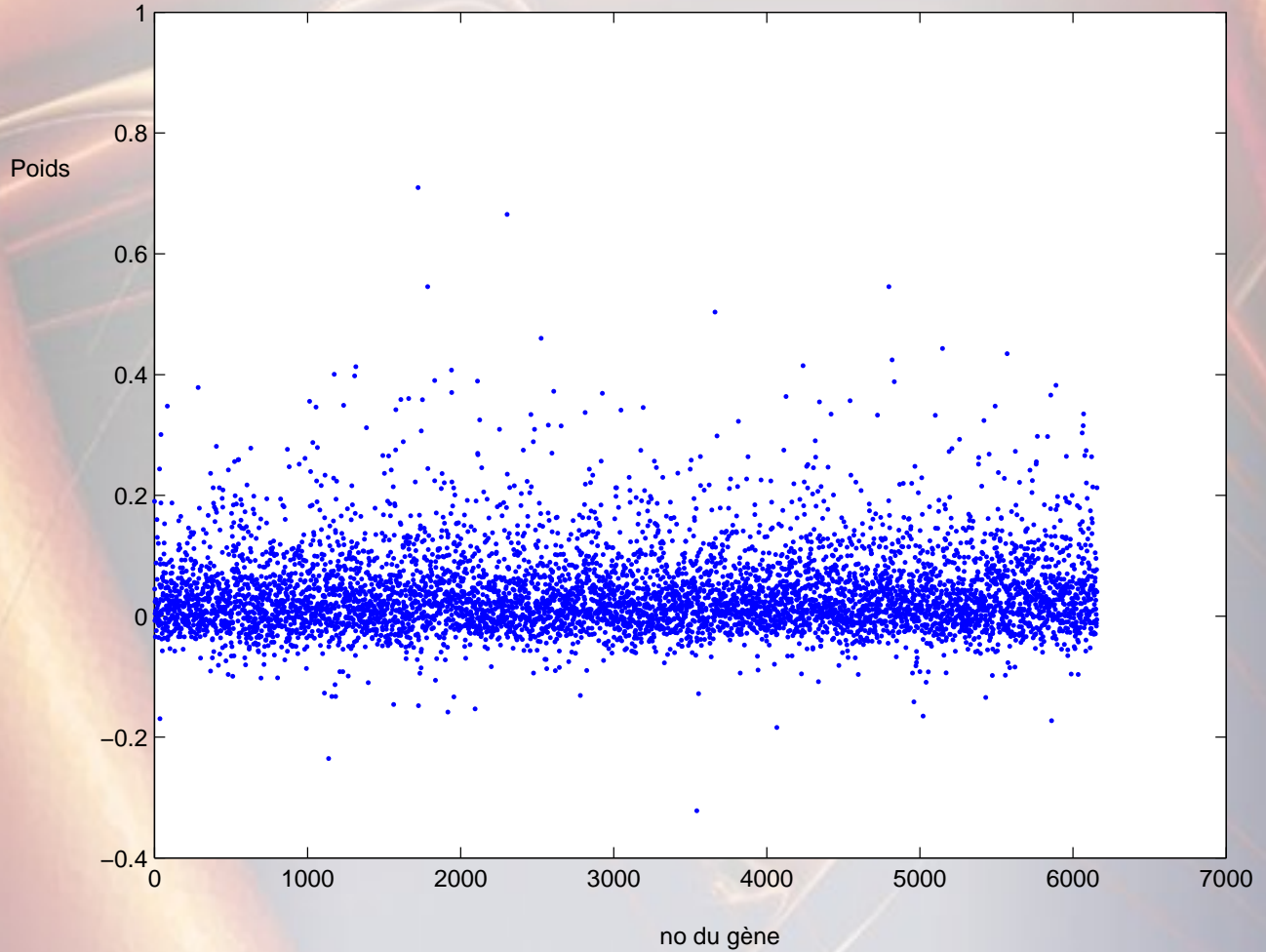
Le gène 1575 : le champion

17



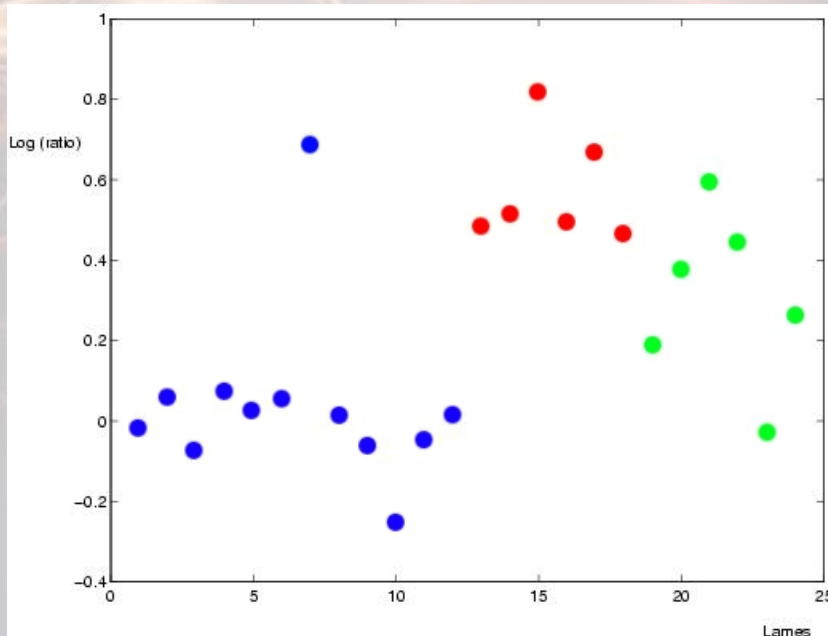
Log(ratio) de l'expression du gène 1575. Les points bleus représentent les valeurs de l'expression du gène pour les levures saines. Les rouges représentent les valeurs pour les irradiés et les verts celles dans l'échantillon à classer.

Poids donnés par RELIEF avec $k = 3$



Le gène 1575 : le champion

19

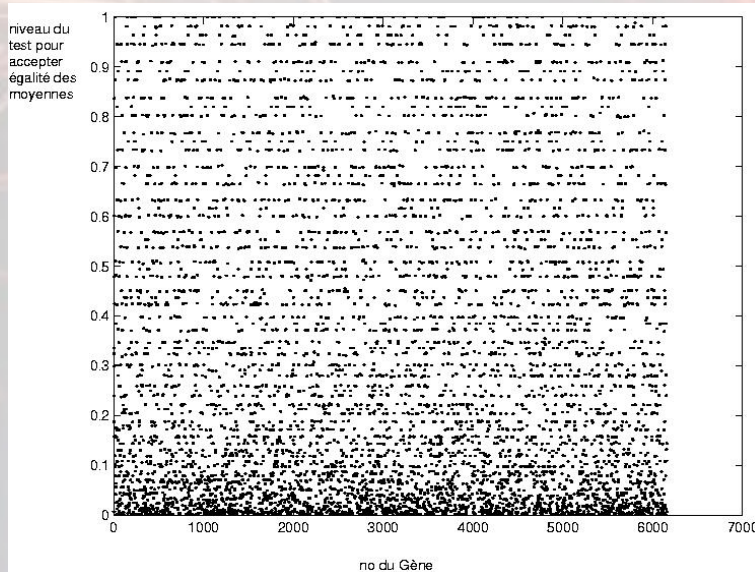


Log(ratio) de l'expression du gène 1575. Les points bleus représentent les valeurs de l'expression du gène pour les levures saines. Les rouges représentent les valeurs pour les irradiés et les verts celles dans l'échantillon à classer.

Comparaison au hasard : le test de Wilcoxon

20

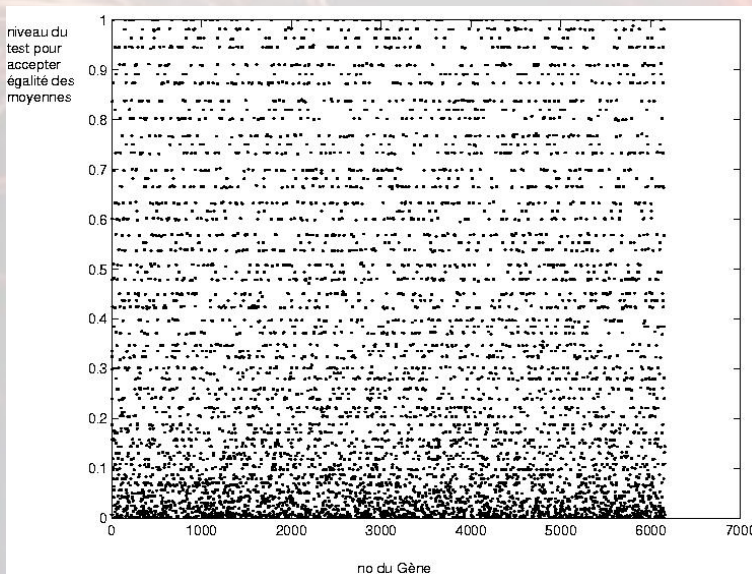
Il s'agit d'un test non paramétrique, indépendant de la densité qui teste l'égalité des lois de deux échantillons.



Comparaison au hasard : le test de Wilcoxon

20

Il s'agit d'un test non paramétrique, indépendant de la densité qui teste l'égalité des lois de deux échantillons.



Au niveau 1% on a 588 gènes.

le nombre de gènes ayant pour le test de wilcoxon un tel niveau est majoré une loi binômiale $\mathcal{B}(\frac{1}{100}, 6157)$. C'est-à-dire que en moyenne, s'il n'y a pas d'influence de la classe ; on observe moins de 61.57 gènes avec cette répartition, ($\sigma = 60.9543$).

Avec le gène 1575

| Lame | Probabilité d'être | |
|-------------|--------------------|----------------|
| | Sain | Irradié |
| 1 | 0.95 | 0.04 |
| 2 | 0.35 | 0.65 |
| 3 | 0.02 | 0.97 |
| 4 | 0.15 | 0.84 |
| 5 | 1 | 0 |
| 6 | 0.82 | 0.17 |

Vote pour la séparation de Gaussiennes (SAM) 22

Prédiction avec 38 gènes (seuil 0.2)

En utilisant la densité produit (et en ne n'ayant jamais fourni de valeur non fiable).

| | Probabilité d'être | |
|------|--------------------|---------|
| Lame | Sain | Irradié |
| 1 | 0 | 1 |
| 2 | 0 | 1 |
| 3 | 0 | 1 |
| 4 | 0 | 1 |
| 5 | 1 | 0 |
| 6 | 0 | 1 |

Vote individuel de chacun des gènes.

| | Probabilité d'être | |
|------|--------------------|---------|
| Lame | Sain | Irradié |
| 1 | 0.53 | 0.47 |
| 2 | 0.46 | 0.54 |
| 3 | 0.5 | 0.5 |
| 4 | 0.47 | 0.53 |
| 5 | 0.65 | 0.35 |
| 6 | 0.55 | 0.44 |

Vote avec les meilleurs gènes de RELIEF

23

En utilisant le produit des densités

| | Probabilité d'être | |
|------|--------------------|---------|
| Lame | Sain | Irradié |
| 1 | 1 | 0 |
| 2 | 0.01 | 0.99 |
| 3 | 0 | 1 |
| 4 | 0 | 1 |
| 5 | 1 | 0 |
| 6 | 1 | 0 |

Vote individuel de chacun des gènes.

| | Probabilité d'être | |
|------|--------------------|---------|
| Lame | Sain | Irradié |
| 1 | 0.70 | 0.30 |
| 2 | 0.54 | 0.45 |
| 3 | 0.36 | 0.63 |
| 4 | 0.47 | 0.52 |
| 5 | 0.78 | 0.22 |
| 6 | 0.72 | 0.28 |