Chapitre 1

Utilisation d'une méthode d'estimation d'attributs pour l'analyse du transcriptome de cellules de levures exposées à de faibles doses de radiation

1.1. Introduction

Tout environnement est influencé par les activités humaines et industrielles qui rejettent dans l'atmosphère différentes sortes de composés chimiques. Par nature ces agents, qui peuvent être hautement nocifs pour la santé publique, sont présents à des doses très faibles et leur détection est rendue malaisée par la présence de mélanges de nombreux composants. Nous sommes partis du postulat que l'exposition à un agent toxique induit un changement des activités de la cellule contribuant à l'élimination de l'agent et la réparation des dommages créés. Ces changements peuvent être estimés par la comparaison des niveaux des ARN messagers codant pour ces activités dans deux populations croissant dans des environnements différents. Nous avons utilisé la technique des biopuces à ADN pour mesurer l'ensemble des ARNm de la levure *Saccharomyces cerevisiae* dans différentes

conditions de croissance. Le but ultime de notre étude est de pouvoir identifier à la fois une réponse commune à tous les agents (permettant ainsi de mettre en place un test général de pollution capable de détecter tout type d'agent toxique) et des réponses spécifiques pour chaque famille d'agents polluants. Dans le cadre de cet article, nous nous limitons à la détermination des effets biologiques de l'exposition à de faibles doses de radiation ionisantes, ce qui est en soi un véritable défi de santé public.

La méthode d'analyse que nous proposons a permis de mettre clairement en évidence qu'une telle réponse transcriptionnelle existe pour de faibles doses d'irradiation. Nous avons pu dégager un ordre de pertinence sur l'ensemble des gènes permettant de différencier les deux conditions (irradié *versus* non irradié) sur les populations observées. Basé sur cet ordre, nous avons isolé un petit ensemble de gènes qui ont permis d'identifier certaines fonctions particulièrement impliquées dans la réponse transcriptionnelle aux faibles doses d'irradiation.

1.2. Données : traitement et choix des doses

Afin d'analyser l'effet de faibles doses de radiations, nous avons utilisé la mesure de l'expression du génome de levures : Non-Irradiées (NI) et Irradiées (I). Pour cela, nous avons fait croître les cellules pendant vingt heures (soit douze cycles cellulaires) en présence de rayonnements ionisants β (1.71 Mev), les cellules étant maintenues en croissance en phase exponentielle durant cette période. Les cultures ont été faites en milieu riche et la distribution de la population dans les différentes phases du cycle a pu être suivie par la morphologie de la cellule : classiquement, on observe des proportions similaires de cellules singlet (G1), de cellules bourgeonnantes (S) et de cellules doublet ou à gros bourgeon (G2/M). D'une façon arbitraire, nous avons choisi de définir la "faible" dose comme la dose d'agent qui n'induit aucun changement cellulaire ou génétique détectable. Ainsi, pour chaque traitement, la croissance de la cellule a été suivie, la morphologie des cellules a été étudiée et les fréquences de mutants et de recombinants en fin de culture ont été mesurées. Les doses auxquelles ont été exposées les cellules de *S. cerevisiae* dans ces expériences sont comprises entre 10 et 30 mGy/h. À de telles doses, aucun changement biologique (retard de croissance), ni génétique (mutagenèse ou recombinaison) n'est observé. Cependant de nombreux gènes montrent un changement transcriptionnel significatif. L'expression de ces gènes a été obtenue à l'aide de puces à lames de verre, produites par Corning, où l'hybridation a été faite avec double marquage fluorescent (Cy3 pour les cADN contrôles et Cy5 pour les cADN étudiés). Pour l'ensemble de l'expérience, les cADN contrôles utilisés proviennent d'un mélange d'ARN extraits de plusieurs cultures indépendantes non traitées. Le même mélange d'ARN sera utilisé tout au long de cette étude pour

générer le cADN contrôle. Des puces développées par la société *Corning*[1] portant la majorité des gènes de la levure *Saccharomyces cerevisiae* ont été utilisées.

Les données ont été générées par analyse après lecture avec un scanner GenePix 4000 (Axon instruments) à l'aide du programme GenePix Pro. Pour les deux fluorescences, les valeurs médianes des pixels à l'intérieur du spot ont été utilisées. Chaque valeur a été soumise à un calcul de contrôle de qualité standard (QCS) basé sur l'estimation de la différence entre la médiane des valeurs à l'intérieur du spot et dans le bruit de fond, corrigée par la somme des écarts types. Les données ayant un QCS faible ont été considérées comme manquantes dans les analyses suivantes. Nous avons travaillé avec des données qui ont été normalisées en utilisant la méthode du LOWESS (LOcally WEighted Scatterplot Smoothing) [YAN 02]. Cette technique corrige les biais introduits par les différences d'intensité et de localisation sur la biopuce, en utilisant une loi de régression locale robuste. Nous avons utilisé la fonction lowess Splus (Insightful) sur chaque bloc d'impression après remise à l'échelle. En effet, pour certaines biopuces, les ratios des deux fluorescences paraissent sous-estimés, pour les intensités très faibles ou très élevées, si on applique une correction linéaire. De plus, la correction par bloc permet de prendre en compte les variations de l'efficacité d'hybridation en fonction de la position sur la biopuce (effet de bord) et de la qualité de l'impression (effet d'aiguilles).

1.3. Problématique bioinformatique et spécificité des données

De façon générale, notre démarche a été, dans un premier temps, de chercher à analyser le transcriptome de ces levures afin d'identifier les gènes concernés par la réponse à une faible irradiation, puis, dans un second temps, de chercher à quels groupes fonctionnels participent les gènes ainsi mis en évidence. Concernant la première étape, plusieurs éléments nous ont intéressés plus particulièrement :

- le nombre de gènes impliqués dans la réponse transcriptionnelle ;
- l'identité des gènes (pour savoir s'il s'agit des mêmes que lors d'une irradiation plus forte) ;
- la capacité de prédiction de la classe (I ou NI) d'une nouvelle levure en prenant seulement en compte son transcriptome.

Il s'agit là d'un problème classique, dit "d'apprentissage supervisé", puisqu'on dispose d'instances d'entraînement de deux classes (levures Irradiées et levures

[1] Les puces Corning se composent de 12 blocs de 24 colonnes et de 24 lignes. Les 6912 spots déposés se répartissent de la façon suivante : 1) 6157 ORF dont 22 répétées deux fois ; 2) 108 spots contrôle (9 par bloc) ; 3) 432 spots vides ; 4) 215 spots FSV (Failed Sequence Verification).

Non-Irradiées) et que l'on souhaite être capable de les distinguer à partir des valeurs d'expression d'un sous-ensemble pertinent de gènes à déterminer.

Cette tâche est cependant rendue difficile pour plusieurs raisons :

– Présence de bruit dans les données correspondant à deux causes :
  - Imprécision de la mesure : il s'agit d'un bruit classique supposé gaussien, bruit qui est très élevé pour certains gènes (ainsi pour certains gènes dont l'expression est mesurée deux fois sur une même puce, les résultats sont parfois très différents);
  - Présence de valeurs aberrantes dues à un problème lors de l'hybridation.
– Nombreux attributs : 6157 gènes
– Très faible nombre d'instances : 12 cultures non-traitées, 6 irradiées
– Les classes sont déséquilibrées (elle ne contiennent pas le même nombre d'éléments)
– Absence d'indépendance conditionnelle probabiliste entre les gènes, puisque l'expression des gènes est corrélée.

Dans un apprentissage idéal, la mesure des niveaux d'expression des gènes devrait suffire pour identifier exactement tous les gènes impliqués dans la réponse à une faible irradiation. Malheureusement, les méthodes d'apprentissage ont des difficultés à chercher un tel ensemble de gènes, eu égard au grand nombre d'attributs (les niveaux d'expression des gènes) et au faible nombre d'instances pour apprendre (les lames étudiées). Nous avons donc dû choisir entre détecter presque tous les gènes impliqués dans la réponse à l'irradiation (on a alors peu de faux-négatifs), quitte à déclarer importants des gènes non-impliqués (nombreux faux-positifs) ou, inversement, ne donner que des gènes impliqués de manière quasi-certaine, avec cette fois le risque d'oublier beaucoup de gènes impliqués. En raison de la difficulté d'interpréter biologiquement les gènes retenus, s'ils sont en trop grand nombre, cette deuxième voie a été choisie. Nous avons pu montrer qu'il est possible de détecter un petit nombre de gènes qui sont impliqués de manière quasi-certaine dans l'irradiation.

1.4 Estimation d'attributs par la méthode BioRelief

Compte tenu des caractéristiques des données soulignées précédemment (grand nombre d'attributs et faible nombre d'instances), il nous a semblé nécessaire de recourir à des techniques qui examinent directement la corrélation de chaque gène avec la classe de l'instance (I ou NI). Il s'agit là d'un problème d'estimation d'attributs pour lequel plusieurs techniques ont été proposées dans les dernières années ([NG 01, TUS 01, XIN 01]). Pour des raisons que nous détaillerons plus loin, nous avons choisi, pour mesurer la pertinence des gènes, une méthode basée sur

l'algorithme RELIEF [KON 94], qui est une technique d'estimation d'attributs qui cherche à détecter les attributs les plus significativement corrélés à la classe à prédire. Son principe consiste à calculer un *poids* compris entre –1 et +1, pour chaque gène, les poids positifs indiquant une corrélation positive entre l'expression du gène sur la lame et la classe de celle-ci. Le poids d'un gène est fonction de la variation de son niveau d'expression au sein d'une même classe, comparée à la variation de ce même niveau entre les deux classes. En effet, l'expression d'un gène semble d'autant plus corrélée avec la classe de l'instance que la variation intra-classe de l'expression du gène est petite comparée à sa variation entre classes.

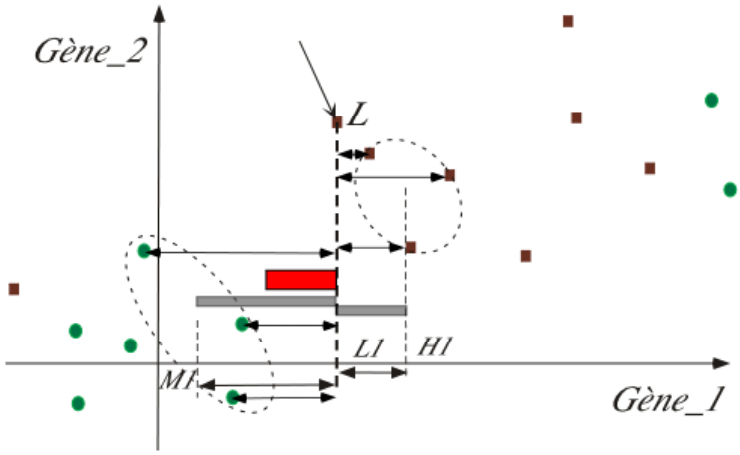

Figure 1.1. *Illustration du fonctionnement de* RELIEF *dans le cas simple de deux gènes seulement, avec un voisinage de* k*=3 voisins. Les lames sont alors représentées par des points dans le plan des niveaux d'expression des deux gènes. Pour évaluer la pertinence du gène_1 par exemple, on prend successivement tous les points en compte. Ainsi pour chaque lame* L*, on détermine ses 3 plus proches voisins de la même classe (carrés), ainsi que ses 3 plus proches voisins de l'autre classe (ronds). Pour chacun des points, on compte comme contribution positive la longueur de la projection sur gène_1 du segment allant de L à chacun des points de la classe opposée (3 contributions positives ici, chacune des contributions sera donc divisée par trois). On compte ensuite comme contribution négative la longueur de la projection du segment allant de L à chacun des points de la même classe (3 contributions négatives, chacune des contributions sera divisée par 3). Le gros segment mesure la contribution résultante de la lame* L *à la pertinence du gène_1 pour différencier les deux classes.*

La procédure de calcul du poids associé à un gène (attribut) est la suivante. Une lame *L* peut être considérée comme un point dans un espace à 6157 dimensions tel que la *i*ème coordonnée correspond au niveau d'expression du *i*ème gène sur la lame. Pour chacune des *m* lames, de classe I ou NI, on calcule ses *k* plus proches

voisins dans la même classe (notés $H_1,\ldots,H_k$). On calcule ensuite ses k plus proches voisins dans l'autre classe (notés $M_1,\ldots,M_k$). On considère alors les projections des points L, H_q et M_q ($1 \leq q \leq k$) selon le gène dont on calcule le poids. Les projections de ces points selon le gène g correspondent aux niveaux d'expression de g pour les lames représentées par ces points. On détermine ensuite, d'une part, la distance entre la projection de L et la projection de chacun des M_q, et, d'autre part, la distance entre la projection de L et la projection de chacun des H_q (voir figure 1.1). La différence entre la somme des valeurs obtenues pour les M_q (divisée par le nombre de M_q utilisés) et la somme des valeurs obtenues pour les H_q (divisée par le nombre de H_q utilisés) fournit la contribution de cette lame (point) au calcul du poids du gène considéré. On répète ce calcul pour toutes les lames et on somme les contributions obtenues en normalisant par le nombre de lames.

Pour k=1, ce calcul correspond à la formule suivante :

$$\text{poids}(gène) = \frac{1}{m}\sum_{L=1}^{m}\{|\text{expr}_{gène}(L)-\text{expr}_{gène}(M_1)| - |\text{expr}_{gène}(L)-\text{expr}_{gène}(H_1)|\} \quad [1.1]$$

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

Le poids calculé pour chaque *gène* est ainsi une approximation (voir [KON 94]) 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)

Dans cette étude, nous avons choisi d'utiliser la distance de Manhattan dans l'espace des lames plutôt que la distance euclidienne classique car cette dernière tend à surpondérer les gènes (attributs) pour lesquels les expressions mesurées sont très bruitées ou présentent des valeurs aberrantes.

Il existe déjà des variantes de RELIEF qui permettent de traiter plus de deux classes. Il en va de même pour les valeurs manquantes. Par ailleurs, le paramètre k (nombre de voisins utilisés) permet de contrôler le compromis entre sensibilité et robustesse au bruit. Une version initiale présentée dans [MAR 03] ne nous permettait pas de choisir un paramètre k différent selon la classe. Comme cette fonctionnalité peut se révéler utile, si l'une des classes comporte beaucoup plus d'instances que les autres, nous avons choisi d'intégrer la possibilité de choisir une valeur de k de manière indépendante pour chacune des classes[2]. Nous proposons donc une autre variante, BioRelief.

[2] Nous remercions J-L. Giavitto pour cette suggestion.

Dans la suite,

- on dispose d'un tableau *puce* de taille *m* x *n*, où *n* est le nombre de gènes et *m* le nombre de lames, qui donne le niveau d'expression de chaque gène sur chaque lame ;

- pour chaque lame, on connaît sa classe (référencée par un entier pour plus de simplicité) ;

- on se donne un tableau d'entiers *K* dont la taille est *nbc*, le nombre de classes, et qui indique le paramètre *k* à utiliser pour chacune des classes. Ainsi la *c*-ème composante de ce tableau *K* contient la valeur de *k* pour la *c*-ème classe.

On commence par calculer une quantité *P(c1,c2)*, *c1≠c2*, qui va servir à pondérer les contributions des classes, lors de l'estimation de la variation entre classes, dans le cas où l'on a plus de deux classes. Représente la probabilité que la lame soit de classe *c2* sachant qu'elle n'est pas de classe *c1*. Remarquons que si le nombre de classes est égal à 2, *P(c1,c2)* =1.

P(c1, c2) = (nb de lames de classe *c2*) / (nb de lames qui ne sont pas de classe *c1*)

On calcule ensuite pour chacun des exemples et des classes quels sont ses plus proches voisins. Ainsi *voisin*(*l*, *c*, *k*) donne la lame de classe *c* qui est la *k*ème plus proche (en ne considérant que les lames de classe *c*) de la lame *l*. Pour cela, nous avons besoin de calculer la distance entre deux lames. On note *kmax* la valeur maximale que peut prendre ce rang *k*, pour toutes les classes. Comme nous l'avons précisé, nous utilisons une variante de la distance de Manhattan modifiée pour traiter les valeurs manquantes. En pratique, pour calculer la distance entre deux lames, nous commençons par supprimer tous les gènes (attributs) pour lesquels une des deux valeurs manque. Ensuite, nous calculons la distance classiquement et enfin nous la divisons par le nombre de gènes utilisés (cette dernière étape sert à autoriser la comparaison entre deux distances qui n'ont pas forcément utilisé le même nombre de gènes).

Il ne reste plus maintenant qu'à obtenir les poids. Il faut cependant adapter la procédure décrite pour tenir compte des données manquantes. Une première méthode consiste à remplacer les valeurs manquantes par 0. Ce choix est dû à la forme des données : il s'agit d'un log ratio, donc biologiquement, si la valeur est à 0, c'est que le gène s'est exprimé de la même manière que dans l'échantillon de référence. Une deuxième méthode consiste à estimer la valeur manquante à partir des données. Le résultat obtenu avec cette seconde méthode étant très similaire à un remplacement par la valeur 0, nous avons retenu la première méthode.

La fonction de calcul de poids BioRelief décrite dans l'algorithme ci-dessous est donnée dans le cas général où il y a plus de deux classes, ce qui est intéressant pour l'étude de plusieurs polluants.

Fonction BioRelief(*i* : entier; *puce* : tableau [1..m, 1..n] de réels; *P* : tableau [1..nbc, 1..nbc] de réels; *K* : tableau [1..nbc] d'entiers; *voisin* : tableau [1..m, 1.. nbc, 1..kmax] d'entiers) :
retourne réel
// La fonction Relief calcule le poids du gène numéro *i* en utilisant *puce*, *P*, *K* et *voisin*, en //supposant connues les classes pour les lames

Début
//Initialisation
Poids = 0 ;
Pour *lame* =1 **jusqu'à** m **faire**

//Traitement de la variation intraclasses
Soit *C* le numéro de la classe de *lame*
//*K(C)* est le nombre de voisins à utiliser dans la classe numérotée *C*
Pour *no_voisin* =1 **jusqu'à** *K(C)* **faire**
Poids = *Poids* - abs(*Puce*(*lame*, *i*) - *Puce*(*voisin*(*lame*, *C*, *no_voisin*), *i*)) / *K*(*C*)
FinPour

//Traitement de la variation entre classes
Pour *DC* parcourant toutes les classes autres que *C* **faire**
Pour *no_voisin* =1 **jusqu'à** *K*(*DC*) **faire**
Poids=*Poids*+*P*(*C*,*DC*)*abs(*Puce*(*lame*,*i*)-*Puce*(*voisin*(*lame*,*DC*, *no_voisin*),*i*))/*K*(*DC*)
FinPour
FinPour

FinPour

Retourne (*Poids / m***)**
Fin BioRelief

La complexité de l'algorithme est en O(m . k_{max} . nbc), m étant le nombre de boucles (nombre de lames), k_{max}, le nombre maximal de voisins considérés, nbc le nombre de classes (ici deux). Le calcul des voisins est lui-même en O(n . m^2). Dans BioRelief, nous précalculons les voisins de chaque point afin d'accélérer les calculs.

Dans le cas de notre étude, il n'y a que deux classes, et il faut alors choisir les « meilleures » valeurs des paramètres k_1 (classe NI) et k_2 (classe I) en fonction des données disponibles. Augmenter la valeur de k permet d'augmenter la résistance au bruit puisque un point aberrant sélectionné comme plus proche voisin aura une influence tempérée par d'autres points. En revanche, une faible valeur de k, en prenant moins en compte la moyenne des points d'une classe, conduit à une plus grande sensibilité aux particularités de la classe considérée. Nous pensons qu'une bonne solution à ce compromis consiste à se comparer au hasard et à essayer d'en être le plus éloigné possible (la figure 1.3 représente une courbe de répartition des poids pour k_1=3, k_2=3, en rouge et en traits pleins et la courbe du « hasard » est bleu et en pointillés). Pour mesurer la distance au hasard, il nous est apparu comme pertinent de considérer l'aire comprise entre les deux courbes. Nous avons donc représenté sur la figure 1.2, pour différentes valeurs de k, l'intégrale de la différence entre la courbe avec les vraies classes et la courbe aux classes permutées. La courbe suggère de prendre k_1=2 (chez les NI) et k_2=5 (chez les I). Les résultats rapportés dans la suite s'appuient cependant sur les valeurs k_1=3, k_2=3 peu éloignées de l'optimum. Ces résultats sont en effet ceux qui ont été utilisés pour l'interprétation biologique [MER 04].

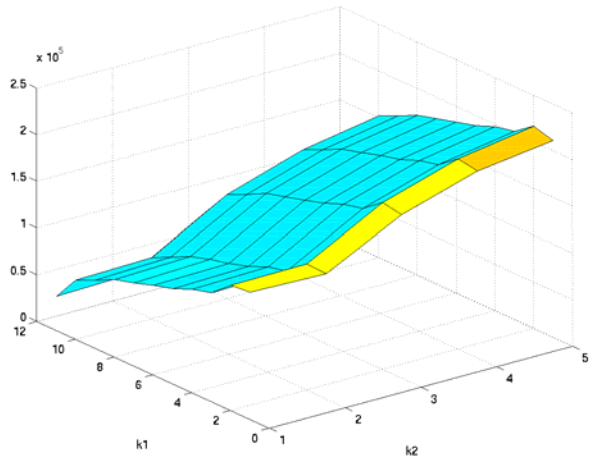

Figure 1.2. *Surface représentant la « différence par rapport au hasard» selon les valeurs de k choisies pour chacune des deux classes (toutes les valeurs possibles pour notre jeu de données (6 lames de la classe I et 12 de la classe NI) sont représentées). La meilleure valeur est obtenue pour (k_1=2, k_2=5). L'axe vertical représente l'aire entre la courbe avec les classes attribuées aléatoirement et la courbe obtenue avec les vraies classes.*

Outre son coût de calcul faible (en optimisant le code), l'un des avantages de BioRelief est que, à la différence de nombreuses autres techniques statistiques, il ne fait aucune hypothèse tant sur l'indépendance entre les expressions des gènes que sur la distribution de leurs valeurs (en particulier on ne suppose pas que la répartition des valeurs d'expression d'un gène suit une gaussienne).

1.5 Détermination de gènes pertinents

Il est difficile de décider combien de gènes sont pertinents pour la détermination de la condition irradiée ou non, en se basant simplement sur les poids trouvés, puisqu'il n'y a pas de seuil évident. Comment alors déterminer à partir de quelle valeur on peut considérer qu'un gène est impliqué dans la réponse à l'irradiation ? Qui plus est, le faible nombre d'instances et le bruit inhérent aux données rendent la tâche encore plus complexe, si bien que par pur hasard, certains gènes pourraient apparaître comme fortement corrélés, alors qu'il n'en est rien. La comparaison avec l'hypothèse nulle selon laquelle il n'y aurait aucune corrélation entre le niveau d'expression d'un gène et la classe (c'est-à-dire si l'irradiation n'entraînait pas de variation de l'expression du génome) permet de mesurer l'information apportée par la connaissance de l'expression des gènes sur la classe des lames.

Pour cela, nous avons constitué de nouveaux jeux de données dans lesquels les classes des lames ont été aléatoirement attribuées (comme dans [TUS 01]), tout en conservant les proportions 6 I et 12 NI - car BioRelief est sensible aux proportions - si bien que le groupe des 6 cultures portant le label (I) n'est pas forcément constitué de 6 cultures irradiées. Nous avons alors réutilisé BioRelief pour obtenir de nouvelles mesures des poids pour chacun des gènes. Ces opérations (mélange des classes puis calcul du nouveau poids pour chaque gène) ont été répétées jusqu'à ce que la courbe moyenne du nombre de gènes ayant un poids supérieur à un certain seuil ne varie plus. On a ainsi itéré le processus 2000 fois, en observant un début de stabilisation vers la 500ème itération. S'il existe une réelle corrélation entre les niveaux d'expression des gènes et l'appartenance à une classe, on devrait observer pour un poids donné (seuil), que plus de gènes apparaissent corrélés avec les vraies données que sous la condition d'hypothèse nulle.

Les résultats obtenus (voir Figure 1.3) montrent qu'effectivement pour tout niveau de corrélation donné (poids calculé par BioRelief), le nombre de gènes franchissant ce seuil est nettement plus élevé dans les données réelles que sous l'hypothèse nulle. Cette observation à elle seule valide l'hypothèse d'une réponse transcriptionnelle aux faibles doses d'irradiation lorsque aucun changement cellulaire ou génétique n'est par ailleurs détectable. Plus précisément :

- En moyenne, aucun gène ne peut atteindre un poids supérieur ou égal à 0.58 par pure chance (la courbe inférieure passant sous le seuil de 1

gène). À ce seuil, on observe cependant 13 gènes corrélés dans les données expérimentales (courbe supérieure). On peut donc estimer qu'il est très probable que ces gènes soient effectivement impliqués dans une réponse transcriptionnelle aux faibles doses d'irradiation

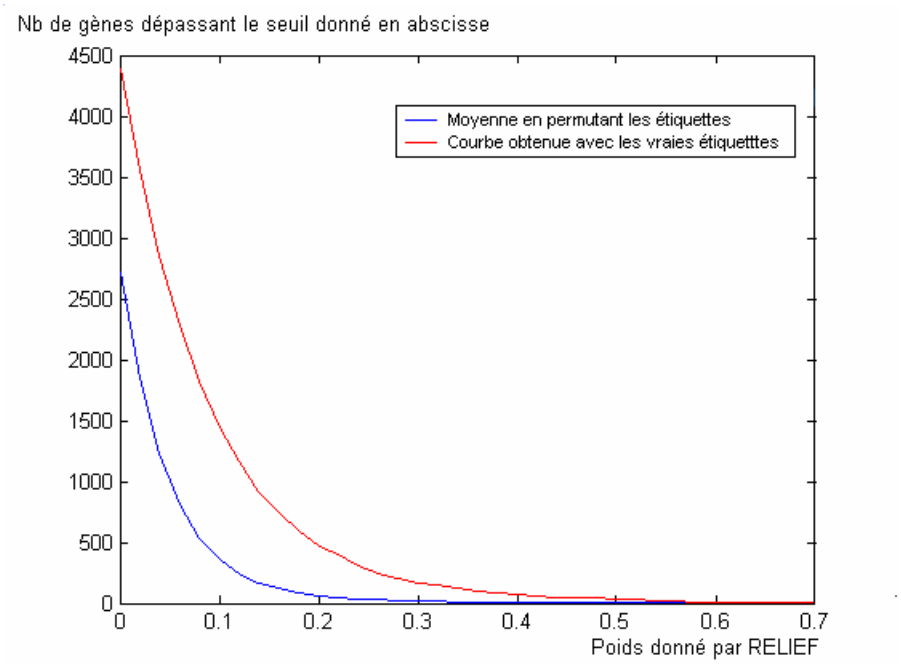

Figure 1.3. *Courbe du nombre de gènes dépassant le seuil de corrélation indiqué en abscisse, avec les vraies classes et courbe moyenne obtenue aléatoirement en permutant les classes, en respectant les proportions des classes initiales. Au seuil de 0.58, la courbe inférieure (hypothèse nulle) passe en dessous de 1, alors qu'elle vaut 13 pour la courbe correspondant aux conditions expérimentales. Pour un seuil de 0.48, il y a 35 gènes correspondant à la courbe expérimentale.*

Il faut noter que d'autres gènes sont très certainement également impliqués. Pour obtenir un ensemble de gènes plus large à des fins d'interprétation biologique, nous sommes prêts à accepter que 10% des gènes retenues soient de faux positifs. Ce pourcentage est obtenu pour un seuil de 0.3 : si nous considérons les gènes avec un poids supérieur à 0.3, 171 gènes sont sélectionnés dans le cas des données expérimentales pour 17 seulement dans la cas aléatoire (cf. figure 1.3). Dans la suite, 171 gènes ayant un poids supérieur à 0.3 sont considérés comme informatifs.

Une question pratique essentielle est alors d'examiner si les gènes classés comme les plus informatifs peuvent suffire à établir un diagnostic d'exposition à de faibles radiations. Nous avons fait des tests à partir d'une méthode de classification

apprise en utilisant un faible nombre de gènes. Les résultats obtenus sont excellents mais se rapportent à un jeu d'essai trop limité. (Le lecteur intéressé pourra se reporter à [MAR 03]).

1.6 Interprétation biologique des résultats

La première étape de notre étude a donc montré que la mesure de l'expression d'une sous-population de gènes est effectivement indicative d'un effet des faibles radiations chez la levure.

La deuxième étape de l'étude consiste à analyser la fonction biologique des gènes qui apparaissent ainsi les plus impliqués dans la réponse transcriptionnelle à ces expositions.

Nous avons étudié les 171 premiers gènes classés par BioRelief, ceux dont le poids est supérieur à 0.3. Ces gènes ont d'abord été regroupés selon les voies métaboliques dans lesquelles ils sont impliqués. La fréquence de ces voies dans la liste des gènes induits ou réprimés par l'exposition aux faibles doses de radiations a été comparée à celle de ces voies métaboliques sur l'ensemble de la lame. Plus précisément, la deuxième colonne de la partie supérieure (respectivement inférieure) de la Table 1.1 indique le nombre d'ORFs induites (resp. réprimées) parmi les 171 sélectionnées, qui participent à l'activité mentionnée en première colonne (e.g. 9 gènes sur les 171 sont induits et correspondent à la fonction d'oxidative phosphorylation). La troisième colonne donne le rapport entre le nombre d'ORFs induites (resp. réprimées) sélectionnées participant à cette activité et le nombre d'ORFs induites (resp. réprimées) sélectionnées, soit 91 (resp. 80) (e.g. 9/91 = 9.9% des 91 gènes induits correspondent à l'oxidative phosphorylation) ; la quatrième colonne indique le pourcentage des ORFs ayant cette activité (on considère alors l'ensemble de tous les gènes de la levure) (e.g. 0.3% de tous les gènes induits correspondent à l'oxidative phosphorylation). La dernière colonne donne alors le taux de sur-représentativité de l'activité considérée dans la population des 171 meilleurs gènes sélectionnés (e.g. la fonction d'oxidative phosphorylation est surreprésentée dans les 91 gènes induits sélectionnés par rapport à sa représentation moyenne dans les gènes induits parmi 6157 : soit 9.9% / 0.3% = 30.5).

Sur la Table 1.1, on voit ainsi que 20 gènes participant au stress oxydatif, à la phosphorylation oxydative et à la synthèse d'ATP sont induits, ce qui représente un excédent de 30 fois par rapport à la population totale.

Sans entrer dans le détail de l'interprétation biologique, en cours, de ces groupes fonctionnels, il est cependant possible d'en conclure que la méthode de sélection de gènes utilisée dans cette étude est validée par le fait que les fonctions mises en

évidence par l'analyse des gènes sélectionnés sont connues pour intervenir dans l'élimination de certains des produits cellulaires des rayonnements ionisants (radicaux libres). (Des informations supplémentaires sur ces fonctions sont publiées dans [MER 04]).

function of 91 induced genes/171	number of ORFs	% in this list	% total ORFS	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

Table 1.1. *Les fonctions des 171 gènes retenus sont présentées ici. La première colonne indique le nombre de gènes concernés par chaque fonction. La deuxième colonne donne le pourcentage relatif des gènes de la fonction considérés dans la liste des gènes retenus. La troisième colonne donne le pourcentage relatif des gènes de la fonction considérée dans l'ensemble de tous les gènes. La quatrième colonne traduit alors la sur ou la sous-représentativité d'une fonction dans les gènes retenus par rapport à l'ensemble des gènes. Ainsi, la fonction de phosphorylation oxydative est largement sur-représentée (30,5 fois) dans les 171 gènes retenus par rapport à sa fréquence globale dans tous les gènes.*

Table 1.1.

1.7 Comparaison avec d'autres méthodes d'analyse

L'utilisation de BioRelief pour la sélection de gènes dans l'analyse du transcriptome est originale. Nous avons souligné dans la section 1.4 les caractéristiques qui en font une méthode à considérer pour ce problème : absence d'hypothèse sur la distribution des instances et sur la probabilité *a priori* des classes, ainsi que robustesse au bruit et aux valeurs aberrantes.

Cette méthode s'inscrit dans la lignée des approches non paramétriques telle que celles présentées dans [GRA 00,PAR 01,TRO 02] (voir section 1.8 ci-dessous).

Nous l'avons par ailleurs comparée avec une méthode très courante : la méthode ANOVA [GLA 00] qui est une généralisation au cas multiclasse de la méthode du test de Student, elle-même à la base de la méthode SAM (Significance Analysis of Microarrays [TUS 01]). L'analyse de la variance a déjà été utilisée pour les données d'expression de gènes [KER 00] et a permis de proposer une normalisation des données des puces à ADN tout en permettant une grande partie de l'analyse des données (y compris la recherche des gènes pertinents). Dans l'étude présentée ici, on l'utilise pour trouver les gènes les plus différentiellement exprimés entre les 2 classes (I et NI) après normalisation de type LOWESS (voir section 1.2).

La méthode ANOVA (ANalysis Of VAriance) teste l'égalité de plusieurs moyennes sur une variable (ici l'expression des gènes) en fonction de variables indépendantes (ici la classe des lames). On suppose que les échantillons de données sont tirés aléatoirement et sont indépendants, que les populations sont distribuées suivant une loi normale et que les populations ont même variance (ce qui est une hypothèse forte et non toujours vérifiée dans les biopuces). Le principe de la méthode est d'estimer la variance des données d'une part en tenant compte de leur classe, et d'autre part, sans en tenir compte. En supposant que la classe ne soit pas corrélée à la variable mesurée, ces deux estimations devraient être proches. La méthode ANOVA est appliquée, dans notre cas, à chacun des 6157 gènes et retourne pour chacun d'eux un nombre mesurant la corrélation statistique avec la classe.

Afin de comparer les statistiques de Fisher données par ANOVA à l'hypothèse nulle de non corrélation entre le niveau d'expression d'un gène et la classe, nous avons utilisé la même approche que pour BioRelief: mélange de classes et calcul des nouvelles statistiques de Fisher (voir section 1.5). Les gènes pertinents sont ceux pour lesquels la statistique de Fisher (pour les vraies classes I et NI) est très élevée par rapport à celles obtenues pour des mélanges de classes. Autrement dit, pour chaque gène, on centre et réduit la statistique obtenue pour les vraies classes par

rapport à l'ensemble des statistiques de Fisher obtenues après mélange de classes. Les gènes sont ensuite classés par valeur décroissante de cette statistique centrée réduite.

Les résultats obtenus ont aussi été comparés à ceux obtenus par le logiciel SAM [TUS 01], qui identifie les gènes les plus significatifs en ordonnant les gènes suivant une statistique basée sur les variations de l'expression d'un gène, ramenées à l'écart-type. Les gènes retenus sont alors ceux pour lesquels le rang est éloigné du rang moyen après mélange des classes. Cette méthode fait ressortir une grande partie des gènes obtenus par notre approche de centrage-réduction du test de Student, puisque parmi les 500 gènes obtenus par les 2 méthodes, 82% sont en commun.

Nous avons alors étudié la corrélation entre les résultats obtenus par la méthode BioRelief et ceux obtenus par ANOVA. Considérons, par exemple, les 500 gènes classés comme les plus pertinents par BioRelief d'une part, et les 500 gènes classés comme les plus pertinents par ANOVA d'autre part. Il y en a 257 en commun : c'est beaucoup, mais est-ce significatif ? Pour cela, nous avons regardé quelle était la probabilité que cette intersection soit au moins aussi grande dans le cas de deux tirages aléatoires indépendants de 500 gènes parmi 6157. On obtient la solution par la mise en œuvre de la loi hypergéométrique qui régit la taille de l'intersection à attendre dans le cas aléatoire : $H(n,N-n,n)$ avec n le nombre de gènes déjà choisis par le premier tirage (500), N le nombre total de gènes (6157), et le troisième paramètre étant le nombre de gènes à choisir par le 2ème tirage aléatoire (500). On montre ainsi que la probabilité d'obtenir par hasard une intersection de taille supérieure à 257 dans deux tirages de 500 parmi 6157 est extrêmement faible (de l'ordre de 10^{-169}). Cette probabilité reste du même ordre pour les différentes intersections mesurées (sur un intervalle de tirages de 100 à 2000 gènes, pour lesquels, à chaque fois, la taille de l'intersection est supérieure à la moitié du nombre de gènes retenus par chaque méthode). On peut donc en conclure que les méthodes BioRelief et ANOVA, bien que basées sur des principes différents, en utilisant les mêmes données, produisent en partie la même information. Une comparaison similaire des classifications par BioRelief, ANOVA et SAM donne les résultats suivants : 409 gènes communs dans les 500 premiers classés par ANOVA et SAM, soit une proportion de 82%, ce qui est remarquable car d'après la loi hypergéométrique, une telle intersection a une probabilité *a priori* de 10^{-160}. Parmi les 35 premiers gènes classés par BioRelief, 8 ont été classés parmi les 35 premiers par SAM et 8 aussi parmi les 35 premiers par ANOVA, et ce sont les mêmes (or la probabilité selon la loi hypergéométrique est de 10^{-12}, ce qui est très significatif). Compte tenu de l'absence d'hypothèse d'indépendance entre les expressions de gènes dans BioRelief et de sa bonne résistance au bruit, l'interprétation biologique rapportée plus haut porte sur les gènes obtenus par BioRelief (cf. Table 1.1).

1.8 BIORELIEF et les méthodes non paramétriques

L'archétype de la méthode paramétrique dans la détection des gènes impliqués dans la réponse biologique à une situation donnée est la méthode du *t*-test. Afin de mesurer le degré d'implication de chaque gène, cette méthode compare l'hypothèse H_1 selon laquelle les moyennes d'expression des gènes sont différentes en fonction de la situation dans laquelle se trouve l'organisme étudié, avec l'hypothèse nulle H_0 selon laquelle ces moyennes sont égales. Pour ce faire, le *t*-test suppose que les degrés d'expression des gènes suivent des lois normales, de variance connue ou inconnue, et égales ou non entre les situations. On peut alors ordonner les gènes en fonction du seuil de rejet de l'hypothèse H_0 leur correspondant.

L'hypothèse paramétrique intervient ici à deux niveaux. D'une part, on suppose que les degrés d'expression des gènes suivent une loi normale. D'autre part, on fonde la comparaison entre l'hypothèse H_1 et l'hypothèse nulle H_0 sur la forme paramétrique (loi normale) de cette dernière.

On peut par conséquent s'éloigner du modèle paramétrique selon deux directions. Soit on peut supposer que les expressions des gènes suivent une loi paramétrique (e.g. normale) mais établir une comparaison avec l'hypothèse nulle par une évaluation directe à l'aide d'une méthode de permutation. C'est le cas de la méthode SAM par exemple [TUS 01]. Soit on peut en plus ne pas supposer que les expressions des gènes suivent une loi normale. C'est le cas de BioRelief.

La plupart des méthodes dites non paramétriques dans la littérature en bioinformatique le sont au premier sens défini ci-dessus. C'est le cas par exemple des méthodes bayésiennes empiriques, de SAM et des méthodes de mélanges de modèles (pour une discussion de ces méthodes, voir par exemple [PAN 03], [ZHA 03]).

En s'affranchissant de toute hypothèse sur le type de distribution des données, la méthode BioRelief est insensible à leur adéquation aux données disponibles. Cela est particulièrement intéressant pour l'analyse du transcriptome pour laquelle les hypothèses paramétriques semblent souvent injustifiées et trop fortes.

Aucune méthode ne pouvant cependant être uniformément supérieure aux autres, il reste cependant à déterminer les champs d'application les plus favorables à BioRelief. C'est l'une des questions que nous étudions actuellement.

1.9 Conclusion

Grâce à l'analyse par BioRelief des niveaux d'expression de gènes d'un échantillon de levures pour lesquelles nous connaissions la classe (irradiée ou non), nous avons pu mettre en évidence qu'une réponse transcriptionnelle à de faibles radiations existe. De plus, nous avons pu dégager un ordre de pertinence sur l'ensemble des gènes permettant de différencier les deux conditions (irradié *versus* non irradié) sur les populations observées. Ce point nous permet d'envisager d'utiliser les données issues du transcriptome à des fins de diagnostic. Par ailleurs, un certain nombre de ces gènes caractéristiques s'avère avoir des fonctions qui participent au même réseau fonctionnel, ce qui est un résultat tout à fait prometteur. La transposition à d'autres études (telles que, par exemple, la classification de tumeurs chez l'homme) de la méthode utilisée dans ce travail fait l'objet d'investigations.

Remerciements

Ce travail a été partiellement soutenu par le contrat Biogen n°74 (BioIngénierie 2001) et l'Institut National de Recherche et de Sécurité (convention n°5011888).

1.10. Bibliographie

[COR 02] Cornuéjols A. & Miclet L. : *Apprentissage artificiel. Concepts et algorithmes*. Eyrolles, 2002.

[GLA 00] S.A. Glantz & B.K. Slinker, *Primer of Applied Regression & Analysis of Variance*, McGraw-Hill/Appleton & Lange, 2nd edition, 2000.

[GRA 00] Grant G., Manduchi E. & Stoeckert C., « Using non-parametric methods in the context of multiple testing to determine differentially expressed genes », in *Methods of Microarray Data Analysis: Papers from CAMDA'00*, eds Lin SM. and Johnson KF., Kluwer Academics, pp.37-55, 2000.

[KER 00] Kerr M.K., Martin M. & Churchill G.A., « Analysis of variance for gene expression in microarray data », *Journal of Computational Biology*, 2000, 7(6), 818-837.

[KON 94] Kononenko I., « Estimating Attributes : Analysis and Extensions of RELIEF », Proc. of the *European Conference on Machine Learning*, ECML-94, 171-182, 1994.

[MAR 03] Mary J., Mercier G., Comet J-P., Cornuéjols A., Froidevaux Ch. & Dutreix M. « An attribute estimation technique for the analysis of microarray data ». In Proc. of the Dieppe Spring school on *Modelling and simulation of biological processes in the context of genomics* (eds. P. Amar, F. Képès, V. Norris and P. Tracqui) Publisher Frontier group, ISBN : 2 91 4601 09 3, 2003, 69-77.

[MER 01] Mercier G., Denis Y., Marc P., Picard L. & Dutreix M., « Transcriptional induction of repair genes during slowing of replication in irradiated Saccharomyces cerevisiae », *Mutation Research* 487, 2001, 157-172.

[MER 04] Mercier G., Berthault N., Mary J., Peyre J., Antoniadis A., Comet J-P., Cornuéjols A., Froidevaux Ch. & Dutreix M. « Biological detection of low radiation by combining results of two microarray analysis methods ». Nucleic Acids Research, 2004, Vol.32 (1), e12, jan. 13, 8 pages.

[NG 01] Ng A. & Jordan M., « Convergence rates of the voting gibbs classifier, with application to Bayesian feature selection », ICML-2001.

[PAN 03] Pan W. "On the use of permutation in the performance of a class of nonparametric methods to detect differential gene expression", *Bioinformatics*, Vol.19,n°11, 2003, pp.1333-1340.

[PAR 01] Park P., Pagano M. & Bonetti M., « A non parametric scoring algorithm for identifying informative genes from microarray data », *Pacific Symposium on Biology*:52-63,2001.

[TRO 02] Troyanskaya O., Garber M., Brown P., Botstein D. & Altman R., « Nonparametric methods for identifying differentially expressed genes in microarray data », *Bioinformatics*, Vol.18, no.11, pp.1454-1461, 2002

[TUS 01] Tusher V., Tibshirami & Chu GG., « Significance analysis of microarrays applied to the ionizing radiation response », *PNAS*, April, 2001, Vol 98, n°9, 5116-5121.

[VAN 02] Van't Veer L., Dai H., van de Vijver M., He Y., Hart A., Mao M., Peterse H., van der Kooy K., Marton M., Witteveen A., Schreiber G., Kerkhoven R., Roberts C., Linsley P., Bernards R. & Friend S., « Gene expression profiling predicts clinical outcome of breast cancer », *Nature*, 415, January, 2002.

[XIN 01] Xing E., Jordan M. & Karp R., « Feature selection for high-dimensional genomic microarray », *Proc. of the Int. Conf. on Machine Learning*, ICML-2001, 601-608.

[YAN 02] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J & Speed TP. « Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation ». *Nucleic Acids Re*s. 2002 Feb 15;30(4):27-28.

[ZHA 03] Zhao Y & Pan W. «Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments», *Bioinformatics*, Vol.19, n°19, 2003, pp. 1046-1054.