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

J. Mary[1], G. Mercier[2], J-P. Comet[3],
A. Cornuéjols[1], Ch. Froidevaux[1] et M. Dutreix[2]

[1] LRI - CNRS UMR 8623 – Bât 490 – Université Paris Sud – F- 91 405 Orsay
[2] Institut Curie – CNRS UMR 2027 – Bât 110 – Centre Universitaire – F- 91405 Orsay
[3] LAMI – Université d'Evry – Tour Evry 2 – 523 Place des terrasses de l'agora – 91000 Evry

[1]{mary, antoine, chris}@lri.fr,
[2]{geraldine.mercier, marie.dutreix}@curie.u-psud.fr,
[3]comet@lami.univ-evry.fr

Résumé : Nous nous intéressons à la détermination des effets biologiques de l'exposition à de faibles doses de radiation ionisantes. Pour cela, nous avons utilisé la technique des biopuces à ADN pour mesurer l'ensemble des ARNm de la levure *S. cerevisiae* dans différentes conditions de croissance. Nous présentons dans ce papier une méthode d'analyse basée sur une technique de sélection d'attributs. Avec cette méthode, nous avons mis clairement en évidence l'existence d'une réponse transcriptionnelle pour de faibles doses d'irradiation. De plus, nous avons dégagé un ordre de pertinence sur l'ensemble des gènes permettant de différencier les deux conditions (irradié *versus* non irradié) sur les populations observées.

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.

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) [13]. 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] 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).

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 :

(1) le nombre de gènes impliqués dans la réponse transcriptionnelle ;

(2) l'identité des gènes (pour savoir s'il s'agit des mêmes que lors d'une irradiation plus forte) ;

(3) 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 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 :

(1) Présence de bruit dans les données à deux niveaux :

i) Imprécision de la mesure : bruit classique supposé gaussien, bruit qui est très élevé pour certains gènes (cf doubles mesures);

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

(2) Nombreux attributs : 6157 gènes.

(3) Très faible nombre d'instances : 12 cultures non-traitées, 6 irradiées.

(4) Les classes sont déséquilibrées (elle ne contiennent pas le même nombre d'éléments).

(5) Absence d'indépendance conditionnelle probabiliste entre les gènes.

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 des difficultés d'interprétation biologique d'un grand nombre de gènes, 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, et qu'ils permettent de construire un système de classification très performant pour le diagnostic de lames inconnues.

4. Sélection d'attributs par l'algorithme RELIEF

Compte tenu des caractéristiques des données vues 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 de sélection d'attributs pour lequel plusieurs techniques ont été

proposées dans les dernières années ([7,10,12]). Pour des raisons que nous détaillerons plus loin, nous avons choisi, pour mesurer la pertinence des gènes, la méthode RELIEF [5], qui est une technique de sélection 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.

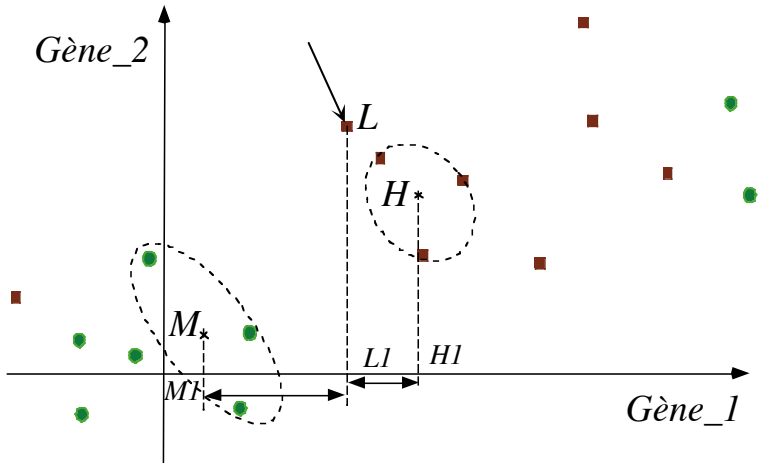

Figure 1. Illustration du fonctionnement de RELIEF dans le cas simple de deux gènes seulement, avec un voisinage de k=3 voisins. Les lames sont alors représentées par des points dans le plan des niveaux d'expression des deux gènes. Pour évaluer la pertinence du gène_1 par exemple, on prend successivement tous les points en compte (somme dans la formule). Ainsi pour chaque lame L, on détermine ses 3 plus proches voisins de la même classe et le barycentre H, ainsi que ses 3 plus proches voisins de l'autre classe et le barycentre M. On reporte ensuite les projections de ces points, L, H et M, sur l'axe du gène_1, soit $L1$, $H1$ et $M1$, afin de calculer la différence : distance($L1$,$M1$) - distance($L1$,$H1$). On voit qu'ici la lame L contribue positivement à la pertinence du gène_1.

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. Pour chacune des m lames, de classe I ou NI, on calcule ses k plus proches voisins dans la même classe et on note H (pour *nearest Hit*) leur barycentre, puis on calcule ses k plus proches voisins dans l'autre classe et on note M (pour *nearest Miss*) leur barycentre. On considère alors les projections des points L, H et M selon le gène dont on calcule le poids. La projection du point L selon le gène g correspond au niveau d'expression de g pour la lame L, tandis que la projection du point H (respectivement M) selon g correspond au niveau d'expression moyen de g pour les k plus proches voisins de L dans la même classe (respectivement dans l'autre classe). On détermine ensuite, d'une part, la distance entre la projection de L et la projection de M, et, d'autre part, la distance entre la projection de L et la projection de H (voir figure 1). La différence entre ces deux valeurs 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.

Ce calcul correspond à la formule suivante :

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

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 (voir [5]) 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 celle-ci tend à surpondérer les gènes (attributs) pour lesquels les expressions mesurées sont très bruitées ou présentent des valeurs aberrantes.

Outre son coût de calcul faible (en optimisant le code), l'un des avantages de RELIEF 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). Qui plus est, le paramètre *k* (nombre de voisins utilisés) permet de contrôler le compromis entre sensibilité et robustesse au bruit. On a choisi *k* de manière à offrir une bonne fiabilité tout en gardant à l'esprit que *k* ne doit pas être trop grand pour ne pas biaiser l'algorithme (on cherche à éviter qu'il ne considère que les moyennes de chacune des deux classes). Ici, plusieurs essais ont été réalisés avec différentes valeurs de *k* (1, 2, 3, 4 et 5) et *k*=3 a été déterminé empiriquement comme le meilleur choix.

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. Il est donc nécessaire de se comparer 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).

Pour cela, nous avons constitué de nouveaux jeux de données dans lesquels les classes des lames ont été aléatoirement attribuées tout en conservant les proportions 6 I et 12 NI - car RELIEF 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 RELIEF 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 2) montrent qu'effectivement pour tout niveau de corrélation donné (poids calculé par RELIEF), 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 lorsqu'aucun changement cellulaire ou génétique n'est par ailleurs détectable. Plus précisément :

(1) Pour les poids positifs (les seuils négatifs ne sont pas représentés car non significatifs), l'écart entre les deux courbes pour chaque poids est supérieur à 500 gènes jusqu'au seuil de poids de 0.2 environ. Par ailleurs, pour des poids compris entre 0.1 et 0.2 , pour lesquels une corrélation significative des gènes avec la classe apparaît, le rapport du nombre de gènes détectés dans les vraies données (courbe supérieure) et du nombre de gènes qui pourraient apparaître corrélés par pure chance (courbe inférieure) est supérieur à 3.

(2) En moyenne, aucun gène ne peut atteindre un poids supérieur ou égal à 0.5 par pure chance (la courbe inférieure passant sous le seuil de 1 gène). À ce seuil, on observe cependant 35 gènes corrélés dans les données vraies (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.

(3)

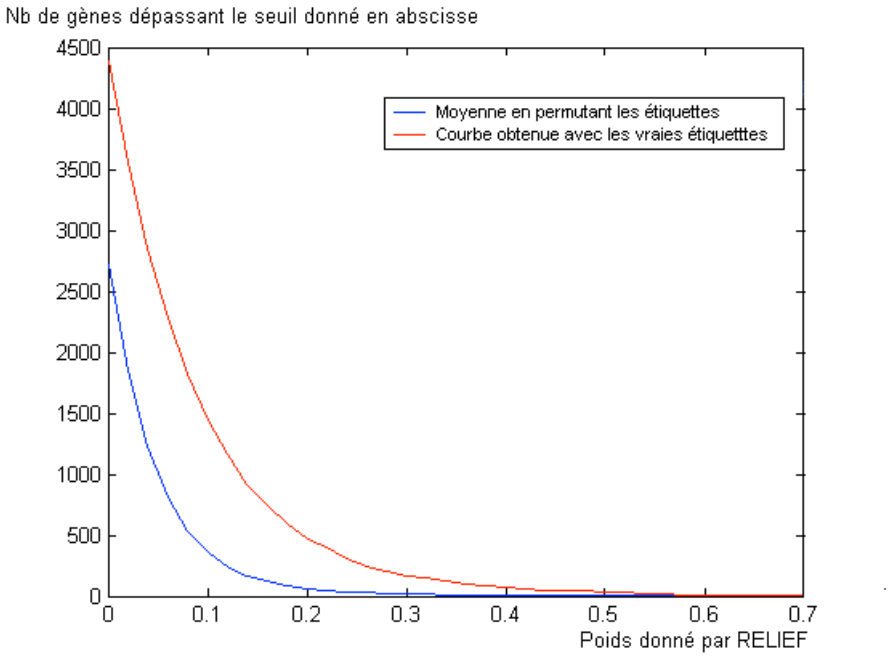

Figure 2- Courbe du nombre de gènes dépassant le seuil de corrélation indiqué en abscisse, avec les vraies classes et courbe moyenne obtenue aléatoirement en permutant les classes, en respectant les proportions des classes initiales.

Nous en concluons que 35 gènes au moins semblent fortement impliqués dans la réponse à une faible irradiation. Notons qu'il s'agit là d'une estimation prudente, minimisant le nombre de faux-positifs parmi les gènes détectés. Retenir les 500 meilleurs selon RELIEF conduirait par exemple à s'exposer à la possibilité d'avoir environ 1/10 de faux positifs d'après la figure 1. En se limitant à 35 gènes, on fait le choix de réduire la probabilité de retenir des gènes non pertinents, au prix de rejeter un grand nombre de gènes pertinents. Une question pratique

essentielle est alors d'examiner si ces 35 gènes peuvent suffire à établir un diagnostic d'exposition à de faibles radiations.

6. Étude de l'utilisation de la méthode en classification

Une fois les gènes (ou encore attributs) sélectionnés, il est possible d'utiliser une méthode classique d'apprentissage supervisé pour obtenir une méthode de décision permettant de classer des lames inconnues. Plusieurs de ces méthodes ont été envisagées ou utilisées (SVM, arbres de décision, *k*-plus proches voisins et utilisation d'une méthode de *k*-means [1, 7]). La préférence des biologistes nous a fait retenir les méthodes de classification bayésienne naïve [1] (on suppose que les classes sont équiprobables et que les attributs sont indépendants). Chaque classe a été caractérisée par une loi gaussienne dans l'espace à 35 dimensions, et la classe d'une nouvelle lame a été calculée par le principe du maximum de vraisemblance. Le taux de reconnaissance est alors de 100% sur les données d'apprentissage.

Les biologistes ont alors fourni un second jeu de données (6 lames correspondant à des traitements non communiqués aux informaticiens) pour voir si la méthode s'étendait à des expériences où les doses d'irradiation étaient beaucoup plus faibles. Les prédictions obtenues répondent aux attentes des biologiste pour les quatre lames ayant subi un traitement de l'ordre de 10 à 0.1 mGy/h. En revanche, la classification obtenue n'est pas correcte pour les deux lames correspondant à des niveaux d'irradiation beaucoup plus faibles (inférieurs à 0.03 mGy/h). Cela peut suggérer une limite de la méthode de détection (approche par microarrays). Mais il est également possible que cela indique une limite biologique, c'est-à-dire un seuil en-dessous duquel l'expression des gènes sélectionnés n'est plus modifiée.

7. Interprétation biologique des résultats

S'il semble possible d'utiliser la méthode de sélection d'attributs pour déterminer des gènes permettant le diagnostic d'une exposition à de faibles doses d'irradiation, il est tout aussi essentiel d'étudier 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 RELIEF, 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 indique le nombre d'ORFs induites (resp. réprimées) parmi les 171 sélectionnées, qui participent à l'activité mentionnée en première colonne. 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), tandis que la quatrième colonne indique le pourcentage des ORFs ayant cette activité (on considère alors l'ensemble de tous les gènes de la levure). La dernière colonne donne alors le taux de sur-représentativité de l'activité considérée dans la population des 171 meilleurs gènes sélectionnés.

Sur la Table 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.

function of 91 induced genes/171	number of ORFs	% in this list	% total ORFS (6	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

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 la fonction mise en œuvre par l'analyse des gènes sélectionnés est connue pour intervenir dans l'élimination de certains des produits cellulaires des rayonnements ionisants (radicaux libres).

8. Comparaison avec d'autres méthodes d'analyse

L'utilisation de RELIEF pour la sélection de gènes dans l'analyse du transcriptome est originale. Nous avons souligné dans la section 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 [3,8,9].

Nous l'avons par ailleurs comparée avec une méthode très courante : la méthode ANOVA [2] 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 [10]). L'analyse de la variance a déjà été utilisée pour les données d'expression de gènes [4] et a permis de proposer une normalisation des données des microarrays tout en permettant une grande part 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 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 RELIEF : mélange de classes et calcul des nouvelles statistiques de Fisher (voir section 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 [10], 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 RELIEF et ceux obtenus par ANOVA. Considérons, par exemple, les 500 gènes classés comme les plus pertinents par RELIEF 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 RELIEF 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 RELIEF, 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 RELIEF, 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 RELIEF et de sa bonne résistance au bruit, l'interprétation biologique que nous avons déjà réalisée porte sur les gènes obtenus par RELIEF (cf. Table 1).

9. Conclusion

Grâce à l'analyse par RELIEF 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. Nous explorons actuellement les propriétés de ces gènes sélectionnés, tout en poursuivant la validation en testant leur pouvoir prédictif sur d'autres lames et avec d'autres outils technologiques. La transposition à d'autres études (telles que, par exemple, la classification de tumeurs chez l'homme) de la méthode utilisée dans ce travail sera également étudiée.

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).

Références

[1] Cornuéjols A. et Miclet L. : Apprentissage artificiel. Concepts et algorithmes. Eyrolles, 2002.

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

[3] Grant G., Manduchi E. et 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.

[4] Kerr M.K., Martin M. et Churchill G.A., Analysis of variance for gene expression in microarray data, Journal of Computational Biology, 2000, 7(6), 818-837.

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

[6] 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.

[7] Ng A. et Jordan M., Convergence rates of the voting gibbs classifier, with application to Bayesian feature selection, citeseer.nj.nec.com/445984.html.

[8] 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.

[9] Troyanskaya O., Garber M., Brown P., Btstein D. et Altman R., Nonparametric methods for identifying differnretially expressed genes in microarray data, Bioinformatics, Vol.18, no.11, pp.1454-1461, 2002

[10] Tusher V., Tibshirami et Chu GG., Significance analysis of microarrays applied to the ionizing radiation response, PNAS, April, 2001, Vol 98, n°9, 5116-5121.

[11] 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. et Friend S., Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, 415, January, 2002.

[12] Xing E., Jordan M. et Karp R., Feature selection for high-dimensional genomic microarray, Proc. of the Int. Conf. on Machine Learning, ICML-2001, 601-608, 2001.

[13] 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 Res. 2002 Feb 15;30(4):27-28.