Une méthode implicative pour l'analyse de données d'expression de gènes

Gérard Ramstein

LINA, équipe COD
Ecole polytechnique de l'université de Nantes
Rue Christian Pauc BP 50609 44306 Nantes cedex 3
gerard.ramstein@univ-nantes.fr

Résumé. Nous présentons une méthode d'extraction d'associations basée sur l'analyse statistique implicative et la notion de rang. Nous avons adapté le concept d'intensité d'implication à des classements pour découvrir des relations partielles robustes vis à vis du bruit et des variations d'amplitude. Appliquée aux données de puces à ADN, cette méthode met en évidence des relations entre des formes d'expressions particulières de gènes. Ces associations peuvent être révélatrices de mécanismes de corégulation génique et donc contribuer à l'analyse biomédicale. Nous montrons que cette définition de l'intensité d'implication apporte une connaissance plus fine des relations entre les gènes que les méthodes usuelles de corrélation et qu'elle permet notamment de discriminer entre différents phénotypes avec une précision comparable aux techniques de classification les plus abouties dans ce domaine.

1 Introduction

La technologie des puces à ADN permet d'analyser l'expression simultanée de milliers de gènes. L'étude du transcriptome représente un enjeu considérable, tant du point de vue de la compréhension des mécanismes du vivant que des applications cliniques et pharmacologiques. Malheureusement, les données d'expression sont entachées de multiples bruits. D'une part, la complexité du protocole expérimental conduit à une réduction de la précision des mesures. D'autre part, la variabilité naturelle de l'activité cellulaire induit des différences notables d'amplitude d'expression entre les gènes, phénomène également perceptible en considérant plusieurs patients présentant le même phénotype.

Cette étude propose une méthode d'analyse implicative des règles d'association sur les données du transcriptome. Elle utilise l'approche de Régis Gras [Gras 1996] en considérant non pas les mesures elles-mêmes, mais le rang des observations. Cette optique permet de s'affranchir des valeurs numériques en considérant des zones de classement dans les mesures d'expression. S'intéresser au classement a l'avantage d'améliorer la robustesse des algorithmes en les rendant insensibles à des transformations monotones des données. On peut émettre une analogie avec un tableau notes scolaires. Chaque enseignant possède son propre système de notation et aura une sévérité différente vis à vis des réponses données par les élèves. Le corps professoral pourra plus facilement s'entendre sur les résultats relatifs des élèves que sur les notes. Un élève situé en haut de classement en mathématique et en physique sera ainsi considéré comme bon, même si sa note diffère entre les deux matières de quelques points. Notre approche peut aisément se généraliser à d'autres domaines. Dans le problème dit du panier de la ménagère, on s'intéressera ainsi au volume d'achat des clients. Les connaissances seront donc relatives à des niveaux de consommation des produits considérés.

Nous allons dans une première partie présenter un état de l'art du domaine d'étude. Nous donnons ensuite le cadre conceptuel des règles d'association utilisé dans notre approche. Nous présentons dans une troisième partie une application à la classification de tumeurs.

2 Etat de l'art

Le traitement des données incertaines ou imprécises a déjà fait l'objet de travaux dans le cadre de l'analyse implicative. Dans [Gras 2001], la méthode proposée consiste à définir une partition optimale des données puis à rechercher des implications entre domaines, ces domaines étant constitués à partir de l'unions d'éléments de la partition obtenue. Une approche parallèle a été développée, basée sur la logique floue [Gras 2005]. Ces deux méthodes reposent sur une partition préalable des mesures avant l'analyse implicative proprement dite. Comme les distributions de nos données se sont avérées monomodales, nous avons préféré rechercher directement l'implication optimale sans passer par un prétraitement des données qui risquerait d'introduire un biais dans l'extraction des règles.

A notre connaissance, l'analyse statistique implicative n'a pas encore été appliquée à l'étude du transcriptome. Cependant, plusieurs études concernent l'extraction de règles d'association à partir de données d'expression. Un travail, basé sur l'algorithme Apriori [Agrawal 1993] et utilisant les notions habituelles de support et de confiance, a été mené sur le génome de la levure [Creighton 2003]. Les données ont été discrétisées

à partir de seuils prédéfinis pour caractériser trois niveaux d'expression (sous-expression, expression normale, sur-expression). [Tuzhilin 2002] présente une analyse sur les biopuces définissant un ensemble d'opérateurs adaptés à de grands volume de règles. Ces deux méthodes présentent l'inconvénient d'être tributaires du paramétrage des seuils de discrétisation. Dans [Carmona-Saez 2006], l'étude est enrichie par la prise en compte de connaissances a priori sur les gènes. Une méthode originale a été développée dans [Cong 2004] à travers un outil dénommé FARMER. Cette méthode recherche des ensembles de règles possédant un support commun et s'appuie sur une technique de discrétisation entropique. Dans [Becquet 2002], les auteurs ont appliqué l'algorithme Min-Ex. L'analyse porte sur des données binaires, la valeur logique associée prenant en compte le fait que le gène est considéré comme sur-exprimé ou non. L'approche la plus proche de la nôtre est certainement celle proposée dans [Jin 2006]. Les auteurs y définissent le concept de patron émergent comme un ensemble d'opérateurs booléens sur des valeurs d'expression. Cette méthode revient à extraire des intervalles d'expression spécifiques à chaque gène en optimisant une mesure entropique. Cette étude diffère de la nôtre par le fait qu'elle suppose la connaissance de classes d'individus et que la procédure est globale : la prise en compte de la totalité des observations induit un risque de perte d'implications statistiquement significatives.

3 Intervalles de rang

Notre étude porte sur une matrice *M(k, l)* de données numériques, où *k* représente un individu, *l* une observation et *M(k, l)* la mesure effectuée. Dans l'exemple cité en introduction, la matrice peut définir un ensemble de notes, *k* désignant une matière et *l* un élève. Pour les données de biopuces, *k* correspond à un gène et *l* à une expérimentation. Dans une étude clinique, *l* représentera par exemple un patient soumis à une condition expérimentale particulière, comme l'absorption d'une médicament. On notera *m* le nombre d'individus et *n* le nombre d'observations. Il est à noter que même si nous traitons de données réelles, notre étude peut s'appliquer à n'importe quelle valeur ordinale. De même, il est possible de transposer la matrice si on s'intéresse à des règles portant sur les observations plutôt que sur les individus. Nous appelons profil d'un individu *k* le vecteur $p(k)=\left(M(k,l),\, l\in[1,n]\right)$ et supposons l'existence d'un opérateur *rang* qui délivre les observations d'un profil dans l'ordre croissant des valeurs d'expression. Plus précisément, nous avons $rang(p(k))=\left(l_1, l_2, \Lambda, l_n\right)$, où les observations $l_1\in[1,n]$ vérifient la condition suivante : $M(k,l_1)\leq M(k,l_2)\leq\Lambda\leq M(k,l_n)$. Soit par exemple un profil représentant les notes en Physique : $p(k)=\left(9,4,17,12\right)$, *rang(p(k))* renvoie alors le vecteur (2, 1, 4, 3). Ce vecteur représente le classement des élèves par ordre croissant dans la matière, le plus mauvais élève (note 4/20) étant désigné par l'indice $l_1=2$ et le meilleur (17/20) par l'indice $l_4=3$.

Nous nous proposons d'extraire des règles d'association entre profils. Ces règles vont mettre en évidence des intervalles d'étude. Un intervalle de rang permet de déterminer des niveaux de classement (i.e. d'expression) sans avoir à définir des seuils de mesure. Pour l'analyse du transcriptome, ces intervalles vont se rapporter à des niveaux d'expression, telles que la sur-expression ou la sous-expression. L'analyse de notes scolaires permet quant à elle d'extraire des règles telles que si un élève est en bas de classement en Mathématique, il sera également en bas de classement en Physique. Sur l'exemple précédent, l'intervalle [1, 3] désignera la zone de classement (2, 1, 4), relative aux trois plus basses notes du profil, à savoir 4, 9 et 12, correspondant aux valeurs $M(k,l),\, l\in(l_1,l_2,l_3)$. Notre approche considère tous les intervalles possibles, à savoir l'ensemble des sous-intervalles de [1, *n*] :

$$I=\left\{\,[p,q],\,(p,q)\in[1,n]^2,\, p\leq q\,\right\} \qquad (1)$$

A un intervalle $i\in I$ est associé un *intervalle de rang* $r_k(i)$, défini comme suit : $r_k(i)=\left(l_j\in rang(p(k)),\, j\in i\right)$. Notons que par raccourci de langage nous désignons sous le terme d'intervalle de rang [*p,q*] les *observations* relatives à cet intervalle, à savoir $\left(l_p, l_{p+1}, \Lambda, l_q\right)$. La table 1 donne deux profils relatifs aux individus A et B. Nous remarquons par exemple que $r_A(i)=(3,5,7,9)$ pour $i=[1, 4]$ et $r_B(j)=(9,7,2,5,3)$ pour $j=[5, 9]$. Ces deux intervalles de rang comporte des observations communes. Ce phénomène semblerait indiquer une association entre le début de classement du profil de A (intervalle i) et la fin de classement du profil de B (intervalle j). Pour savoir si ce phénomène est statistiquement significatif, nous allons reprendre l'approche de [Gras 1996] et considérer deux ensembles α et β de mêmes tailles respectives que $r_A(i)$ et $r_B(j)$. Ces deux ensembles sont évidemment inclus dans l'ensemble des observations O. Notons que dans notre approche l'hypothèse nulle consiste à affirmer que

l'opérateur de classement *rang* n'apporte aucune information utile. Dans ce cas, prendre deux intervalles de rang reviendrait à sélectionner les ensembles α et β au hasard, i.e. sans tenir compte du classement des observations.

Nous reprenons le concept d'intensité d'implication telle qu'il est défini dans [Gras 1996] ainsi que la mesure de qualité φ(α,β), où α et β représentent les ensembles vérifiant respectivement la prémisse et la conclusion d'une règle A → B. Cette mesure définit l'étonnement statistique d'observer si peu de contre exemples dans une règle d'association. Nous étendons la définition originelle reposant sur le cardinal $\alpha \cap \overline{\beta}$ aux intervalles de rang :

$$\varphi(A \rightarrow B) = \max(\varphi(r_A(i), r_B(j)), (i, j) \in I^2) \qquad (2)$$

L'expression (2) indique que la qualité de la règle A → B est définie comme la plus grande intensité d'implication entre deux intervalles de rang, le premier étant issu du profil p(A) et le deuxième du profil p(B). Cette définition revient à rechercher les intervalles i et j qui maximisent $\varphi(r_A(i), r_B(j))$. Dans l'exemple de la table 1, les intervalles i = [1, 4] et j = [5, 9] correspondent à la valeur maximale d'intensité d'implication : $\varphi(A \rightarrow B) = 0{,}86$.

profil	1	2	3	4	5	6	7	8	9
p(A)	5	6	1	8	2	9	3	7	4
p(B)	13	17	19	14	18	12	16	11	15

TAB 1 – *Exemples de profils. Les numéros de colonne désignent les indices relatifs à neuf observations effectuées sur deux individus A et B.*

4 Intérêt de l'approche implicative pour l'étude des données d'expression

En matière d'étude du transcriptome, les analyses les plus couramment menées par les biologistes sont basées sur des mesures de corrélation entre profils d'expression. Ces mesures présentent l'inconvénient d'être globales dans le sens où elles font intervenir l'ensemble des observations, alors que la définition (2) recherche une correspondance optimale entre des sous-ensembles d'observations. Pour expliciter la différence entre ces deux approches, nous allons considérer un exemple de règle d'association entre deux gènes (figure 1). Ces gènes appartiennent à l'espèce Saccharomyces cerevisiae, communément dénommée levure du boulanger. Nous avons repris les données de puces sélectionnées dans [Gasch 2002]. Nous avons retenu 89 différentes conditions expérimentales correspondant à différents stress induits tels que le choc thermique. La figure 1 représente l'implication CHA1 → SAM1. Le gène CHA1, intervenant dans le catabolisme de la threonine, est clairement sous-exprimé en réponse à un signal de déficience en acide aminé, et dans une moindre mesure, en nitrogène. SAM1, un gène interférant dans le métabolisme de la methionine, est sur-exprimé pour le même jeu d'observations. Ce jeu correspond à environ 9 % des conditions.

Méthode	Valeur
Intensité d'implication	0,9992
Indice de corrélation de Pearson	-0,16
Indice de corrélation de Kendall	-0,0089

TAB 2 – *Comparaison de mesures effectuées sur l'exemple de la figure1.*

Comme le montre la table 2, Les indices usuels de corrélation ne peuvent déceler de telles associations. Les valeurs obtenues sont trop faibles pour être retenues dans une analyse alors que la mesure implicative exprime que le risque de rencontrer une association de même nature au hasard est inférieur à un pour mille.

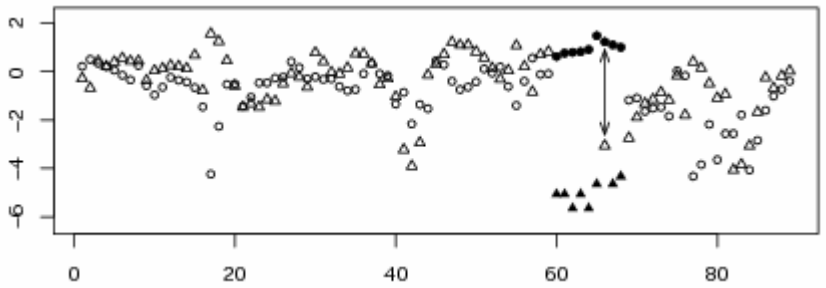

FIG. 1 - *Profils des gènes CHA1 et SAM1. L'axe des abscisses représente 89 conditions expérimentales. L'axe des ordonnées représente les mesures d'expression. Le profil du gène CHA1 (YCL064C) est représenté par des triangles et celui de SAM1(YLR180W) par des cercles. Les figures pleines correspondent aux observations appartenant aux intervalles de rang qui maximisent la valeur de l'intensité d'implication. On remarque que ces observations sont communes à une exception près, indiquée par la double flèche : l'observation de SAM1 qui est moins sous-exprimée, n'apparaît pas dans le groupe alors qu'il est bien présent dans celui de CHA1.*

5 Une mesure implicative du pouvoir discriminant des gènes

La section précédente a montré que la méthode implicative peut être utilisée dans le cadre non supervisé. Nous supposerons désormais que les observations appartiennent à un ensemble *C* de classes prédéfinies. Par la suite, une classe correspondra à des tissus provenant de patients présentant un phénotype commun. Nous proposons une technique de sélection des gènes les plus pertinents pour la discrimination de tumeurs basée sur la méthode implicative. La recherche de gènes informatifs est un enjeu majeur en bio-informatique. Il s'agit d'extraire dans un jeu de plusieurs milliers de gènes ceux dont l'expression est la plus significative. L'analyse la plus commune est celle de l'expression différentielle, à savoir la recherche de profils d'expression qui varient d'une classe d'observation à une autre. Nous montrons à partir d'une expérimentation sur deux types de leucémie que l'approche que nous proposons s'avère particulièrement performante. Nous proposons enfin une technique originale de représentation visuelle, appliquée à un jeu de données portant sur des tumeurs cérébrales.

5.1 Une définition implicative du concept de gène discriminant

Les données de puces concernent généralement un nombre important de gènes. La plupart d'entre eux n'aident pas à discriminer les classes, soit parce que leurs expressions ont une faible amplitude de variation, soit parce que leur expression est indépendante des classes d'observations. Une étape de sélection des gènes est donc nécessaire. La technique la plus communément partagée repose sur la puissance discriminative des tests statistiques telles que le test de Student ou ANOVA (voir [Chen 2005] pour une synthèse des mesures utilisées sur les données d'expression). Nous proposons d'utiliser l'intensité d'implication pour déterminer les gènes les plus discriminants. On appellera fonction d'étiquetage la fonction $L(o_k)=c_l$, où o_k est une observation de *O* et $c_l \in C$. Notre approche est fondée sur des règles de classification de la forme : $r_g(i) \rightarrow Oc$, où la conclusion désigne les observations de classe *c* et la prémisse l'intervalle de rang associé au gène *g* et à l'intervalle *i*. Cette règle peut être explicitée de la façon suivante : si, pour une observation, sa mesure d'expression relative au gène g figure dans la zone de classement défini par l'intervalle i, alors cette observation appartient probablement à la classe *c*. Nous restreignons par la suite le domaine des intervalles défini en (1) en imposant des niveaux d'expression ayant une signification biologique précise, à savoir la sous-expression (intervalle de la forme [1, q]) ou la sur-expression (intervalle [p, n]). L'ensemble I des intervalles est donc désormais : $I=\{[p,q],(p,q)\in[1,n]^2,\ p\leq q\ ,p=1\wedge q=n\}$.

La règle $r_g(i) \rightarrow Oc$ peut dès lors s'interpréter plus simplement : elle signifie que si on observe une dérégulation du gène *g* sur un patient *o*, alors ce patient possède probablement le phénotype *c*. Cette implication correspond bien à la notion de gène discriminant.

Le pouvoir discriminant'un gène g vis à vis d'une classe c sera exprimée par l'expression suivante :

$$\varphi_c(g)=\max(\ \varphi(\ r_g(i), O_c),\ i\in I,\ O_c=\{o, L(o)=c\ \}) \qquad (3)$$

La mesure $\varphi_C(g)$ est définie par la maximisation de l'intensité d'implication des règles de classification de type $r_g(i) \rightarrow Oc$.

Il est à noter que la définition (3) diffère uniquement de l'expression (2) par une modification de l'ensemble conclusion : $r_B(j)$ est remplacé par l'ensemble O_c des observations de classe c. Une remarquable propriété de l'intensité d'implication est de prendre en compte le nombre d'observations appartenant à O_c: pour un intervalle fixe i, le fait d'accroître la population O_c entraîne une diminution de la qualité de la règle. Il est en effet normal de retrouver une proportion notable d'observations de classe c si cette dernière est sur-représentée.

5.2 Méthode de sélection

Soient G l'ensemble total des gènes définis dans l'ensemble des expérimentations O et M la matrice de données d'expression associée. Soient C les classes d'observations et L la fonction d'étiquetage de O vers C. L'algorithme présenté en figure 2 extrait les K gènes les plus discriminants pour chaque classe de C.

```
Algorithme Sélection
Entrée :
M, G : matrice de valeurs d'expression et ensemble des genes considérés
C, L : ensemble des classes et fonction d'étiquetage
K    : le nombre de gènes discriminants à retenir
Sortie :
gd   : ensemble {(g,c,φ)} des gènes discriminants où g est le gène retenu,
       c la classe discriminée et φ la mesure de qualité définie en (3)
Début

Pour chaque classe c ∈ C faire
        ListeGènes ← Ø
        Pour chaque gene g ∈ G faire
                φ ← φc(g)
                ListeGènes ← ListeGènes ∪ {(g,c,φ)}
        fait
        trier les triplets de ListeGènes par ordre décroissant de φ
        Soit sélection l'ensemble des K premiers triplets de la liste triée
        gd ← gd ∪ sélection

fait

Fin.
```

FIG. 2 – *Algorithme pour l'extraction de gènes discriminants*

Notons que cet algorithme considère chaque classe de manière indépendante. Il peut arriver qu'un gène soit considéré comme discriminant pour plusieurs classes. On remarquera que la mesure $\varphi_C(g)$ sera différente selon la classe considérée. Il peut en effet par exemple arriver qu'un gène soit nettement sous-exprimé pour une classe donnée et qu'il soit au contraire sur-exprimé dans une autre, à un degré moindre. Il faut aussi noter que dans le cas particulier d'une expérimentation à deux classes (e.g. patients sains versus patients malades), on pourrait s'attendre à ce qu'un gène discriminant pour une classe le soit aussi pour l'autre. En ce qui concerne notre approche, ce fait n'est pas forcément vérifié. Il peut en effet arriver que les deux extrêmes du classement des observations n'aient pas la même homogénéité. Dans ce cas, les valeurs de $\varphi_{C_1}(g)$ et de $\varphi_{C_2}(g)$ différeront, et selon la qualité relative du gène vis à vis des autres gènes candidats, il se peut fort bien qu'il soit retenu pour une classe et rejeté pour une autre.

5.3 Application à une étude portant sur deux types de leucémie

Nous appliquons notre algorithme de sélection sur une étude portant sur deux types de leucémie [Golub 1999]. La leucémie se caractérise par une prolifération maligne de cellules d'origines hématopoïétiques peu matures et rapidement diffusantes. Cette maladie se caractérise par une atteinte massive de la moelle osseuse, due au développement de lymphomes malins. On distingue les leucémies aiguës lymphoblastiques (notées ALL par la suite pour *Acute Lymphoblasic Leukemia*) des leucémies aiguës myéloblastiques (notées AML pour *Acute Myeloid Leukemia*). La distinction entre ces deux formes est essentielle pour le succès des thérapies envisagées : le traitement diffère selon l'une ou l'autre de ces deux classes de leucémie. Le jeu de données comporte 38 patients (27 patients ALL et 11 patients AML) et concerne 3571 gènes.

Nous nous proposons d'extraire dans ce jeu les gènes les plus discriminants en appliquant l'algorithme de sélection décrit précédemment. Nous avons étudié le pouvoir discriminantes gènes du jeu complet sur la base de la mesure (3), dont nous donnons ici une version logarithmique : $\lambda_c(g) = -\log_{10}(1-\varphi_c(g))$. Cette transformation présente plusieurs avantages : l'indice de qualité n'est plus borné et les valeurs sont plus facilement interprétables. Une règle de qualité $\lambda_c(g)$ possède ainsi un risque de $10^{-\lambda_c(g)}$ d'être dû au hasard. L'analyse révèle que 10% des gènes ont un pouvoir discriminant important ($\lambda_c(g) > 3{,}5$ soit $\varphi_C(g) > 0{,}9997$).

Bien que 300 gènes peuvent être considérés comme discriminants, les auteurs de l'étude n'ont retenu qu'une liste de 50 gènes les plus informatifs, tout en indiquant le caractère arbitraire de ce nombre. Comme ils ont développé une technique originale de sélection, il est intéressant de comparer leur approche avec la nôtre. Nous avons donc fixé le paramètre K de l'algorithme de la figure 2 à 25, puisque nous avons deux groupes de gènes pour les classes ALL et AML. Nous avons obtenu une liste de gènes discriminants dans laquelle figurent 14 gènes appartenant à la liste publiée par les auteurs. La table 3 compare le pouvoir discriminant des deux jeux. On remarque le gène le plus discriminant est commun aux deux jeux de gènes et que la moyenne est du même ordre de grandeur. On observe cependant une plus grande dispersion dans le jeu de Golub et al. De même, la valeur médiane est significativement plus faible par rapport à notre liste.

Liste de gènes	max	médiane	min	moyenne	variance
Golub et al.	9.08	5.44	4.55	5.12	10.16
notre liste	9.08	6.47	5.49	5.92	11.80

Tab 3 – Comparaison des valeurs de $\lambda_c(g)$ sur les 50 gènes sélectionnés par Golub et al. et selon notre méthode.

Pour comparer la puissance discriminative de ces deux jeux de gènes, nous avons procédé à une validation croisée sur les données en utilisant le même classifieur, à savoir les 3 plus proches voisins. Cette technique a été retenue parce qu'elle prend en considération la distance entre points dans tout l'espace des gènes, contrairement à d'autres méthodes qui vont privilégier certains gènes (arbres de décision, forêt aléatoire, séparateurs à vastes marges) et qui apporteraient donc un biais pour la comparaison. Nous avons appliqué l'algorithme sur la matrice d'expression réduite comportant les mêmes patients et portant sur le jeu de gènes considéré. La table 4 montre que notre sélection donne des résultats supérieurs à celle des auteurs.

% test	Golub & al.	Nos résultats
50%	3.11	0.79
25%	1.67	0.11
10%	1.00	0.00
2.6%	0.00	0.00

Tab 4 – Comparaison des taux d'erreurs en validation croisée. La première colonne indique le pourcentage considéré en test sur le jeu de données. Les deux autres colonnes donnent les taux d'erreurs en pourcentage sur 100 jeux de validation aléatoires.

5.4 Représentation des données basée sur l'intensité d'implication

Nous avons utilisé l'intensité d'implication pour déterminer l'importance individuelle de chaque gène vis à vis d'une pathologie. Comme cette mesure permet également de définir l'association entre individus, il est possible de caractériser les relations liant les gènes entre eux et de projeter ces derniers dans un espace de représentation abstrait. L'échelonnement multidimensionnel (ou *multidimensional scaling* [Kruskal 1978]) est en effet une technique assurant une représentation des données à partir d'une matrice de dissimilarité. Nous proposons de définir une telle mesure à partir de l'intensité d'implication. On considère pour se faire la similarité $sim(g_1,g_2)$ entre deux gènes g_1 et g_2 définie comme l'intensité d'implication maximale entre deux gènes : $sim(g_1,g_2)=\max(\lambda(g_1\rightarrow g_2),\lambda(g_2\rightarrow g_1))$ où $\lambda(g_1\rightarrow g_2)=-\log_{10}(1-\varphi(g_1\rightarrow g_2))$. La dissimilarité est obtenue par la relation suivante : $dissim(g_1,g_2)=\max(sim(g_1,g_2),(g_2,g_1)\in G)-sim(g_1,g_2)$. Nous pouvons dès lors grâce à l'échelonnement dimensionnel associer à un gène g un point $P(g)$ du plan. Il nous a paru intéressant de visualiser également le lien existant entre une observation o et l'ensemble des gènes discriminants. Cette relation peut être spécifiée par la présence de l'observation o dans l'intervalle de rang associé à une règle de sélection. A partir de l'ensemble des règles $r_g(i)\rightarrow Oc$ issues de l'algorithme de sélection, on définit G_o comme l'ensemble des gènes g de G qui vérifient $o\in r_g(i)$. Pour représenter la position d'une observation dans l'espace des gènes, nous associons à o le point P(o) défini par la moyenne arithmétique des gènes de G_o :

$$P(o)=\frac{1}{|G_o|}\sum_{g\in G_o}P(g).$$

Nous avons appliqué cette méthode de représentation sur un jeu de données complexe faisant intervenir cinq types de tumeurs du système nerveux central. Le jeu comporte 42 échantillons de tissus, répartis comme suit : 10 medulloblastomes, 10 gliomes mains, 10 tumeurs rhabdoïdes et tératoïdes atypiques, 8 tumeurs neuroectodermiques primitives (PNETs) et 4 tumeurs cérébelleuses. Les données brutes sont disponibles sur le site : *Whitehead Institute Center for Genomic Research* (http://www-genome.wi.mit.edu/cancer). Après une étape de filtrage, le jeu de données correspond à 5 597 gènes.

La figure 3 montre l'espace des 50 gènes les plus discriminants (10 gènes retenus par classe). On observe une séparation spatiale des gènes relativement à chaque catégorie, même si certaines marges inter-classes se révèlent étroites. Les chiffres représentent les classes des observations o projetées sur l'espace des gènes. On remarquera que dans une situation idéale, les observations devraient coïncider avec le barycentre des gènes de leur classe d'appartenance. L'éloignement des points par rapport à leur position idéale met en valeur l'attraction de gènes portant sur des règles de classification impropres. De telles déviations montrent en effet que l'observation considérée vérifie des règles dont la conclusion porte sur une classe externe, autrement dit qu'elle y figure en tant que contre-exemple. La visualisation proposée permet donc d'analyser la représentativité des observations vis à vis des classes. On observe notamment sur la figure 3 une bonne homogénéité des classes 1 et 2 ainsi qu'un déplacement des individus de la classe 4 vers la classe 5.

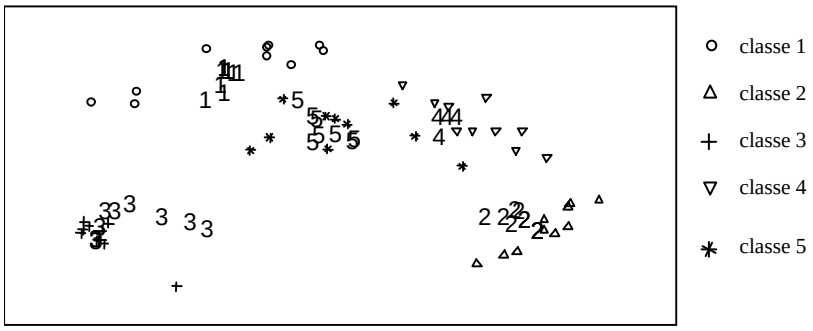

FIG. 3 – *Représentation des observations dans l'espace des gènes. Les symboles graphiques représentent les gènes discriminants pour la classe indiquée en légende. Les chiffres représentent les observations.*

6 Application à la classification de tumeurs

La section précédente a permis de définir une mesure du pouvoir discriminant des gènes. Comme nous avons basé notre algorithme de sélection sur des règles de classification de type $r_g(i) \rightarrow Oc$, il est naturel d'envisager d'utiliser celles-ci pour prédire la classe d'un patient d'après son profil d'expression. Cette catégorisation est un des enjeux majeurs de la technologie des biopuces : elle permet de diagnostiquer l'existence d'un cancer à un stage précoce, lorsque la maladie s'exprime dans les cellules sans qu'on observe encore des signes cliniques manifestes. Par ailleurs, la prédiction basée sur l'expression des gènes permet de distinguer entre différents types de tumeurs même si leur apparence morphologique tumorale est identique. Nous allons dans un premier temps présenter les algorithmes mis en œuvre pour classer des observations, puis nous allons présenter les jeux de données ainsi que les méthodes utilisées pour comparer nos résultats avec ceux de la littérature.

6.1 Algorithmes pour la classification

Pour caractériser la capacité prédictive des règles extraites, nous proposons une approche supervisée comprenant un jeu d'apprentissage $\mathcal{A}$ = { *G*, *M*, *O*, *L*, *C*}, où *G* est un ensemble de gènes, *M* les mesures effectuées sur un ensemble *O* d'observations, et *L* la fonction qui attribue à chaque observation une classe de *C*. La figure 4 présente l'algorithme d'extraction des règles de classification, qui est analogue dans son principe à celui qui nous a servi à sélectionner les gènes. On notera que le cardinal de l'ensemble des règles extraites $\mathcal{R}$ est $|\mathcal{R}|$=K.|*C*|, K étant un paramètre d'entrée de l'algorithme et |*C*| le nombre de classes considérées.

```
Algorithme Extraction des règles de classification
Entrée :
𝒜       : jeu d'apprentissage
K       : le nombre de règles souhaitées par classe
Sortie :
ℛ       : l'ensemble des règles extraites
Début

Pour chaque classe c ∈ C faire
        ListeRègles ← Ø
        Pour chaque gene g ∈ G faire
                Rechercher l'intervalle de rang r_g(i) qui maximise
                φ( r_g(i), O_c), O_c = {o, L(o) = c}
                ListeRègles ← ListeRègles ∪ {( r_g(i) ,φ_c(g)) }
        fait
        trier les couples de ListeRègles par ordre décroissant de φ_c(g)
        Soit sélection l'ensemble des K premiers couples de la liste triée
        ℛ ← ℛ ∪ sélection

fait
```

FIG. 4 – *Algorithme pour l'extraction de règles de classification*

Le principe de la prédiction d'un échantillon à partir d'une expérimentation peut être décrit comme suit : soit *o* une observation nouvelle de classe inconnue sur laquelle a été effectuée une mesure d'expression sur l'ensemble des gènes *G* et soit *p(o)* le profil d'expression correspondant à ces mesures, il s'agit de relever les prémisses des règles de classification que respecte *o* et définir *L(o)* comme étant la classe la plus souvent rencontrée en conclusion de ces règles.

On ne peut en pratique vérifier directement si *o* respecte la prémisse *P* d'une règle $r_g(i) \rightarrow Oc \in \mathcal{R}$. En effet, $P = r_g(i)$ est défini à partir de l'ensemble O $\in \mathcal{A}$, ensemble dans lequel ne figure pas l'observation *o*. Pour pallier à cette difficulté, nous recherchons où se situerait la nouvelle mesure d'expression M[g,o] relative à un ensemble

$O'=O \cup \{o\}$. Autrement dit, on cherche à savoir si la nouvelle observation s'insère entre les rangs du classement effectué au moment de l'apprentissage. La pratique opératoire est donc la suivante : soit $s(P)=\min(M[g,o],o\in P)$ et $S(P)=\max(M[g,o],o\in P)$, on dira que o respecte P si la condition suivante est réalisée : $s(P)\leq M[g,o]\leq S(P)$.

La figure 5 décrit l'algorithme de prédiction d'une observation o. Son principe repose sur un consensus : la classe attribuée est celle qui a recueilli le maximum de suffrages, les votants étant les règles et les votes leurs conclusions.

On notera que la pertinence du vote est liée à la qualité de la règle : plus celle-ci comporte de contre exemples et plus le risque d'erreur de classification augmente. Il est envisageable de tenir compte de cette propriété en attribuant un poids de vote proportionnel à l'intensité d'implication de la règle. En pratique, nous n'avons pas observé d'améliorations notables en procédant de la sorte. La raison en est vraisemblablement que toutes les règles extraites sont suffisamment pertinentes vis à vis des jeux de données considérés. Cela est dû au nombre K=25 relativement réduit qui a été appliqué pour la sélection des règles lors de l'apprentissage.

```
Algorithme Prédiction

Entrée :
M       : matrice d'expression du jeu d'apprentissage
p(o)    : profil d'expression mesuré sur la nouvelle observation o
ℛ       : l'ensemble des règles extraites lors de l'apprentissage

Sortie :
χ       : la classe prédite pour l'observation o

Variable intermédiaire :
count  : vecteur de taille |C| pour compter les occurrences de
         prémisses satisfaites par l'observation o

Début

Pour chaque classe c ∈ C faire
        count[c] ← 0
        Pour chaque règle r ∈ ℛ faire
                soit g le gène relatif à la règle r,
                soit pg(o) la mesure d'expression relative à g dans le profil p(o),
                soit P la prémisse de la règle r, s(P) et S(P) les valeurs
                minimales (resp. maximales) dans M relativement à g et à la
                prémisse de la règle r.
                si s(P)≤pg(o)≤S(P) alors count[c] ← count[c]+1

        fait
fait
χ ← argmax_{c∈C}(count)

Fin.
```

FIG. 5 – *Algorithme de prédiction d'une observation o.*

6.2 Etude comparative des performances de classification

Outre les jeux de données sur la leucémie et sur le cerveau, qui ont déjà été présentés, nous appuyons notre étude sur des données portant sur le cancer du colon. ce jeu contient des données d'expression sur des tissus du colon. L'étude a été faite sur 62 tissus, dont 22 sains et 40 tumoraux. L'expression de 6 500 gènes a été analysée. Les données sont accessibles sur le site : *Colorectal Cancer Microarray Research* (http://microarray.princeton.edu/oncology). La table 5 résume les différentes caractéristiques des jeux que nous avons étudiés.

Jeu de données	Publication	# tissus	# classes	# gènes	type d'étude
Cerveau	[Pomeroy 2002]	42	5	5 597	sous-types tumoraux
Colon	[Alon 1999]	62	2	2 000	tissus sains / malades
Leucémie	[Golub 1999]	72	2	3 571	sous-types tumoraux

Tab 5 – Présentation des jeux de données publiques sur le cancer utilisés dans notre étude.

Nous comparons les performances de notre classifieur avec les résultats obtenus, soit dans la littérature, soit obtenus en utilisant des classifieurs génériques. Pour le premier type de classifieur, nous avons retenu deux études majeures dans le domaine de l'analyse du transcriptome, à savoir l'algorithme *Gene clustering* [Dettling 2002] et l'algorithme *Fuzzy c-means* [Wang 2003]. Outre ces méthodes spécifiquement dédiées au traitement de données d'expression, nous avons également retenu des techniques de classification supervisée habituellement utilisées dans ce domaine.

Les classifieurs utilisés dans l'étude sont les suivants :

k-plus-proches-voisins : cette méthode requiert les faveurs des biologistes pour sa simplicité d'interprétation [Yeang 2001]. Le classifieur recherche les k plus proches voisins d'un échantillon inconnu en fonction d'une mesure de disance d(x,y). La métrique la plus courante en bioinformatique est le coefficient de Pearson absolu, la distance étant défini par :

$$d(x,y)=1-\frac{\left|\sum_{i=1}^{n}(x_i-\mu(x_i)(y_i-\mu(y_i)\right|}{(n-1)\sigma(x)\sigma(y)}$$

où μ et σ désignent respectivement la moyenne et l'écart-type des profils d'expression. Le classifieur attribue à l'échantillon inconnu la classe majoritaire de ses k voisins (k étant impair, souvent fixé à 3 dans la littérature). Comparée à des techniques plus élaborées, cette méthode donne des résultats satisfaisants, à condition de disposer d'un jeu de gènes pertinents. Dans le cas contraire, la grande dimension de *G* peut être un élément défavorable et rendre peu significatif le calcul de distance entre observations.

Forêt aléatoire : ce classifieur est composé d'un très grand nombre d'arbres de décision. Chacun de ces arbres reçoit les données relatives à un choix aléatoire d'observations (tirage avec remise). A chaque nœud, l'arbre sélectionne l'attribut le plus pertinent parmi un choix aléatoire de gènes. L'algorithme utilise ensuite une prédiction basée sur le consensus, à l'instar de notre méthode : la classe retenue est celle qui a été le plus souvent prédite par les arbres. Cette technique s'est montré particulièrement efficace sur les données du transcriptome. Elle possède un grand pouvoir de prédiction, même quand le nombre d'observations est réduit par rapport au nombre de gènes [Diaz-Uriarte 2006].

Machine à vecteurs de support : cette technique est un des plus performantes en matière d'apprentissage automatique [Vapnik 1995]. Elle est particulièrement efficace dans le cas d'espace de données de haute dimension. Le principe du classifieur consiste à rechercher un hyperplan de séparation optimale entre deux classes d'échantillon dans un espace de caractéristiques. Cette méthode se prête bien à la classification tumorale à partir de biopuces [Lee 2003].

La table 6 présente les résultats obtenus selon la technique du *leave-one-out* (validation croisée en prenant l'ensemble des observations en apprentissage moins une, cette dernière servant de jeu de test, ce principe étant répété pour chaque observation). Sur les trois jeux de données décrits précédemment, notre méthode atteint des performances comparables aux autres classifieurs. Ce résultat est d'autant plus remarquable que notre algorithme est relativement frustre, puisqu'il s'agit d'un simple comptage de règles de classification. Malgré sa simplicité, il rivalise parfaitement avec des techniques sophistiquées.

méthode	cerveau	colon	leucémie
Règles de classification	14.3	12.9	2.8
Gene clustering	11.9	16.1	2.8
Fuzzy c-means	14.3	11.4	4.1
Forêt aléatoire	19.0	14.5	2.8
Séparateurs à vastes marges	11.9	12.9	2.8
3 plus proches voisins	23.8	22.6	1.4

Tab 6 – Comparaison des méthodes de classification. Le tableau indique les taux d'erreurs selon la technique du leave-one-out. Bien évidemment, la méthode basée sur les règles de classification effectue à chaque test une extraction de règles sur le jeu d'apprentissage ; le candidat testé a été retiré de ce jeu.

7 Conclusion

L'approche implicative, appliquée aux données d'expression, présente plusieurs avantages. En premier lieu, elle est plus fine que les techniques qui mesurent des relations de similarité globale. Par mesure globale, nous entendons des estimations basées sur l'ensemble des observations. Il est clair que si un certain nombre d'observations ne participent pas à une relation, ces informations apportent un bruit qui masque l'association entre gènes, comme nous l'avons montré sur un exemple d'expérimentations sur la levure. De la même manière, une association entre conditions expérimentales peut être masquée par des gènes qui ne sont pas régulés de manière coordonnée avec un autre groupe de gènes. C'est la raison pour laquelle une analyse implicative est plus performante que des techniques basées sur des corrélations.

Un deuxième intérêt de notre approche est liée à la robustesse de l'analyse de rang. On remarquera que l'analyse de classement est invariante par rapport à toute transformation monotone des données. Cette propriété est particulièrement utile dans le cadre de données de puces qui subissent un grand nombre de prétraitement (transformation logarithmique, normalisation, …).

Enfin, on rappelle que l'implication est une information orientée, contrairement aux techniques de similarité qui sont symétriques. Cette propriété peut être exploitée dans le cadre du transcriptome. On sait en effet que les gènes sont activés par le biais de facteurs de transcription, qui sont eux-mêmes exprimés dans la cellule. Découvrir des relations de causalité entre l'expression de gènes est un enjeu majeur en bioinformatique. Les jeux de données jusqu'à présent ne permettaient pas de telles analyses, puisque la biopuce n'est que la photographie de l'activité de la cellule à un instant donné. L'accumulation des expérimentations et leur libre diffusion au sein de la communauté scientifique offrent depuis peu la possibilité d'opérer des méta-analyses : la comparaison de multiples jeux de données permet dès lors d'inférer des relations d'implications entre gènes (lorsque tel gène est exprimé, tel autre gène est exprimé, et non l'inverse). C'est une voie d'application prometteuse pour la fouille de données.

Nous avons proposé une méthode originale de sélection des gènes informatifs. La pertinence de la méthode a été vérifiée en démontrant le pouvoir prédictif du jeu sélectionné. Une forme originale de représentation visuelle des données a été proposée pour analyser les gènes d'intérêt et la représentativité des observations. La découverte de gènes discriminants est d'une grande importance pour les applications cliniques, car elle permet de définir des méthodes de diagnostic fiables et relativement peu coûteuses.

Nous avons développé un algorithme d'extraction de règles de classification. L'avantage d'une méthode de classification basée sur les règles est qu'elle délivre une information aisément interprétable par un expert biologiste, contrairement à des méthodes abstraites telles que les machines à vecteurs de support. Malgré sa simplicité de mise en œuvre, notre algorithme s'est révélé aussi performant que les techniques les plus éprouvées dans ce domaine.

Références

Agrawal, R., Imielinski, T., Swami, A., 1993, Mining association rules between sets of items in large databases, In:SIGMOD'93: Proceedingsof the1993 ACMSIGMOD international conference on Management of data.Volume22., ACM Press, pp 207–216.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J, 1999.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96(12) pp 6 745–6 750.

Becquet, C., Blachon, S., Jeudy, B., Boulicaut, J.F., Gandrillon, O., 2002, Strong association-rule mining for large-scale gene-expression data analysis: a case study on human sage data. Genome Biol 3(12).

Carmona-Saez, P., M. Chagoyen, A. Rodríguez, O. Trelles, J. M. Carazo, et A. D. Pascual- Montano, 2006, Integrated analysis of gene expression by association rules discovery, BMC Bioinformatics 7, 54.

Chen, D., Z. Liu, X. Ma, et D. Hua, 2005, Selecting genes by test statistics, Journal of Biomedicine and Biotechnology 2, 132-138.

Cong, G., A. Tung, X. Xu, F. Pan, et J. Yang, 2004, Farmer: Finding interesting rule groups in microarray datasets.

Creighton, C., Hanash, S., 2003, Mining gene expression databases for association rules, Bioinformatics 19 (1) pp 79–86

Dettling, M. et P. Bühlmann , 2002, Supervised selection of genes. Genome. Biol. Res. 3(12), Research, 0069.1-0069.15.

Diaz-Uriarte, R., Alvarez de Andres, S., 2006 : Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7.

Gasch, A. et M. Eisen, 2002, Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, et E. S. Lander, 1999, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, pp 531-537.

Gras, R., Couturier, R., Guillet, F et F. Spagnolo, 2005, Extraction de règles en incertain par la méthode statistique implicative. In 12èmes Rencontres de la Société Francophone de Classification, Montréal, pp. 148-151.

Gras, R., E. Diday, P. Kuntz, et R. Couturier (2001). Variables sur intervalles et variables-intervalles en analyse statistique implicative. In Société Francophone de Classification (SFC'01), Univ. Antilles-Guyane, Pointe-à-Pître, pp. 166-173.

Gras, R., 1996, L'implication Statistique, Grenoble : La Pensée Sauvage.

Jin, X., X. Zuo, K.-Y. Lam, J. Wang, et J.-G. Sun, 2006, Efficient discovery of emerging frequent patterns in arbitrary windows on data streams. In ICDE, p 113.

Lee,Y., Lee,C.K., 2003 : classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics 19(9) pp 1132–1139.

Kruskal, J. B., and Wish, M. (1978), Multidimensional Scaling, Sage University Paper series on Quantitative Application in the Social Sciences, 07-011. Beverly Hills and London: Sage Publications.

Pomeroy, S.L.,Tamayo,P.,Gaasenbeek, M.,Sturla, L.M.,Angelo, M.,McLaughlin, M.E., Kim, J.Y.H., Goumnerova, L.C., Black, P.M., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S., Golub, T.R., 2002 : Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415(6 870) pp 436–442

Tuzhilin, A. et G. Adomavicius, 2002, Handling very large numbers of association rules in the analysis of microarray data. In KDD, pp. 396-404.

Yeang,C.H.,Ramaswamy,S.,Tamayo,P.,Mukherjee,S.,Rifkin,R.M.,Angelo,M., Reich, M., Lander, E., Mesirov, J., Golub, T., 2001 : Molecular classification of multiple tumor types. Bioinformatics 17 Suppl 1 316–322

Vapnik,V.N., 1995 : the nature of statistical learning theory, Springer-Verlag, NewYork, Inc., USA.

Wang, J., T. H. Bø, I. Jonassen, O. Myklebost, et E. Hovig, 2003, Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data, BMC Bioinformatics 4, 60.

Summary

We present a rule extraction method based on the statistic implicative analysis and the concept of ranking. We adapt the definition of the intensity of implication to ranked data. This definition permits to discover partial relationships inside ordered observations. Applied to microarray DNA data, our method extracts relationships between particular forms of gene expressions. These associations may reveal underlying mechanisms of gene coregulation and help biological analysis. We show that our definition of the intensity of implication gives a finer knowledge of gene relations than correlation-based techniques. Our tool can discriminate different phenotypes with a precision comparable to the most performing classifiers.