Motifs Séquentiels Discriminants pour les puces ADN

Paola Salle, Sandra Bringay, Maguelonne Teisseire

MSDAdn

Motifs Séquentiels Discriminants pour les puces ADN

Paola Salle* — **Sandra Bringay*,**** — **Maguelonne Teisseire*,*****

* *LIRMM - Université Montpellier 2 - CNRS*
161 rue Ada, 34392 Montpellier Cedex 5
{salle,bringay,teisseire}@lirmm.fr

** *Departement d'enseignement MIAP - Université Montpellier 3*
*** *Cemagref, UMR TETIS, 500 rue Jean-François Breton, F-34093 Montpellier*

RÉSUMÉ. *Découvrir de nouvelles informations sur les groupes de gènes impliqués dans une maladie est un véritable challenge. Les puces ADN sont des outils puissants pour l'analyse des expressions de gènes. Elles mesurent l'expression de milliers de gènes dans différentes conditions biologiques. Dans cet article, nous proposons une nouvelle approche mettant en évidence des relations d'ordre entre les expressions de gènes. Tout d'abord, nous extrayons des motifs séquentiels qui peuvent être utilisés comme matériel d'étude par les biologistes. Or, comme la densité des bases issues des puces à ADN rend difficile l'extraction de ces motifs, nous introduisons une source de connaissances pendant le processus de fouille. De cette manière, l'espace de recherche est réduit et les résultats obtenus sont plus pertinents d'un point de vue biologique. Les expérimentations sur des données réelles soulignent la pertinence de notre proposition.*

ABSTRACT. *Discovering new information about groups of genes implied in a disease is still challenging. Microarrays are a powerful tool to analyze gene expression. They provide an expression level for genes under given biological situations. In this paper, we propose a novel approach outlining relationships between genes based on their ordered expressions. First, we propose to use a new material, called sequential patterns, to be investigated by biologists. But, due to the expression matrice density, extracting sequential patterns from microarray datasets is far away from easy. Second, we propose to introduce a knowledge source during the mining task. By this way, the search space is reduced and more relevant results (from a biological point of view) are obtained. Results of various experiments on real biological data highlight the relevance of our proposal.*

MOTS-CLÉS : *Fouille de données, Motif séquentiel, Puces ADN, Pathway*

KEYWORDS: *Data mining, Sequential patterns, Gene expression data, Pathway*

1. Introduction

Grâce à l'emergence de nouvelles biotechnologies telles que les puces ADN, il est maintenant possible d'avoir une vue globale de la cellule (Piatetsky-Shapiro *et al.,*, 2003). Les puces ADN sont devenues le moyen le plus utilisé par les biologistes pour comprendre le fonctionnement des gènes. Elles permettent d'étudier le comportement d'un organisme au niveau du génome en mesurant le niveau des expressions des gènes d'une cellule dans un état particulier. Pour mieux comprendre les maladies, les biologistes étudient comment les mécanismes cellulaires contrôlent et régulent les expressions des gènes et dans ce contexte, l'identification des relations entre les expressions des gènes est pertinente.

Découvrir de nouvelles relations d'ordre entre les expressions des gènes, telle que "Le GeneA a une expression inférieure à l'expression des GeneB et GeneC qui ont une expression similaire" sont des corrélations clefs. Ce type de relations peuvent être représentées par des motifs séquentiels. Introduits par (Agrawal *et al.,*, 1995), ils ont été étudiés par la communauté de fouille de données et appliqués sur des bases de données contenant un nombre très élevé de lignes et peu élevé de colonnes. Dans le cas des puces ADN, les données ont une configuration différente. Les bases de données sont très denses car elles contiennent un nombre important de gènes. Chaque gène a une valeur pour chaque puce. Les données sont également très corrélées. Les approches traditionnelles d'extraction de motifs trouvent alors leurs limites. C'est pourquoi nous proposons une nouvelle approche efficace de recherche de motifs permettant aux biologistes de pourvoir comprendre et manipuler des corrélations malgré le gros volume des données manipulées. Tout d'abord, nous introduisons une source de connaissances servant de référence à l'expert et nous utilisons une mesure d'intérêt pour limiter l'espace de recherche.

L'article est organisé de la façon suivante. Section 2 nous présentons les définitions préliminaires associées aux motifs séquentiels. Section 3, nous exposons nos motivations et nos propositions en détaillant l'algorithme MSDAdn (Motifs Séquentiels Discriminants pour les puces ADN). Section 4, nous soulignons la pertinence des résultats obtenus sur des données réelles associées à la maladie d'Alzheimer. Enfin, avant de conclure, Section 5, nous comparons notre approche aux approches existantes.

2. Motifs séquentiels

Les motifs séquentiels ont été introduits par (Agrawal *et al.,*, 1995). Ils correspondent à des séquences d'itemsets (ensembles non vides d'items présents dans une base de données) fréquemment associés sur une période de temps donnée.

Soit BD une base de transactions, chaque transaction est composée de : un identifiant, une date et un ensemble d'items. Une séquence $s = <\text{it}_1\ \text{it}_2 ...\ \text{it}_k >$ est une liste d'itemsets it_i contenant un ensemble non vide d'items. Pour évaluer une telle séquence, nous calculons son support, c'est à dire le pourcentage de clients qui supportent s.

Dans notre contexte, les identifiants sont les puces, les dates sont les expressions des gènes et les items sont des gènes (cf. Tableau 1). La relation proposée est une relation binaire d'ordre strict qui correspond à une relation "est inférieur à". Elle est non réflexive, antisymétrique et transitive.

Exemple 1 *Soit une séquence* $s =< (GeneA)\ (GeneB\ GeneC)\ (GeneD) > [100\%]$*, cette séquence signifie que pour toutes les puces, l'expression de GeneA est strictement inférieure aux expressions de GeneB et GeneC qui sont similaires mais strictement inférieures à l'expression de GeneD.*

Puces (identifiants)	Expressions (date)	Gènes (items)
8	1081,6	GeneA
	1195,5	GeneC
	1358,6	GeneD
	3466,1	GeneB
42	1037,4	GeneD
	1245,2	GeneC
	1824	GeneA
	3346,8	GeneB
43	782,4	GeneD
	1158,2	GeneA
	1379,2	GeneC
	2491,6	GeneB

Tableau 1. *Base de données issues de puces ADN*

L'extraction de motifs séquentiels consiste à chercher toutes les séquences maximales (non incluses dans une autre séquence) dont le support est supérieur à un seuil minimum fixé par l'utilisateur.

La plupart des approches d'extraction de motifs séquentiels se basent sur le principe "Générer-Elaguer" qui est composé de deux phases :

1) Génération : Les k-séquences candidates (avec k la longueur de la séquence) sont générées à partir des (k-1)-séquences fréquentes.

2) Elagage : Pour chaque séquence candidate, nous calculons son support et celles qui ne respectent pas la contrainte de support minimal spécifiée par l'utilisateur sont élaguées.

3. Motivations et proposition

Les puces ADN génèrent une masse d'informations très dense rendant difficile toute extraction de motifs séquentiels avec des méthodes classiques basées sur le principe "Générer-Elaguer". En effet, l'espace de recherche associé est trop grand et les

résultats obtenus sont trop nombreux. Pour y remédier, nous introduisons des connaissances lors de la fouille, pour réduire l'espace de recherche. Pour limiter la génération, nous utilisons une mesure d'intérêt, le coefficient de corrélation de Pearson afin de ne générer que des séquences contenant des gènes corrélés. Finalement, nous cherchons les motifs séquentiels les plus discriminants entre deux classes en ajoutant une contrainte d'élagage.

3.1. *Introduction de connaissances dans la fouille*

Les biologistes étudient les gènes susceptibles d'intervenir dans le développement d'une maladie. Par exemple, A2M est connu pour être impliqué dans la maladie d'Alzheimer. Dans ce contexte, nous proposons de mettre en évidence des liens entre ces gènes connus (gènes "guides") et d'autres gènes encore inconnus. Nous utilisons les gènes "guides" pour orienter l'extraction de motifs séquentiels. Ainsi, nous fouillons un sous-espace de recherche défini à partir de ces gènes "guides". Ce sous-espace est plus petit que l'espace de recherche total. Les motifs séquentiels générés contiennent des gènes guides et d'autres gènes. L' intérêt est double : l'espace de recherche est réduit et les motifs obtenus sont plus pertinents d'un point de vue biologique.

3.2. *Mesure d'intérêt*

Du fait de la densité de la base de données, les séquences candidates sont très nombreuses. Nous utilisons une mesure d'intérêt pour limiter cette génération. Nous avons choisi le coefficient de corrélation de Pearson, souvent employé par les biologistes pour étudier l'existence d'une relation entre deux gènes (Eisen *et al.*,, 1998). Ce calcul permet d'établir si deux variables (X et Y) varient de la même manière. Il est défini comme le rapport de leur covariance et du produit non nul de leurs écarts types.

$$\alpha_{XY} = \frac{Cov(X,Y)}{\sigma x \sigma y}$$

Cov(X,Y) correspond à la covariance des deux variables, σx (resp. σy) correspond à l'écart type (en anglais *standard deviation*) de X (resp. Y) et $-1 \leq \alpha_{XY} \leq 1$. Lorsque le résultat est proche de $\alpha_{XY} = 1$ (resp. $\alpha_{XY} = -1$), la corrélation entre les deux variables est très forte. En revanche, lorsque le résultat est proche de 0 alors les variables sont considérées comme linéairement indépendantes.

3.3. *Utilisation de classes pour les motifs séquentiels*

Les biologistes cherchent des gènes qui ont un comportement différent selon les classes (par exemple malade/sain). Dans notre cas, nous cherchons des motifs séquentiels discriminants, c'est-à-dire spécifiques à une classe.

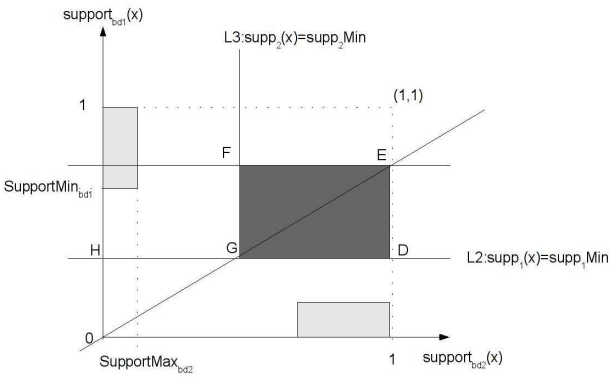

Figure 1. *Motifs séquentiels discriminants*

Pour cela, nous nous sommes inspirés des travaux réalisés sur les "motifs émergents" (EPs) proposés par (Dong *et al.*,, 1999). Ces auteurs ont introduit le calcul d'un taux d'accroissement pour un motif d'une base de données BD_1 vers une base de données BD_2 (*growth rate*). Ce taux correspond au support du motif dans BD2 divisé par son support du motif dans BD_1. Si ce taux est supérieur à un seuil minimal alors le motif est dit émergent de la base BD_1 vers la base BD_2 (le motif caractérise BD_2 par rapport à BD_1). Les cas recherchés sont illustrés Figure 1 en gris foncé.

Dans notre contexte, nous recherchons des motifs séquentiels qui ont un support élevé dans BD_1 et faible dans BD_2. Ces motifs sont caractéristiques de BD_1 par rapport à BD_2. Pour cela, nous avons défini deux types de supports :
supportMin$_{BD1}$ et supportMax$_{BD2}$.
Si le support d'une séquence s est supérieur (ou égal) au supportMin$_{BD1}$ mais inférieur (ou égal) au supportMax$_{BD2}$ alors cette séquence est considérée comme fréquente dans BD_1 et discriminante entre BD_1 et BD_2 (support(s)$_{BD1} \geq$ supportMin$_{BD1}$ et support(s)$_{BD2} \leq$ supportMax$_{BD2}$). Les cas recherchés sont illustrés Figure 1 en gris clair.

Exemple 2 *Considérons le tableau 2. Les puces 1,2,4 correspondent à des sujets jeunes et la puce 3 correspond à un sujet âgé. Nous définissons le supportMin = 2 et le supportMax = 0. Soit une séquence s= <(3 4)(2)> avec un support$_{BD1}$(s)=3 $\geq$ supportMin et un support$_{BD2}$(s)=0 $\leq$ supportMax alors s est considérée comme frequente dans BD_1 et discriminante entre BD_1 et BD_2.*

3.4. *Algorithme MSDAdn : Description*

3.4.1. *Principe*

L'algorithme que nous proposons repose sur le principe "Générer-Elaguer". Pour réduire l'espace de recherche et pour assurer le passage à l'échelle, les 2-séquences candidates sont générées à partir des gènes guides (items fréquents et sélectionnés grâce à la source de connaissance). Puis, lors des itérations suivantes, nous étendons les (k-1)-séquences (avec k la longueur de la séquence et k>2) fréquentes en utilisant les autres gènes présents (items fréquents et non sélectionnés grâce aux connaissances) dans la base. Pour étendre ces séquences, nous adoptons le principe des extensions proposées par (Wang *et al.,*, 2004).

Extensions : Nous avons défini quatre extensions :

– "Avant" (respectivement Après) : un item fréquent est ajouté **Avant** (respectivement **Après**) les itemsets de la sous-séquence,

– "Entre" : un item fréquent est ajouté **Entre** les itemsets de la sous-séquence,

– "Dans" : un item fréquent est ajouté **Dans** un itemset de la sous-séquence.

Exemple 3 *Soit deux séquences fréquentes composées de gènes guides (items fréquents et sélectionnés grâce à la source de connaissances) s = <(1)(2)> et s'= <(3)(4)> et un item fréquent "5" de la base de données. En étendant s et s' avec les extensions **Après** et **Dans**, nous obtenons les candidats : {<(1)(2)(5)>, <(1 5)(2)>, <(1)(2 5)>, <(3)(4)(5)> , <(3 5)(4)>, <(3)(4 5)>}. En revanche, nous avons perdu les candidats : {<(5)(1)(2)>, <(1)(5)(2)>, <(5)(3)(4)>, <(3)(5)(4)>}. Pour lever cette limitation, chaque séquence fréquente est étendue en utilisant les quatre extensions.*

Corrélation de Pearson : Lors de la génération des k-séquences candidates à partir des (k-1)-séquences fréquentes, le coefficient de corrélation de Pearson permet de limiter le nombre de candidats. Plus précisemment, nous choisissons l'item utilisé pour étendre une séquence en fonction de sa corrélation avec les items déjà présents dans cette séquence. L'intérêt est d'obtenir des séquences fréquentes contenant uniquement des gènes corrélés selon un seuil spécifié par l'expert.

Exemple 4 *Soit s = <(1)(2)> une séquence fréquente et un item fréquent "5" de la base de données. Nous mesurons le coefficient de corrélation de Pearson entre l'item 2 et l'item 5. Si $\alpha_{25} \geq$ PearsonMin, nous étendons la séquence candidate selon les 4 extensions (e.g. avec l'extension après : s'=<(1)(2)(5)>), sinon nous ne générons pas de nouvelle séquence candidate.*

Classes : Lors de la phase d'élagage, afin de ne conserver que des séquences discriminantes, nous avons introduit la notion de classe et deux seuils : supportMin et supportMax. Nous ne gardons que les séquences qui sont très fréquentes dans une classe et non fréquentes dans une autre classe, c'est-à-dire toutes les séquences s tel que $support_{classe1}(s) \geq supportMin$ et $support_{classe2}(s) \leq supportMax$.

Le processus est réitéré de la même manière, c'est à dire, toutes les (k-1)-séquences fréquentes sont étendues en utilisant les quatre extensions définies, jusqu'à qu'il n'y ait plus de séquences candidates.

3.4.2. *MSDAdn : Pas à Pas*

Dans la suite de cette section, nous utilisons les notations présentées tableaux 3 et 4. Nous détaillons les étapes de l'algorithme MSDAdn et nous les illustrons par un exemple simple en utilisant la base de données BD de la Table 2. Pour les deux première itérations, nous n'utilisons que la contrainte d'élagage classique sur le support dans BD. A partir de l'itération 3, nous utilisons la contrainte d'élagage définie selon les classes dans les bases BD_1 et BD_2. Cette base est découpée en deux bases $BD_1 = \{1, 2, 4\}$ et $BD_2 = \{3\}$ correspondant aux deux classes.

Puces	Expressions	Genes	Classe
1	2	1 3 4	jeune
	4	2	
2	2	1 3 4	jeune
	4	2	
3	1	4	âgé
	2	1 3	
	3	2	
4	1	1	jeune
	2	3 4	
	4	2	

Tableau 2. *Exemple d'une base de données (BD = $BD_1 \cup BD_2$)*

Première itération : Soit L_1^{guides} = $BD \cap L^{guides}$ l'ensemble de tous les gènes guides (items fréquents) en provenance d'une source de connaissances et qui sont présents dans la base de données BD avec $BD = BD_1 \cup BD_2$. Les autres gènes de BD sont ajoutés dans $L_1^{BD-L^{guides}}$. Il n'est pas nécessaire de calculer le support de

$<it_1\ it_2 ...\ it_k>_{support}$: une séquence avec son support
L^{guides} : l'ensemble des gènes guides
L_1^{guides} : l'ensemble des items fréquents présents dans la source de connaissance et dans la BD.
$L_1^{BD_1-L_1^{guides}}$: l'ensemble des items fréquents dans BD_1 - L_1^{guides}
n-seqCand : l'ensemble des n-séquences candidates
n-seqFreq : l'ensemble des n-séquences fréquentes
L^{freq} : l'ensemble de toutes les séquences fréquentes maximales de BD

Tableau 3. *Notations utilisées*

$\bowtie$: opérateur de jointure classique
$\rhd$: nous ajoutons chaque item i $\in \mathrm{L}_1^{BD_1 - L_1^{guides}}$ dans chaque itemset de chaque séquence de (j-1)-seqFreq
$\lhd$: nous ajoutons chaque item i $\in \mathrm{L}_1^{BD_1 - L_1^{guides}}$ entre chaque itemset de chaque séquence de (j-1)-seqFreq

Tableau 4. *Opérateurs de jointure*

chaque item car ils sont tous fréquents. En effet, l'une des spécificités de nos données est que pour chaque puce, chaque gène (item) s'exprime une fois. Par conséquent, le support de chaque item est égal au nombre total de puces. $\forall i \in \mathrm{L}_1^{guides}$, $\forall i \in \mathrm{L}_1^{BD - L^{guides}}$ $support(i) = |nombre\ de\ puces\ total|$.

Exemple 5 *Soient* L^{guides} *={2,3,5,6} et* BD*={1,2,3,4},* L_1^{guides} *= {<2>$_3$, <3>$_3$} et* $L_1^{BD-L^{guides}}$ *= {<1>$_3$, <4>$_3$}*

Deuxième itération : Pour obtenir l'ensemble 2-seqCand, nous réalisons une jointure classique ($\mathrm{L}_1^{guides} \bowtie \mathrm{L}_1^{guides}$). Puis, pour chaque séquence candidate s $\in$ 2-seqCand, nous calculons son support dans BD. Toutes les séquences de 2-seqCand vérifiant la contrainte du supportMin sont ajoutées dans 2-seqFreq.

Exemple 6 *Soit 2-seqCand = {<(2)(3)>, <(2 3)>, <(3)(2)>}*
Avec un supportMin = $2/3$ *pour* BD*, nous obtenons : 2-seqFreq = {<(3)(2)>$_3$>}. Les séquences fréquentes de longueurs 2 contiennent uniquement des items de* L_1^{guides}*.*

Troisième itération : Pour obtenir les 3-seqCand, nous étendons les 2-seqFreq avec des items de $\mathrm{L}_1^{BD-L^{guides}}$ selon les quatre extensions définies précédemment. Pour chaque item i $\in \mathrm{L}_1^{BD-L^{guides}}$, nous générons une nouvelle séquence candidate en ajoutant l'item i avant (resp. après) chaque séquence s $\in$ 2-seqFreq (extension avant et après). Pour étendre s avec l'extension "Entre", nous créons un nouvel itemset contenant l'item i que nous ajoutons entre les 2 itemsets de s. Pour étendre s avec l'extension "Dans", nous ajoutons l'item i dans chaque itemset de s et pour chaque ajout, une séquence candidate est générée. Finalement, nous ajoutons toutes ces séquences candidates dans 3-seqCand. Pour chaque séquence candidate s $\in$ 3-seqCand, nous calculons deux supports de s dans BD_1 et BD_2. Toutes les 3-seqCand vérifiant la contrainte du supportMin dans BD_1 et du supportMax dans BD_2 sont ajoutées dans 3-seqFreq.

Exemple 7 *Après la seconde itération, nous avons une seule séquence fréquente de longueur 2, s = {<(3)(2)>$_3$} $\in$ 2-seqFreq. Il existe 4 extensions possibles avec*

Algorithme 1 MSDAdn

ENTRÉES: BD_1 la base de données d'une classe, BD_2 la base de données de la classe complémentaire, L^{guides} les gènes guides, $supportMin$ le seuil minimal du support dans BD_1 et $supportMax$ le seuil maximum de support dans BD_2, $PearsonMin$ le seuil utilisé pour la corrélation des gènes

SORTIES: Ensemble des séquences fréquentes maximales de BD_1

// Première itération

$L_1^{guides} \leftarrow$ tous les items i $\in L^{guides}$

$L_1^{BD_1-L^{guides}} \leftarrow$ tous les items j $\in (BD - L_1^{guides})$

// Seconde itération

2-seqCand $\leftarrow$ candidats de 2-sequences de $L_1^{guides} \bowtie L_1^{guides}$

pour chaque ($candidat \in$ 2-seqCand) **faire**

  si ($supp(candidat) \geqslant suppMin$) **alors**

    2-seqFreq $\leftarrow$ 2-seqFreq + $candidat$;

  fin si

fin pour

// j^{eme} itération

tantque ((j-1)-seqFreq != $\varnothing$) **faire**

  si (correlation($(j-1) - seqFreq$, candidat de $L_1^{BD_1-L^{guides}}$) $\geqslant PearsonMin$) **alors**

    j++ ;

    j-seqAvant $\leftarrow$ candidats de $L_1^{BD_1-L^{guides}} \bowtie$ (j-1)-seqFreq ;

    j-seqAprès $\leftarrow$ candidats de (j-1)-seqFreq $\bowtie L_1^{BD_1-L^{guides}}$;

    j-seqEntre $\leftarrow$ candidats de (j-1)-seqFreq $\lhd L_1^{BD_1-L^{guides}}$;

    j-seqDans $\leftarrow$ candidats de (j-1)-seqFreq $\rhd L_1^{BD_1-L^{guides}}$;

    j-seqCand $\leftarrow$ j-seqAvant $\cup$ j-seqAprès $\cup$ j-seqEntre $\cup$ j-seqDans ;

    pour chaque ($candidat \in$ j-seqCand) **faire**

      si ($supp_{BD1}(candidat) \geqslant supportMin$ et $supportMax \geqslant supp_{BD2}(candidat)$) **alors**

        j-seqFreq $\leftarrow$ j-seqFreq + $candidat$;

      fin si

    fin pour

  fin si

fin tantque

$L^{freq} \leftarrow L^{freq} \cup$ {séquences fréquentes maximales dans j-seqFreq}

les items de $L_1^{BD-L^{guides}}$ = {1,4}. Les séquences candidates sont illustrées dans le tableau 5.

Si l'on utilise uniquement la contrainte classique sur le support, supportMin = $2/3$ dans BD_1, nous obtenons deux séquences fréquentes : 3-seqFreq = {<(1 3)(2)>$_2$, <(3 4)(2)>$_3$}. Si nous prenons en compte les classes avec la deuxième contrainte sur le

Extensions	Items	
	$<1>_3$	$<4>_3$
Avant	$<(1)(3)(2)>_1$	$<(4)(3)(2)>_0$
Après	$<(3)(2)(1)>_0$	$<(3)(2)(4)>_0$
Entre	$<(3)(1)(2)>_0$	$<(3)(4)(2)>_0$
Dans	$<(1\ 3)(2)>_2$ $<(3)(1\ 2)>_0$	$<(3\ 4)(2)>_3$ $<(3)(4\ 2)>_0$

Tableau 5. *Les extensions*

supportMax = 0 *dans* BD_2*, nous obtenons une seule séquence fréquente : 3-seqFreq = {<(3 4)(2)>$_2$}. Ainsi, la recherche de séquences fréquentes discriminantes permet d'élaguer des séquences fréquentes.*

Itérations suivantes : nous réitérons le processus jusqu'à ne plus obtenir de séquences candidates.

4. Expérimentations

4.1. *Contexte de l'étude*

La maladie d'Alzheimer est une maladie neurodégénérative qui entraîne la perte progressive et irréversible des fonctions mentales. La cause exacte de cette maladie est encore inconnue, mais les biologistes suspectent des facteurs environnementaux et génétiques. Dans le cadre du projet Gene-Mining, nous travaillons avec des biologistes du laboratoire MMDN[1] qui étudient la maladie d'Alzheimer. Ils travaillent avec une population de lémuriens Microcebus murinus (ou microcébe murin). Ces primates sont phylogénétiquement proches de l'homme et très souvent utilisés dans les études portant sur le vieillissement ou la maladie d'Alzheimer. Pour étudier les expressions des gènes dans des cellules du cortex temporal, les biologistes utilisent des puces ADN Affymetrix U133 plus 2.0. Les analyses via ces puces ADN ont été réalisées pour 14 animaux : 5 jeunes, 7 âgés et 2 porteurs de signes histologiques de la maladie d'Alzheimer. Chaque puce mesure l'intensité de 19 653 gènes.

4.2. *Mise en oeuvre*

Classes : Pour les expérimentations, nous avons cherché des motifs séquentiels caractéristiques de la classe "jeune" (5 sujets) par rapport à la classe "âgé" (9 sujets).

1. "Molecular mechanisms in neurodegenerative dementias" de l'Université de Montpellier 2 www.mmdn.univ-montp2.fr

Sélection des données : nous avons choisi de réaliser nos expérimentations sur des gènes qui ont une différence d'expression significative entre les deux classes. Nous avons sélectionné 579 gènes en réalisant une analyse SAM (V. Tusher *et al.*,, 2001) puis une analyse de la variance (ANOVA (Kerr *et al.*,, 2000)).

Choix de la source de connaissances du domaine : Nous avons choisi comme gènes guides, les gènes du pathway de la maladie d'Alzheimer issus de la base KEGG[2]. Le nombre total de gènes guides (présents dans la base de données et dans le pathway) est de 7.

4.3. *Protocole*

Pour réaliser les expérimentations, nous avons comparé l'algorithme MSDAdn à l'algorithme PSP (Masseglia *et al.*,, 1998) d'extraction de motifs séquentiels. Toutes les expérimentations ont été réalisées sur un PC avec I Intel(R) Pentium(R) 4 CPU 3.00GHz et 2GB de RAM. L'algorithme MSDAdn est codé en C++.

Nous avons défini expérimentalement un seuil ρ sur le nombre maximal de séquences générées. Lorsque le nombre de (k-1)-séquences fréquentes (avec k la longueur de la séquence) dépasse ce seuil ($|$(k-1)-seqFreq$| < \rho$) alors nous générons les k-séquences candidates uniquement en utilisant les items représentant les gènes guides (L_1^{guides}). En faisant cela, nous réduisons de manière ponctuelle l'espace de recherche.

Nous avons fixé deux paramètres :

– supportMax qui est la contrainte du support maximal dans la base de données des "âgés" (supportMax=0),

– maxFreqSeq qui est la contrainte sur le nombre de séquences générées.

Nous avons fait varier deux paramètres :

– supportMin qui est la contrainte classique du support minimal dans la base de données des "jeunes". Nous avons fait varier ce paramètre de 1 à 0.4.

– PearsonMin qui est la contrainte que doivent respecter deux gènes. Nous avons fait varier ce paramètre sur deux valeurs 0.6 et 0.9.

Finalement, nous avons testé notre approche : sans le seuil maxFreqSeq et sans la contrainte PearsonMin ; avec le seuil maxFreqSeq mais sans la contrainte Pearson-

2. "Pathway" est un terme générique correspondant à une représentation graphique des différentes interactions entre les gènes. Cette représentation est utilisée par les experts pour modéliser les relations existantes entre les gènes impliqués dans une maladie. http ://www.genome.jp/kegg/

supportMin	PearsonMin	longueur	nombre (motifs séquentiels)	nombre de gènes guides dans un motif séquentiel
1	0.6	6	1763	2
	0.9	6	100	2
	/	6	1778	2
0.8	0.6	6	20304	3
	0.9	7	663	2
	/	6	24525	3
0.6	0.6	7	80953	4
	0.9	9	29722	3
	/	7	83618	4
0.4	0.6	6	185240	4
	0.9	8	47764	4
	/	6	188017	4

Tableau 6. *Résultats obtenus en faisant varier SupportMin et PearsonMin*

Min ; avec le seuil maxFreqSeq et avec la contrainte PearsonMin = 0.6 ; et enfin avec le seuil maxFreqSeq et avec la contrainte PearsonMin = 0.9.

4.4. *Résultats*

Les résultats de nos expérimentations montrent que PSP ne permet pas d'extraire des motifs séquentiels à partir de ces données, et ceci même avec supportMin=1 alors que le nombre de motifs est censé être le plus faible. Concernant les expérimentations avec notre algorithme MSDAdn, les figures montrent le temps d'exécution (Fig.2) et le nombre de motifs séquentiels discriminants obtenus (Fig.3) lorsque nous faisons varier supportMin. Lorsque supportMin = 1.0 et supportMax = 0, nous obtenons 1778 motifs séquentiels discriminants. Si nous ne prenons pas en compte le seuil maxFreqSeq, le nombre de séquences fréquentes obtenues est alors trop important. Il n'est donc pas possible de faire varier le supportMin. Par contre, nous constatons sensiblement les mêmes résultats lorsque nous utilisons ou non la corrélation de Pearson (PearsonMin = 0.6). Ceci s'explique car les données de base sont corrélées. En revanche, lorsque nous fixons PearsonMin = 0.9, le temps d'exécution est inférieur et les résultats obtenus sont nettement moins nombreux. Le Tableau 6 résume les résultats en faisant varier les différents paramètres. Nous constatons que la longueur des motifs séquentiels et le nombre de gènes guides obtenus croissent lorsque supportMin diminue.

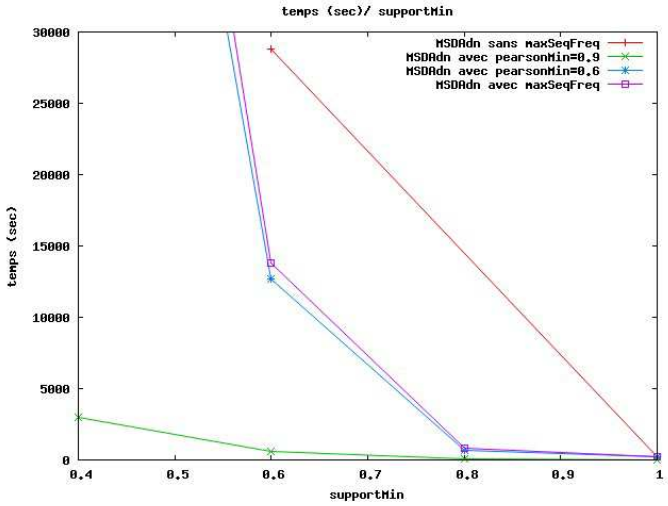

Figure 2. *Temps d'exécution en fonction de SupportMin*

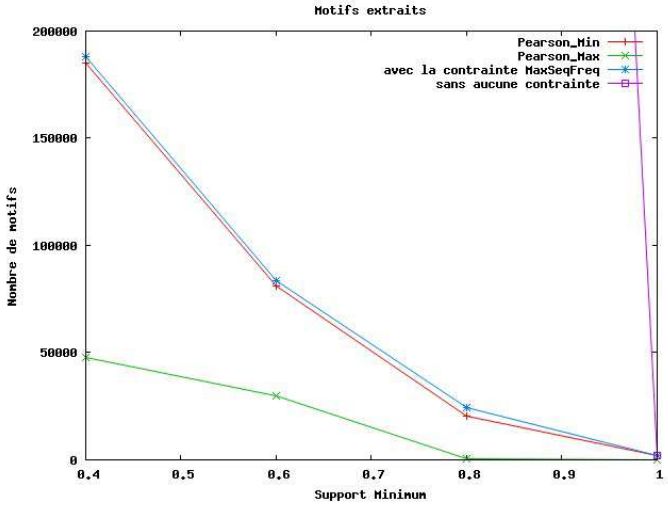

Figure 3. *Nombre de motifs séquentiels discriminants en fonction de SupportMin*

5. Etat de l'Art

Des méthodes statistiques ont été proposées pour comparer les expressions de gènes entre plusieurs conditions biologiques dans le but de sélectionner ceux qui ont une différence significative d'expression. Ainsi, en général, les biologistes réalisent une analyse SAM (V. Tusher *et al.,*, 2001) sur leurs données. Puis, pour diminuer le nombre de faux positifs, ils complètent leur première analyse par une analyse de la variance (ANOVA) (Kerr *et al.,*, 2000). Plusieurs méthodes d'Intelligence Artificielle ont été proposées pour sélectionner des gènes discriminants grâce à une classification et une phase d'apprentissage. Parmi ces méthodes, nous trouvons les réseaux de neurones (Khan *et al.,*, 2001), les K-plus-proches-voisins (Li *et al.,*, 2001), les support vector machines (Brown *et al.,*, 2000) et l'analyse discriminante linéaire (LDA) (Dudoit *et al.,*, 2000). Les méthodes de clustering ont également été utilisées pour identifier des goupes de gènes qui s'expriment de la même manière. Plusieurs approches, essentiellement du clustering hiérarchique, ont été proposées (Eisen *et al.,*, 1998), carte de Kohonen (en anglais self organizing map) (Tamayo,, 1999) ou encore des graphes théoriques (Hartuv *et al.,*, 2000). (Eisen *et al.,*, 1998) proposent un clustering hiérarchique sur les données préalablement discrétisées selon la distinction "sur et sous-exprimé". Puis, ils regroupent les gènes selon une mesure de similarité.

Basé sur les des règles d'association, (Pan *et al.,*, 2003), (Rioult *et al.,*, 2003) sont les premiers à proposer un algorithme adapté à la configuration particulière des bases de données de puces ADN. Ils proposent d'extraire des motifs fermés en réalisant une énumération non pas sur les colonnes (gènes) mais plutôt sur les lignes (puces). (Xu *et al.,*, 2004), (Cong *et al.,*, 2004) catégorisent des règles d'associations dans des groupes de règles du type $R = GeneA, GeneB \rightarrow Cancer$. (Pensa *et al.,*, 2004) réalisent une extraction sous contraintes en utilisant des propriétés sur les gènes. Pour cela, ils font appel à une source de connaissances extérieure Gene Ontology. (Tanasa *et al.,*, 2004) proposent également de transposer la base de données. Ensuite, ils discrétisent les valeurs selon trois degrés d'expression. Finalement, ils cherchent des séquences contenants des niveaux d'expressions ordonnés selon des conditions biologiques et qui sont supportées par une grande majorité de gènes de la base. Ces approches prennent en compte des expressions de gènes discrétisées (2 ou 3 degrés d'expressions).

L'originalité de notre approche est d'extraire des motifs séquentiels qui mettent en avant des relations d'ordre d'expression entre les gènes. Et la recherche de motifs séquentiels discriminants telle que nous la proposons permet d'identifier un groupe de gènes ordonnés selon leurs expressions spécifiques à une condition biologique. Pour les biologistes, la recherche de motifs séquentiels discriminants est plus pertinente que la recherche de motifs séquentiels classique. En effet, savoir qu'une séquence s = <(GeneA) (GeneB) (GeneC)> est fréquente dans une classe (chez tous les jeunes) et non fréquente dans l'autre classe (chez tous les âgés) est plus pertinent pour l'expert que savoir qu'une séquence s'= <(GeneB) (GeneA) (GeneC)> est fréquente dans la base de données. Pour aider les biologistes à analyser ce nouveau matériel d'étude, nous avons proposé et développé un outil de visualisation facilitant la découverte de nouveaux gènes potentiellement impliqués dans la maladie (Salle *et al.,*, 2009). Grâce

à notre approche et à cet outil, ils peuvent exploiter de nouvelles pistes de recherche sur les interactions entre les gènes.

6. Conclusions et perspectives

Dans cet article, nous avons proposé une nouvelle méthode pour analyser les données issues des puces ADN. Cette méthode consiste à chercher des motifs séquentiels discriminants. Dans ce contexte, nous avons proposé un algorithme MSDAdn prenant en compte la spécificité des données. En effet, grâce à l'intégration de gènes guides, l'espace de recherche est réduit et les résultats obtenus sont plus pertinents pour l'expert. Ces gènes peuvent provenir de représentations de connaissances formelles ou informelles : proposition de l'expert, connaissances consensuelles du domaine comme les pathways ou les ontologies. Aussi, pour réduire le nombre de résultats, nous avons proposé d'utiliser une mesure d'intérêt. Pour les biologistes, ce type de motifs séquentiels discriminants met en évidence de nouvelles corrélations entre gènes qui sont des pistes de recherche permettant de compléter des pathways existants. Concernant l'analyse des motifs obtenus, nous avons développé un outil de visualisation afin de rendre accessible les corrélations pertinentes aux biologistes. Afin d'améliorer les performances de MSDAdn, nous envisageons maintenant de comparer l'approche actuelle à une approche hybride mélant à la fois les avantages d'une génération par niveau puis par profondeur. Nous souhaitons également intégrer les notions de la théorie des sous ensembles flous pour proposer des motifs plus informatifs du type <a *proche de* b *très surexprimé* c>. Pour finir, quelque soit la méthode choisie, le nombre de motifs obtenus reste très élevé, d'autres techniques de visualisation doivent donc accompagner ces propositions pour faciliter l'interprétation par les experts.

7. Bibliographie

Agrawal R., Srikant R., « Mining sequential patterns », *in* , I. C. S. Press. (ed.), *Eleventh International Conference on Data Engineering*, p. 3-14., 1995.

Brown M. P., Grundy W. N., Lin D., Cristianini N., Sugnet C. W., Furey T. S., Ares M., Haussler D., « Knowledge-based analysis of microarray gene expression data by using support vector machines », vol. 97, p. 262-267, January, 2000.

Cong G. A., Tung X., F.Pan, J.Yang., « Farmer : Finding interesting rule groups in microarray datasets. », *in* , ACM (ed.), *SIGMOD Conference*, p. 143-154, 2004.

Dong G., Li J., « Efficient mining of discriminant patterns : discovering trends and differences », *in* , ACM (ed.), *KDD '99 : Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 43-52, 1999.

Dudoit S., andJ. F., Speed T. P., Comparison of discrimination methods for the classification of tumors using gene expression data., Technical report, Department of Statistics, University of California, Berkeley ., 2000.

Eisen M., P.Spellman, P.Brown, D.Botstein, « Cluster analysis and display of genome-wide expression patterns. », *Proceedings of the National Academy of Science*, p. 14863-14868, 1998.

Hartuv E., Schmitt A., Lange J., Meier-Ewert S., Lehrach H., Shamir R., « An algorithm for clustering cDNA fingerprints. », *Genomics*, p. 249-256, 2000.

Kerr M. K., Martin M., Churchill G. A., « Analysis of Variance for Gene Expression microarray Data », *Journal of Computational Biology*, vol. volume 7, p. 819-837, 2000.

Khan J., Wei J. S., Ringner M., Saal L. H., Ladanyi M., Westermann F., Berthold F., Schwab M., Antonescu C. R., Peterson C., Meltzer P. S., « Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks », *Nature Medicine*, p. 673-679, 2001.

Li L., Weinberg C. R., Darden T. A., Pedersen L. G., « Gene selection for sample classification based on gene expression data : study of sensitivity to choice of parameters of the GA/KNN method. », *Bioinformatics,*, p. 1131-1142, 2001.

Masseglia F., Cathala F., Poncelet P., « The PSP Approach for Mining Sequential Patterns », *PKDD '98 : Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery*, p. 176-184, 1998.

Pan F., G.Cong, Tung A., J.Yang, Zaki. M., « Carpenter : finding closed patterns in long biological datasets., », *in* , P. L.Getoor, T.E. Senator, , C. (Eds.) (eds), *KDD*, ACM, p. 637-642, 2003.

Pensa R. G., Besson J., Boulicaut J.-F., « A methodology for biologically relevant pattern discovery from gene expression data. », *in* , E. Suzuki, , S. A. (Eds.) (eds), *Discovery Science*, vol. 3245, Lecture Notes in Computer Science, Springer, p. 230-241, 2004.

Piatetsky-Shapiro G., Tamayo P., « Microarray data mining : facing the challenges », *SIGKDD Explor. Newsletter.*, vol. vol 5, 2003.

Rioult F., Boulicaut J.-F., Crémilleux B., Besson J., « Using transposition for pattern discovery from microarray data. », *8th ACM SIGMOD Workshop on research issues in Data Mining and Knowledge Discovery DMKD'03*, p. 73-79, 2003.

Salle P., Bringay S., Teisseire M., « Mining Discriminant Sequential Patterns for Aging Brain », *AIME '09 : Proceedings of the 12th conference on Artificial Intelligence in Medicine*, 2009.

Tamayo P., « Interpreting patterns of gene expression with self-organizing map : Methods and application to hematopoietic differentiation. », *Proc. of the National Academy of Sciences of the United States of America*, p. 2907-2912, 1999.

Tanasa D., Lopez J. A., Trousse B., « Extracting Sequential Patterns for Gene Regulatory Expressions Profiles », *Knowledge Exploration in Life Science Informatics, International Symposium,KELSI*, p. 46-57, 2004.

V. Tusher R. T., Chu C., « Significance Analysis of microarrays Applied to Ionizing Radiation Response », *Proceedings of the National Academy of Science*, vol. 98, p. 5116-5121, 2001.

Wang J., Han J., « BIDE : Efficient Mining of Frequent Closed Sequences. », *ICDE '04 : Proceedings of the 20th International Conference on Data Engineering*, IEEE Computer Society, p. 79-90, 2004.

Xu X., G.Cong, Ooi B., Tan K.-L., Tung A., « Semantic mining and analysis of gene experession data », *Proceedings 2004 VLDB Conference,*, p. 1261-1264, 2004.