Tendances dans les expressions de gènes : Application à l'analyse du transcriptome de *Plasmodium Falciparum* [1]

Philippe Collet, Vincent Derozier, Gérard Dray,
François Trousset, Pascal Poncelet, Michel Crampes

EMA/LGI2P
Ecole des Mines d'Alès
Site EERIE, Parc Scientifique Georges Besse
30035 Nîmes, cedex 1, France
{Prénom.Nom}@ema.fr

Résumé. L'étude de l'expression des gènes est depuis quelques années révolutionnée par la technique des puces à ADN (*DNA arrays*). Cette méthode offre de nombreux avantages par rapport aux méthodes usuelles (Northern-blotting, quantitative RT-PCR, real-time RT-PCR) comme la mesure quantitative de l'expression d'un très grand nombre de gènes de façon simultanée sur des micro volumes d'échantillons. La masse d'information fournie par les scanners de puce à ADN impose un traitement informatique spécifique. Les méthodes habituellement mises en œuvre pour analyser ces données s'appuient sur des algorithmes de partitionnement, comme les clustering hiérarchiques, et sur une hypothèse communément admise qui associe à un ensemble de profils d'expression similaires, une fonction identique. Cette analyse classique étudie l'ensemble des gènes sans distinction. L'approche que nous proposons permet de prendre en compte la quantité d'informations reliées aux gènes et de les diviser en deux catégories : gènes connus ou putatifs.
Pour chaque gène n'ayant pas d'information rattachée, nous étudions son voisinage afin d'y trouver des motifs fréquents (itemsets). Ensuite, l'Analyse est guidée par l'interprétation biologique afin de faire émerger des propriétés intéressantes.
Un premier jeu de test sur *P. Falciparum* nous a permis de mettre en évidence, en nous intéressant aux items relatifs à la glycolyse, un transporteur de nucléosides qui intervient au niveau énergétique dans la phase ring (précoce) du parasite.

[1] Ce travail est supporté par un projet pluridisciplinaire GEMBIO du groupement des Ecoles Des Mines dans le domaine de la Bio-informatique.

1 Introduction

L'étude de l'expression des gènes est depuis quelques années révolutionnée par l'approche des puces à ADN (*DNA arrays*). Cette méthode offre de nombreux avantages par rapport aux méthodes usuelles (Northern-blotting, quantitative RT-PCR, real-time RT-PCR) : la possibilité de quantifier l'expression de plusieurs milliers de gène simultanément permet d'obtenir un « instantané » du transcriptome de la cellule. Un avantage que nous n'avons pas encore exploité (mais qui est à considéré pour nos futurs travaux) et d'utiliser l'information contenue dans les puces à ADN afin d'en étudier les relations potentielles entre les différents gènes ainsi que l'aspect coordination de leurs comportements [LiYa04]. D'autre part la miniaturisation des volumes d'échantillon utilisés (quelques µL) permet de travailler sur des densités d'échantillon très importante (de l'ordre de plusieurs dizaines de milliers de dépôts par lame) et ainsi de pouvoir observer un génome complet.

L'apport de cette technologie à l'étude de la malaria est intéressant. En effet, le génome de *Plasmodium* étant connu depuis peu [HofSub02], l'entière connaissance fonctionnelle des gènes fait défaut. A l'heure actuelle seul 40% des gènes se sont vu attribuer une fonction (les 60% restant sont souvent appelés « gènes hypothétiques » ou putatifs).

Nous avons donc entrepris, à partir d'une étude de puces à ADN préalablement réalisée [MaGl01] d'extraire de la connaissance à partir d'un tableau d'expression de gènes, afin de pouvoir attribuer une fonction aux gènes hypothétiques. Le tableau d'expression est composé d'un ensemble de 943 gènes auxquels une condition temporelle est affectée.

A l'heure actuelle, pour analyser les puces, des techniques de clustering sont utilisées [JiPe04, MaOl04], par exemple, pour :

- Regrouper des gènes selon leur expression en fonction de différentes conditions ;
- Regrouper des conditions expérimentales en fonction de leurs profils d'expression sur chaque gène ;
- Déterminer la fonction de gènes putatifs grâce à l'expression de gènes connus déjà réunis en clusters.

Récemment, la notion de *bicluster* a été introduite dans [ChCh00] et dont un état de l'art complet est proposé dans [MaOl04], consiste à rechercher des clusters qui ne soient pas uniquement sur les lignes ou les colonnes de la matrice d'expression mais plutôt qui tienne compte des deux dimensions. Dans ce contexte, il devient alors possible de rechercher par exemple des sous ensembles de gènes qui exhibent des motifs d'expressions similaires au sein de la matrice, i.e. des expressions discrètes et temporelles. De la même manière dans [AgPe04], les auteurs s'intéressent à différents types de règles qui peuvent être obtenus par rapport aux données d'expressions. Dans ce cadre, ils proposent des règles de sémantiques différentes.

Cependant l'utilisation des méthodes de clustering ne fait pas la différence entre les gènes porteurs d'informations (gènes connus) et ceux non porteurs d'informations (gènes putatifs) et traitent l'ensemble des gènes de façon similaire. La distinction que nous introduisons permet d'obtenir des regroupements plus pertinents puisque basés sur le contenu informatif des gènes. Ainsi nous pouvons pour les gènes putatifs, suivant une hypothèse biologique

donnée, obtenir les items les plus fréquents dans un voisinage informatif (de gènes connus). L'avantage de cette démarche consiste donc à guider le processus par la connaissance a priori d'un expert biologiste.

L'article est organisé de la manière suivante. La section 2 présente plus formellement la problématique étudiée. La section 3 décrit l'approche proposée pour rechercher les tendances dans les expressions des gènes. La section 4 présente les différentes expériences menées avec différents jeux de données. Enfin, en conclusion, nous présentons les perspectives de ce travail.

2 Problématique

Soit $E = \{e_1, e_2, \ldots, e_n\}$ l'ensemble des conditions expérimentales temporelles.
Soit $G = G_{Known} \grave{E} G_{Unknown}$ l'ensemble des gènes où G_{Known} représente l'ensemble des gènes connus et $G_{Unknown}$ représente l'ensemble des gènes hypothétiques ou putatifs.
Soit $F = G \ X \ E \ ® \hat{A}$, une fonction qui représente le niveau d'expression d'un gène pour une condition donnée (i.e. dans notre cas à un instant donné).
Soit $M\ (G, E)$ une matrice d'expression de gènes où chaque colonne correspond à un condition expérimentale et où chaque ligne correspond à un gène. Chaque élément de M prend ses valeurs dans F. La figure 1 représente un exemple de données d'expression de gènes.

Nom du gène	e1	e2	e3	e4	e5
n98138	5,60	4,11	0,77	1,03	0,48
t02493.1	5,19	4,11	1,12	1,52	0,69
t02496.1	5,15	4,42	1,07	1,09	0,52
t02499.1	4,07	5,20	0,84	0,90	0,24
n98171	0,72	17,30	0,96	0,48	0,26
n98196	1,88	2,21	1,28	2,18	0,16

Fig. 1 – Exemple d'expressions de gènes

La problématique de l'analyse de tendances dans des données d'expression consiste à proposer une approche d'analyse temporelle du comportement de gènes hypothétiques ou putatifs par rapport à des gènes connus. En d'autres termes, nous recherchons pour chaque gène $gp_i \ \hat{I} \ G_{Unknown}$, quels sont ceux pour lesquels nous possédons le plus d'information, *i.e.* quels sont les gènes connus ayant les comportements les plus similaires au cours du temps.

3 Analyse de tendances

Dans cette section, nous présentons l'approche que nous avons utilisée pour examiner l'expression des gènes. Contrairement aux approches classiques de clustering utilisées pour définir des regroupements de gènes, nous préférons rechercher des clusters dont le centre est un gène putatif (gp_i). En effet, en se focalisant sur les gènes putatifs, il est plus simple de déterminer les gènes connus qui lui sont associés et de plus cela permet de réduire considérablement l'espace de recherche. En outre cette approche offre également l'avantage de rapprocher des gènes connus entre eux (s'ils participent à la même fonction par exemple)

mais relativement aux gènes putatifs. La figure 2 illustre le principe général de l'approche en identifiant les différentes étapes ainsi que les différents choix qui peuvent être effectués. Dans la suite de cette section, nous décrivons les différentes étapes.

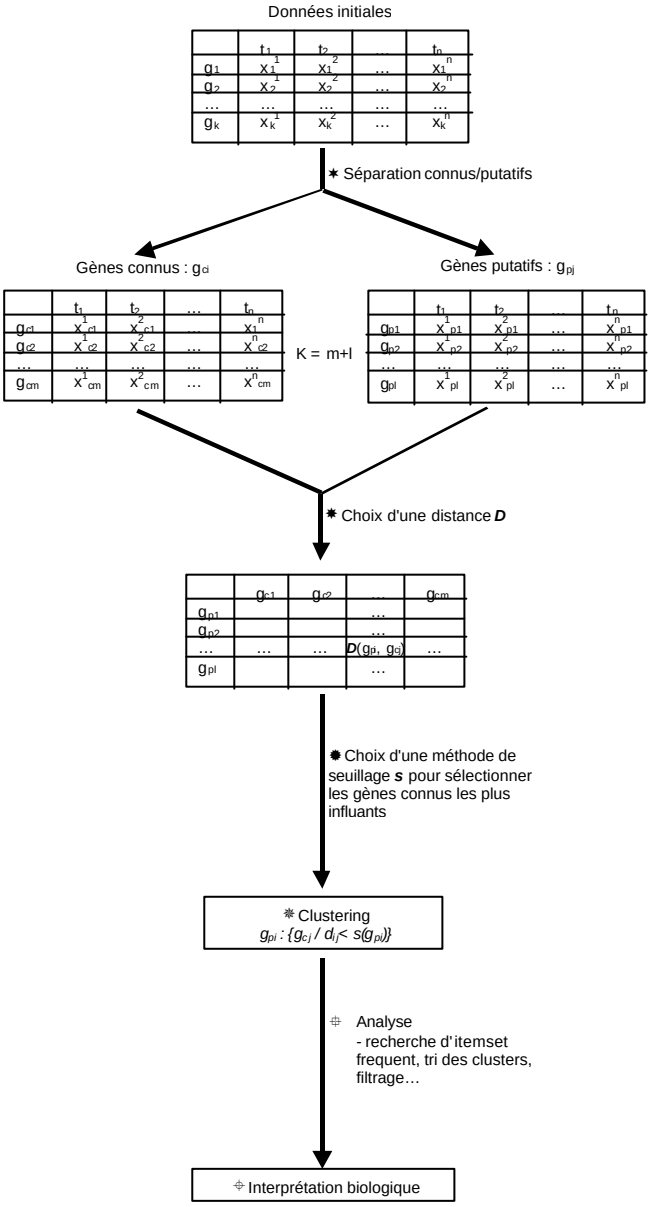

Fig. 2 – Principe général

3.1 Extraction de connaissances

Avant d'extraire la connaissance, il est tout d'abord nécessaire de séparer l'ensemble des gènes en deux sous ensembles (l'un contenant les gènes connus et l'autre les gènes putatifs).

La quantité d'information reliés aux fragments d'ADN (communément appelés séquences d'ADN) est exponentielle au cours du temps. Ces informations sont stockées sur une base de données (http://www.ncbi.nlm.nih.gov), et permettent à l'utilisateur de se renseigner sur la fonction d'une séquence d'ADN, à l'aide d'une analyse de type BLAST (*Basic Local Alignement Search Tool*) [AltGis90]. Dans le cadre de nos expérimentations, nous avons donc identifié les gènes putatifs (*gp*) ou connus (*gc*) par les résultats de l'analyse BLAST fournis par l'étude que nous avons prise comme référence [MaGl01]. Bien entendu, d'autres types de critères de sélection peuvent être utilisées sans modifier l'approche générale.

A l'issue de ce traitement, nous obtenons donc deux ensembles pour lesquels nous avons les données d'expression associées à chacun des gènes qu'ils contiennent. Le problème consiste maintenant à rechercher, pour chacun des gènes putatifs, quels sont les gènes connus qui possèdent le même comportement. Pour cela, nous considérons chaque expression de gènes comme une fonction discrète $C_{gi}=\{x_i / x_i=F_{gi}(t_i)\ i\ \hat{I}\ [1..n]\}$. Par exemple, dans le cas de la Fig.1, la courbe correspondant au gène n98138 a comme valeur C_{n98138}={5,60 ; 4,11 ; 0, 77 ; 1,03 ; 0,48}.

Afin d'extraire le sous ensemble des gènes connus ($gc_i\ \hat{I}\ G_{Known}$) associés à un gène putatif ($gp_i\ \hat{I}\ G_{Unknown}$), nous choisissons une fonction *D* de calcul de distance entre deux courbes et nous générons la matrice de distance de tous les gènes putatifs par rapport aux gènes connus.

La Fig. 3 illustre un exemple de matrice de distance dont l'ensemble associé à chaque gène putatif est trié par ordre croissant de distance.

gp_1	0.1, gc_1	0.3, gc_2	0.6, gc_3	4, gc_4
gp_2	0.2, gc_4	2, gc_3	3, gc_1	4, gc_2
gP_3	2, gc_2	3, gc_4	4, gc_1	5, gc_3

gp_1	0.1, gc_1	0.3, gc_2	0.6, gc_3	4, gc_4
gp_2	0.2, gc_4	2, gc_3	3, gc_1	4, gc_2
gp_3	2, gc_2	3, gc_4	4, gc_1	5, gc_3

Fig. 3 – Un exemple de matrice de distance entre gènes putatifs et connus

De cette matrice, il faut maintenant extraire, pour chaque gène putatif, un sous ensemble de gènes connus qui soient représentatifs de sa classe. Cela revient, en fait, à rechercher pour chaque gène putatif, quels sont les gènes connus qui sont à une distance inférieure à un seuil $seuil_1$ relativement à *D*.

Afin que l'ensemble des seuils ne soient pas constitués de valeurs trop divergentes (un tel cas correspondrait à obtenir des groupes ne contenant que des gènes connus trop éloignés pour être significatifs), il est nécessaire de pondérer cette valeur à l'aide d'un autre seuil

($seuil_2$) obtenu sur l'ensemble de la matrice. Le seuil retenu est alors celui correspondant au minimum entre les deux seuils ($seuil_1$ et $seuil_2$).

Les zones grisées dans la figure 3 illustre les gènes connus considérés comme trop éloignés par rapport aux valeurs de seuils pour être retenus. Dans la matrice supérieure, nous considérons la valeur du seuil $seuil_2$ égale à 1. La matrice du dessous présente les résultats pour le seuil $seuil_1$ uniquement en prenant la moyenne des distances.

Les groupes obtenus via la matrice des distances correspondent alors à l'ensemble des informations pertinentes retenues (i.e. les gènes connus) pour chaque gène putatif vis à vis de la distance *D* choisie.

Algorithme: *ExtractKnowledgeForUnknownGenes*
Input: *M* est la matrice d'expression des gènes; *nbgenesmax* est le nombre maximal de gènes par cluster.
Output: MatDist : la matrice qui associe à chaque gène putatif, la liste des couples gènes connus et distance qui lui sont les plus proches.

```
Begin
  // Init
     G_Unknown ¬ {Extract unknown genes};
     G_Known = M-G_Unknown;

  // Find Closed Genes
     Foreach gp_i Î G_Unknown do
        Foreach gc_i Î G_Known do
           MatDist[gp_i, gc_j] ¬{gc_i,D (gp_i, gc_j)};
        Done
        sort (MatDist[gp_i]); // sort by distances
        S[gp_i] ¬ Seuil(MatDist[gp_i], nbgenesmax);
     Done
     S=seuil(MatDist, nbgenesmax);

     Foreach gp_i Î G_Unknown do
        S[gp_i] ¬ min(S,S[gp_i]);
        Suppress(MatDist[gp_i,i]; pour tout i tel que MatDist[gp_i,i]>S[gp_i];
     Done
End
```

La fonction de seuil est exprimée de la manière suivante :
$S_i = e_i$ (ensemble des courbes des gènes connus) tel que :
$D(gp_i, gc_j) < 2*EcartType\ (D(gp_k,gc_l))\ \ "\ k,l$
et $D(gp_i, gc_j) < 2 * EcartType\ (D(gp_i, gc_k))\ \ "\ k$
et $card\ (e_i) < nbgenesmax$ (avec comme choix de critère $D(gp_i, gc_j)$ minimum).

Discussion sur le choix de la distance entre courbes

Il existe de nombreuses possibilités de calculer les distances entre les différentes courbes (une présentation des différentes distances utilisées pour analyser les données de puces à ADN est proposé dans [Drag03]). Le choix d'une distance peut, en fait, être séparé en deux étapes :

1) normalisation des courbes
2) calcul d'une norme entre deux courbes

La première étape permet d'identifier un certain type de courbes que nous souhaitons voir appartenir à la même classe. Par exemple, nous pouvons vouloir regrouper des courbes d'expressions similaires mais dont les amplitudes diffèrent ou regrouper des courbes similaires mais exprimées à des endroits différents de l'espace : décalage suivant les abscisses (décalage temporelle) ou les ordonnées.

La seconde étape permet d'obtenir une valeur réelle exprimant la distance entre les deux courbes normalisées et permet de définir les classes par leur taille. Cependant, selon la méthode de normalisation utilisée, les classes ne possèdent pas la même « forme » et peuvent donc regrouper des éléments différents. Dans l'espace affine associé aux courbes, une norme euclidienne définit des classes sphériques alors que la distance de Manhattan formera des classes cubiques.

Chacun des deux éléments permet de sélectionner des familles de courbes ayant le même comportement. Bien entendu, ce choix doit être guidé par une expertise de façon à obtenir des résultats significatifs dans le domaine considéré. Dans notre cadre, pour analyser l'expressions de gènes similaires au cours du temps, il est nécessaire de trouver une ou plusieurs distances qui soient capables d'exprimer que les courbes d'expression des gènes se comportant de manière similaire apparaissent dans la même classe. Ce résultat ne peut être obtenu que par apprentissage sur les jeux de données manipulés et par l'expert.

Dans le cadre de nos expérimentations, après analyse du jeu de données, nous avons considéré l'identité pour la normalisation et une norme euclidienne comme mesure de distance et des premiers résultats significatifs ont déjà été obtenu (Cf. Section 4). Toutefois, en intégrant plus de connaissances dans le choix de la distance, nous obtiendrons des résultats encore plus significatifs.

3.2 Interprétation des résultats

De manière générale, les différents résultats obtenus lors de l'étape précédente peuvent être classés selon différents critères. Ceux-ci peuvent être par exemple :

1) Les gènes putatifs pour lesquels nous disposons de la plus grande quantité de gènes connus proches. Dans ce cas, un simple tri sur les données de la matrice nous permet de classer les gènes putatifs en fonction du nombre de gènes connus associés.
2) Les gènes putatifs dont les gènes connus sont à une distance minimale. Comme précédemment il suffit de trier la matrice mais en fonction des distances.
3) Les gènes putatifs dont nous disposons d'une sémantique sur une fonction biologique connue, *i.e.* ceux qui contiennent des gènes connus proches disposant de cette information. Dans ce cadre les résultats tiennent véritablement compte de l'expert du

domaine et permettent de se focaliser plus directement sur des fonctions déjà reconnues de manière à définir plus rapidement les fonctions des gènes putatifs.

4) Les ensembles de gènes putatifs qui partagent les comportements de même gènes connus. En considérant qu'un gène connu est un item, nous pouvons appliquer directement des algorithmes de recherche de règles d'association soit pour rechercher uniquement les plus fréquents ou plus long itemsets, soit pour générer des règles comme dans [AgPe04].
5) Etc.

Bien entendu ces différents critères peuvent être utilisés ensemble de manière à faciliter le travail d'interprétation de l'expert biologiste. Par exemple dans les expériences menées, nous avons combiné à la fois le critère 3 en sélectionnant des fonctions connues et ensuite sur ces résultats nous avons appliqué des techniques de recherche de fréquents itemsets (critère 4).

Entre ces différentes possibilités, il est utile de rappeler ici l'intérêt du modèle choisi, qui est multiple. Dans le cadre du *Plasmodium Falciparum* nous avons déjà souligné le fait que les gènes inconnus sont prépondérants dans cet organisme. D'autre part, il possède un cycle de développement identifié en plusieurs phases, dont chacune d'elles possèdent des caractéristiques fonctionnelles établies (possédant une sémantique). Ces phases sont informatives et nous avons voulu les relier aux distances calculées précédemment. Ainsi, en recherchant des itemsets associés aux gènes putatifs, nous pourrons attribuer des fonctions aux gènes inconnus précédemment établies sur des gènes connus.

4 Expérimentations

De manière à valider notre approche, différentes expérimentations ont été réalisées. Notre approche a été développée en PERL et la recherche d'itemsets fréquents utilise l'implémentation d'Apriori développée par [Borg03].

Le premier jeu de données correspond aux valeurs d'expression de 943 gènes de *Plasmodium Falciparum*. Elles sont organisées en un tableau comportant une ligne par gène composée de 5 valeurs prises à 5 instants (temps) différents du cycle de développement du parasite. Le jeu de données ayant servi à cette étude est disponible à l'URL : http://www.microbiology.wustl.edu/dept/fac/goldberg/PfArray1.xls.

Cette étude portant sur l'expression des gènes de *Plasmodium Falciparum* est de première importance. En effet, le Paludisme est la maladie causée par ce parasite unicellulaire qui se multiplie dans le foie puis dans les globules rouges, suite à la piqûre d'un moustique nommé *anophèles gambiae*.

Plusieurs centaines de millions de personnes sont infectées par ce parasite chaque année dans le monde, et ce dernier cause le décès de plus d'un million de personnes par an. Cette maladie est aujourd'hui présente en Amérique du sud, en Afrique, en Inde et Asie du Sud-Est. Autrefois présente en Europe, notamment en Espagne et en Italie, Elle constitue une pandémie, touchant spécifiquement les pays en voie de développement. Seul le moustique femelle héberge ce parasite. Ce protozoaire se multiplie par un cycle reproductif composé de deux phases successives : une phase sexuée chez le moustique et l'autre asexuée chez l'Homme.

Le paludisme aussi appelé malaria est donc du à la présence d'un parasite unicellulaire (*Plasmodium falciparum*) dans la circulation sanguine. Les érythrocytes (globules rouges) infectés se sont révélés être une source intéressante pour l'étude du métabolisme parasitaire, et de son transcriptome. Des études biochimiques furent menées de longue date [Sher79], dans un effort concerté de recherche de solutions thérapeutiques.

Néanmoins, le génome de *Plasmodium* étant connu depuis peu [HofSub02], l'entière connaissance fonctionnelle des gènes fait défaut. A l'heure actuelle 60% de ces gènes n'ont pas de fonction connues. Nous avons donc entrepris, à partir d'une étude préalablement réalisée [MaGl01] d'extraire de la connaissance à partir d'un tableau d'expression de gènes, afin de pouvoir attribuer une fonction aux gènes putatifs.
L'expérience décrite ici a consisté à vérifier si notre approche était capable de retrouver facilement des comportements déjà connus. Pour cela, nous avons utilisé les données d'expressions de l'article de référence [MaGl01]. Les caractéristiques de ce jeu de données sont les suivantes : il comporte 943 gènes en totalité que nous avons subdivisé en 631 putatifs et 312 connus.

Dans ce jeu de données, nous avons tout d'abord recherché quels étaient les gènes qui possédaient des comportements communs. L'idée sous jacente était de voir si ces comportements communs entre expressions de gènes nous permettrait de proposer une fonction aux gènes putatifs. La figure 4 illustre un exemple de courbe obtenue pour le gène putatif t02499.1. Nous pouvons constater que seuls les gènes connus suivants : T025667 n98007, t18099, t02580 et n97774 appartiennent à la classe du gène putatif en fonction des seuils précédemment décrits.

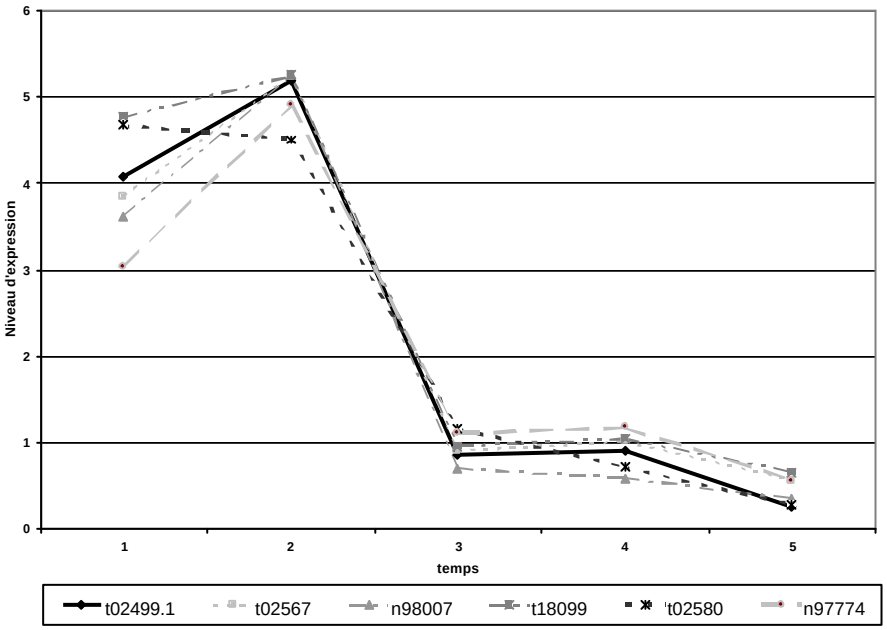

Fig. 4 – Un exemple de regroupement obtenu pour le gène putatif t02499.1

En ce qui concerne l'interprétation des résultats obtenus après traitement, le choix des items à rechercher fut dirigé par la sémantique des gènes. Il est déjà connu que pour des raisons énergétiques, le parasite favorise l'oxydation du glucose lors des premières heures de l'infection. C'est pourquoi les items reliés à l'oxydation du glucose furent recherchés en priorité. La figure 5 illustre les items recherchés et leur sémantique associée :

Sémantique du gène	Items
Lactate déshydrogénase (LDH)	T02580
Enolase	n97850
Glycéraldéhyde-3-phosphate déshydrogénase (GAPDH)	t18024
Glucose phosphate isomérase (GPI)	t18099
Aldolase	t18186
Pyruvate Kinase (PK)	N97635

Fig. 5 – Liste des items recherchés et leur sémantique associée

En filtrant les résultats obtenus précédemment avec ces items significatifs nous avons pu obtenir les différents gènes putatifs qui leur sont associés. Nous avons alors appliqué l'algorithme de recherche de fréquents itemsets afin d'obtenir les associations d'items les plus fréquentes et les plus longues. La figure 6 illustre les différentes apparitions des items dont nous connaissons la sémantique et sur lesquels nous avons recherché les itemsets.

Gènes putatifs	1er gène	2ème gène	3ème gène	4ème gène
n98138	t18197	t18073	t02580	t18094
t02493.1	t18073	t18197	t02580	t18099
t02496.1	t18197	t02580	t18099	t18073
t02499.1	t02567	n98007	t18099	t02580
n98171	t18024	N97635	t18172	n97850
n98196	n97699	n97697	n97892	t18186
n98198	t18092	N97635	t02606	t18019
t17984	t18186	t02626	n97770	n97892
t18006	t02580	t18197	t18099	t02567
t18059	t18099	t02567	t02580	n98007
t18163	t18016	t18172	t18092	t18099
t18175	n97892	t18102	t18186	n97697
N97630	t18024	n97850	n97635	T18172
n97865	t18186	n97892	n97795	t18102
n98071	t02580	n98181	t18143	t02567

Fig. 6 – Recherche des items au sein des gènes connus proches des gènes putatifs

Suite à l'obtention de ces résultats, nous obtenons le tableau de la figure 7, dans lequel les gènes ont été remplacés par leur sémantique.

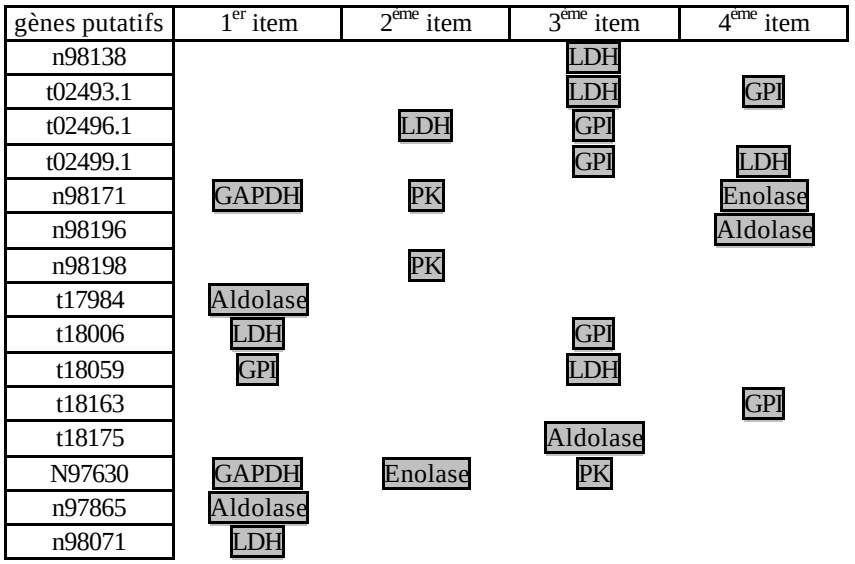

Fig. 7 – Représentation du sens suspecté des gènes putatifs

A ce stade du travail, nous sommes en mesure de proposer une fonction aux gènes putatifs. L'observation du tableau de la figure 7 nous permet une première approche de leur distribution. A plus grande échelle, nous avons pu observer que deux items dont la sémantique est opposée ne sont jamais associés au même gène putatif. Riche de ce constat, nous avons poursuivi l'attribution des fonctions aux gènes putatifs par un ensemble de protocoles Bio-informatiques communément utilisés :

Premièrement, pour les gènes putatifs, proches des items GPI et LDH, nous n'avons pas observé de similitude ou d'identité de séquence. Cette analyse fut réalisée par le logiciel CLUSTALW [ThoHi94] en accès anonyme à l'URL http://www.infobiogen.fr. Les résultats de cette analyse sont présentés en figure 8. Ces gènes ne partageant aucune similitudes de séquences, ils représentent bien des ARN messagers putatifs différents, réunis par la proximité d'items recherchés communs.

Les lignes 1 à 5 œrrespondent respectivement aux gènes T18006, T02493, T18059, T02499, T02496. Les chiffres indiquées en fin de chaque ligne, indiquent la cote de chaque séquence à cette position.

```
           10        20        30        40        50        60
1  TTTTTTTTTTTTTTTTAAACAAGGACATTTTCATATAAATATGAATGAGATTTTNCATAC   60
2  ------------------------ACG---------AAATATAACTNA------------   15
3  ----------------------TCCGTTT---------TACAATTAG------------   16
4  ----------------------CACGAGTT--------TTTTNACAG------------   17
5  ----------------------CTCGAGTT--------TTTTTTT--------------   15
C  ......................:+C*::*+.....:::T+*:+:*::............
           70        80        90       100       110       120
1  AAATATTAGTCACTAATGCCAAAAAAAAANTATTAT-GNTAC-ATTTNCTT--CAAAGAAA  116
2  AGA-----------------AAAAAAAGGTAAAAT-CAAAA-ATCTCTNA--TTAAAATA   54
3  ------------------TTAATACACATTCTAAT-GGTTCTGTATCATA--TAAAACTC   55
4  --------------------ATACACAATTTGATACATATTGTTTAGTAGCTTTAATTG   56
5  ---------------------------TTTTTTTTTTTTTCNTTTTTTTTTTTTTGCTT   47
C  :.:.................+*:*:*:+:T:*:*T.::*+:::T+T::*+..*++*+:*:
          130       140       150       160       170       180
1  TAATCAATAA-GAATTGA-----TCTCAAAACT---AAAA--ATGTAAAACT----ATG-  160
2  TA-TAAATA--GTAGCGC-----TGTAAAAACA---CAAA--ATNCATCANTTA--ATG-   98
3  TA--GAATGT-GTATTAACTAATTGTAAAGNTT---TATATGATACAGAACC----ATT-  104
4  TNATTGGTATAGTGGCAGGATTAGCTCAAACCATTGCATTTAATATAGGATCAACCATGG  116
5  TA---AATTTTGTNATAT--AATTTTCNAGGTTTT-NATTTTNAATATAATTNAT-ATAT  100
C  T*:+.**T++.G*+:++:..:::*:T+*A+:++::.:A+++.**++A:+A:+.+..AT+.
          190       200       210       220       230       240
1  -ACCTAAAATCCT--GCTATCAAGCAAGANTAGGGGGGGGCACTCCAGGGNA-AACTACT  216
2  -ANCTTNAAA-----AATAT--AGTAATAACGANGNNTG-----------TA-ANGNATN  138
3  -AGNTGAAAATTTAAATTATGAATCAAAAA---AGGTTGA---------ATATAAGTNTT  151
4  AAGATAATATGGGTGGTTATATGTCAGCAGGTATTGGTAT--------ATCAGGAGTATT  168
5  -ATATAAANN-----NTTNNNNNAANANNNCNNANACNNN----------NATATATNTN  144
C  .A::T+***:..:..:+T**.:+:+**.*::.::++:++...........:A:*++*+*+
          250       260       270       280       290       300
1  CATCCCTGGNGACCANTCTGCTGATTTATTGCTTCTCTTTTCAAGTGCCCCAATAAATAA  276
2  TAT-------GGCTAGT-----AATTTAATGCNTCTCC-----------------GTAT  168
3  CTT--------GCAATTGCNCTGATT--GTCCTTGTCCTATC------------CAAAAT  189
4  TATT-----TTNTTAAT----TAATTTATTACTTGANC-----------------AATTC  202
5  NAN--------NCCAT-------GTNTATNACATTNGN-----------------GNNT  171
C  :**.......::*:A:*...:+:*T***+*:C+T:+++:.::.............+++++
          310       320       330       340       350       360
1  GCAAATCCCTATTAATA--A---TAT------TAATAAAAT-----AATTTTGTGTATAA  320
2  GCAGATTGGTNTANNTG--G---T-T------TANNANGG------AANCTTTTNTG---  207
3  GAAGATTCTGAAAAAGA--AAAANCTGAAACGCNAGAAGA------AAGTTTAAATA---  238
4  GTATCTCCCGAAAANCATTATGGTGT------TAATAAAGCAAAGTAATTATATTTATAT  256
5  TTAAATATTTATAATGNT-----TAT------TNATNTAN------NNATNTATTNAN--  212
C  *:A:*T:+:+*+**::+:.+...*:T......*+*+*++:......**:*+T+*:**::.
          370       380       390       400       410
1  AAAATAAGGAATTTTAG----------GGGG--GCTTTGGGG-----------          350
2  -----GAGTNNTCTTGG----------GNNG--GTTTCACCGTCCTGNTTCTT          243
3  -----AAGAAGTTGTTG----------AAGGATGTTCAACTGGTATGTGCCCA          276
4  ATAACCNGTGANCTTNGTTTAATATTAGCTATAGTATNTGNGGTATGTAATT-          308
5  -----NNATAANNNTNCCCCCCC-CCNNNAACGGNAATNATNGTNCGG-----          254
C  :.::.:+*++++:+T:*.........+::+..G+++::::*++:+*:..::.
```

Fig. 8 – Alignement des séquences (sur 413 nucléotides) des gènes putatifs associés aux items GPI et LDH (analyse par CLUSTALW).

Le consensus (C) est matérialisé par une séquence de caractères qui, selon la situation rencontrée en une position donnée, seront les suivants :"." si le caractère majoritaire a une représentation >= 20%, ":" si le caractère majoritaire a une représentation >= 40%, "+" si le caractère majoritaire a une représentation >= 60%, "*" si le caractère majoritaire a une représentation >= 80%, le caractère lui-même, si celui-ci est identique pour l'ensemble des séquences.

Parmi l'ensemble des gènes putatifs, nous en avons découvert 15 associés à la sémantique que nous recherchions (Cf. Figure 5). Parmi ces 15 gènes, nous pouvons remarquer que 5 d'entre eux sont reliés à la fois aux items GPI et LDH.
Il est utile de rappeler à ce point la sémantique de GPI et LDH qui sont associés à l'oxydation du glucose chez *Plasmodium Falciparum* et participent au stockage d'énergie intracellulaire. Ces items sont associés à la première phase de développement du parasite. Nous pouvons à ce stade, supposer que les 5 gènes putatifs assurent des rôle proches. Afin de vérifier ceci, nous avons effectué une analyse avec BLASTX sur le gène t02499.1,dont les résultats sont montrés en figure 9 [GisStat93]. Cette analyse utilisant les connaissances du code génétique, nous informe sur les protéines putatives que peut générer un fragment d'ADN. Cette analyse est couramment utilisée à l'heure actuelle dans les laboratoires de génomique et de biologie, lorsqu'un gène putatif est étudié.

A.

Peptide obtenu sur la phase 1 :
HEFFXRYTI*YILFSSFNCXWYSGRISSNHCI*YRINHGR*YGWLYVSRYWYIRSIYFXN*FIT*XIRISRKXLWC**SKVIIFIYNX*
XXFNISYSXXGM*

Peptide obtenu sur la phase 2 :
TSFXTDTQFDTYCLVALIVIGIVAGLAQTIAFNIGSTMEDNMGGYMSAGIGISGVFIXLINLLLXQFVSPEXHYGVNKAK*LYLYITXX
LXLILAIVXXVCN

Peptide obtenu sur la phase 3 :
RVFXQIHNLIHIV**L*LXLV*WQD*LKPLHLI*DQPWKIIWVVICQQVLVYQEYLFX*LIYYLXNSYLPKXIMVLIKQSNYIYI*PVX
XV*Y*L*YXXYVI

Peptide obtenu sur la phase 4 :
NYIPXIL*LILNXXHXLYINIITLLY*HHNXFREIRIXQVIN*LXK*ILLIYQYLLTYNHPYYLPWLILY*MQWFELILPLYQXQLKLL
NNMYQIVYLXKNS

Peptide obtenu sur la phase 5 :
ITYXXYYS*Y*TKXTGYI*I*LLCFINTIMXFGRYELXK**IN*XNKYS*YTNTC*HITTHIIFHG*SYIKCNGLS*SCHYTNXN*SY*
TICIKLCICXKTR

Peptide obtenu sur la phase 6 :
LHTXXTIANIKXXSXVIYKYNYFALLTP*XFSGDTNXSSNKLIXKINTPDIPIPADI*PPILSSMVDPILNAMV*ANPATIPXTIKATK
QYVSNCVSVKKLV

B.

```
  5  SFXTDTQFDTYCLVALIVIGIVAGLAQTIAFNIGSTMEDNMGGYMSAGIGISGVFIXLINLLLXQFVSPEXHYGVNKAK*LYLYITXXLXLILAIVXXVCN307
     SF TDTQFDTYCLVA IVIGIVAGLAQTIAFNIGSTMEDNMGGYMSAGIGISGVFI +INLLL QFVSPE HYGVNKAK LYLYI   L LILAIV  VCN
109  SFFTDTQFDTYCLVAFIVIGIVAGLAQTIAFNIGSTMEDNMGGYMSAGIGISGVFIFVINLLLDQFVSPEKHYGVNKAKLLYLYIICELCLILAIVFCVCN209
```

Fig. 8 –Résultats de l'analyse BLASTx.

A. la déduction des six protéines putatives à partir d'une séquence d'ADN :

Le gène t02499.1 est traduit successivement dans les 6 phases de lecture.On obtient ainsi différentes protéines putatives codées par le fragment d'ADN analysé. La séquence de 308 nucléotides du gène t02499.1 peut coder pour des peptides de 102 acides aminés. Le symbole * représente un codon STOP, signal de fin de synthèse de la protéine.

B. leurs comparaison à une banque de données protéique :

Le peptide de la phase 2 (ligne soulignée), soumis à la comparaison de séquence sur une banque protéique, nous identifie une homologie forte avec une protéine (transporteur de nucléoside, en gris). La ligne intermédiaire indique le symbole de l'acide aminé lorsque il y a identité entre ces deux séquences, ou le signe « + » lorsque les acides aminés sont différents mais possèdent les même propriétés. Les nombres indiqués de part et d'autre de l'alignement des deux protéines, indiquent les côtes des séquences nucléique (gène t02499.1) et protéique (pour la protéine identifiée).

Cette analyse nous a permis de découvrir une fonction associée à l'un des 5 gènes putatifs isolés parmi les 943 initiaux. En effet, ce gène t02499.1 code pour une protéine de type transporteur de nucléoside, dont le rôle est d'économiser de l'énergie cellulaire en important par transport les nucléosides dont la synthèse est coûteuse en énergie. Cette énergie est d'autant plus précieuse à conserver qu'elle est nécessaire au parasite, pour qu'il puisse rapidement réinfecter une autre cellule.

En analysant les tendances au cours du temps, nous avons donc pu trouver par comparaison des comportements de plusieurs gènes le fait que le gène t02499.1 suivait le même comportement que les gènes GPI et LDH. D'autre part, notre analyse identifie dans l'ensemble des gènes putatifs, 5 gènes de fonction similaire aux items associés.

Cette méthode rapide et élégante nous permet d'assigner des fonctions à des gènes putatifs, grâce à l'existence d'un jeu réduit d'items dont on dispose de la sémantique. Dans le contexte actuel, il est fréquent d'étudier de nombreux transcriptomes pour lesquels on dispose de beaucoup de gènes putatifs et de peu de gènes sémantiquement caractérisés. C'est à cette situation que notre méthode propose une solution.

Suite à ces expérimentations, nous réalisons actuellement une autre étude sur un nouveau jeu de données du *Plasmodium Falciparum* [BoLl03] possédant les caractéristiques suivantes :

- un nombre de gènes supérieur à 6000,
- des données d'expression existant pour chaque gène sur 48 conditions temporelles,
- un nombre de gènes sémantiquement caractérisés plus important.

5 Conclusion

Dans cet article, nous avons proposé une nouvelle approche d'analyse des expressions de gènes basée sur l'analyse des tendances temporelles. L'originalité de l'approche est de se focaliser sur les gènes putatifs de manière à déterminer parmi les connus ceux qui leurs sont le plus associés. Nous avons utilisé l'approche proposée dans le cadre de données d'expression de gènes concernant le *Plasmodium Falciparum*. Les résultats obtenus ont permis de faire ressortir des comportements intéressants du transcriptome de *Plasmodium Falciparum*.

Une seconde originalité de l'approche est également d'être suffisamment générique pour pouvoir rechercher des comportements différents (e.g. des gènes qui s'expriment de la même manière mais avec une amplitude différente, des gènes qui se comportent de la même manière mais avec un décalage au cours du temps, …). En effet, comme nous l'avons montré dans la section 3, le choix de la mesure de distance permet de faire varier les différents comportements recherchés. A l'heure actuelle, nous poursuivons ces travaux en association avec des biologistes pour permettre d'analyser d'autres types de comportements en utilisant différentes distances.

Références

[AgPe04] M. Agier, J. M. Petit, V. Chabaud, Y-J Bignon et V. Vidal. « Vers différens types de règles pour les données d'expression de gènes – Application à des données de tumeurs mammaires ». Actes du 22ièmes Congrès Inforsid (INFORSID'04), Biarritz, mai 2004.

[BoLl03] Z. Bozdech, M. Llinas, B. L. Pulliam, E. D. Wong, J. Zhu and J. DeRisi. « « The Transcriptome of the Intraerythocytic Developmental Cycle of Plasmodium Falciparum ». In PloS Biology, Vol. 1, N. 1, 2003.

[Borg03] C. Borgelt. « Efficient Implementations of Apriori and Eclat ». In Proceedings of Workshop of Frequent Item Set Mining Implementations (FIMI 2003), Melbourne, FL, USA, 2003.

[ChCh00] Y. Cheng and G.M. Church. « Biclustering of Expression Data ». In Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'2000), San Diego, USA, pp. 93-103.

[Drag03] S. Draghici. « Data Analysis Tools for DNA Microarrays». Chapman & Hall, CRC Mathematical Biology and Medecine Series, 2003.

[JiPe04] D. Jiang, J. Pei, M. Ramanathan, C. Tang and A. Zhang. « Mining Coherent Gene Clusters from Gene-Sample-Time Microarray Data ». In Proceedings of the Internation Conference on Knowledge Discovery and Data Mining (KDD'04), Seattle, USA, August 2004.

[LiYa04] J. Liu, J. Yand and W. Wang. « Biclustering in Gene Expression Data by Tendency ». In Proceedings of 2004 IEEE Computational Systems Bioinformatics Conference (CSB'2004), Stanford, USA, August 2004.

[MaGl01] C.B. Mamoun, I.Y. Gluzman, C. Hott, S.K. MacMillan, A. S. Amarakone, D. L. Anderson, J. M. Carlton, J. B. Dame; D. Chakrabarti, R. K. Martin, B. H. Browstein, R. K. Martin, B. H. Brownstein and D. E. Goldberg. « Co-ordinated Program of Gene Expression During Asexual Intraerythrocytic development of the Human Malaria Parasite Plasmodium Falciparum Revealed by Microarray Analysis ». In Molecular Microbiology, Vol. 39, N. 1, 2001, pp. 26-36.

[MaOl04] S.C. Madeira and A. L. Oliveira. « Biclustering Algorithms for Biological Data Analysis: A Survey ». In ACM Transactions on Computational Biology and Informatics, Vol.1, N. 1, January 2004, pp. 24-45.

[Sher79] I. W. Sherman. «Biochemistry of Plasmodium (Malarial Parasites) ». In Microbiological Reviews, Vol. 43, N. 4, December 1979, pp. 453-495.

[AltGis90] S.F. Altschul, W. Gish,W. Miller,E.W Myers and D.J. Lipman. «Basic local alignment search tool ». In Journal of Molecular Biology, 1990, Vol 215, pp. 403-410.

[GisStat93] W. Gish, D.J. States. «Identification of protein coding regions by database similarity search. » In Nature Genetics, Vol. 3, 1993, pp. 266-272.

[HofSub02] SL. Hoffman, GM. Subramanian, FH.Collins, J.C. Venter. « Plasmodium, human and Anopheles genomics and malaria ». In Nature, Feb 2002, Vol 415 (6872), pp. 702-9.

[ThoHi94] J.D Thompson, D.G.Higgins, T.J. Gibson, « CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice ». In Nucleic Acids Research, 1994, Vol. 22, pp. 4673-4680.