Tendances dans les expressions de gènes : application à l'analyse du transcriptome de *Plasmodium Falciparum* *

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 les puces à ADN. 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 étudie l'ensemble des gènes sans distinction. L'approche que nous proposons deux catégories de 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 *Plasmodium Falciparum* (agent de la Malaria) 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.

Summary. Gene expression studies have been tremendously changed by DNA arrays technology. Based on the sentence « same profil, same function », severals algorithms have been developped such as partioning or hierarchical clustering. However, genes are taken as a whole. In this paper, genes are sorted by their cellular gene function, giving two classes of genes (known and putatives). Thus, frequent items are searched in the neighbourhood of each unknown genes, and gives clues about their hypothetical functions.
As a first step towards, a first experiment with energetical items on *Plasmodium Falciparum* (Malaria's agent) identify a new function to a nucleosid transporter, expressed in the early stage of infection.

* Ce travail est supporté par un projet pluridisciplinaire GEMBIO du groupement des Ecoles des mines dans le domaine de la Bio-informatique.
** P.COLLET, Docteur en Biologie Structurale, Moléculaire et Cellulaire, actuellement en Post Doc au centre de recherche LGI2P.

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), notamment la possibilité de quantifier l'expression de plusieurs milliers de gène simultanément permet d'obtenir un « instantané » du transcriptome de la cellule. Nous avons donc entrepris, à partir d'une étude de puces à ADN sur le génome de *Plasmodium* préalablement réalisée (Mamoun 2001) d'extraire de la connaissance afin d'attribuer une fonction aux gènes. A l'heure actuelle, pour analyser les puces, de nombreuses techniques d'extraction sont utilisables pour obtenir des règles de sémantiques différentes (Agier et al. 2004) ou pour former des groupes de gènes. Ainsi des techniques de clustering (Jiang et al. 2004, Madeira et Oliveira 2004] sont utilisées pour, regrouper des gènes selon leur expression en fonction des différentes conditions, regrouper des conditions expérimentales en fonction des profils d'expression sur chaque gènes, ou déterminer la fonction de gènes putatifs grâce à l'expression de gènes connus déjà réunis en clusters. Cependant l'utilisation de 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 connus. Ainsi nous pouvons, pour les gènes putatifs et suivant une hypothèse biologique donnée, obtenir les items les plus fréquents dans un voisinage informatif (i.e. 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 expériences menées. 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 $G = G_{Known} \cup G_{Unknown}$ l'ensemble des gènes où G_{Known} représente l'ensemble des gènes connus et $G_{Unknown}$ l'ensemble des gènes hypothétiques ou putatifs. Soit $F = G \times E \rightarrow \Re$, 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 $M(G, E)$ matrice d'expression de gènes où chaque colonne correspond à une condition expérimentale et où chaque ligne correspond à un gène. Chaque élément de M prend ses valeurs dans F. 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 $gp_i \in G_{Unknown}$, 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.

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). Dans le cadre de nos expérimentations, nous avons identifié les gènes putatifs (*gp*) ou connus (*gc*) par les résultats de l'analyse BLAST (Altschul et al. 1990) fournis par l'étude que nous avons prise comme référence (Mamoun et al. 2001). 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 $C_{gi}=\{x_i / x_i=F_{gi}(t_i)\ i \in [1..n]\}$. 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 ($gc_i \in G_{Known}$) associés à un gène putatif ($gp_i \in G_{Unknown}$), 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. De cette matrice, il faut maintenant extraire, pour chaque gène putatif, un sous ensemble de gènes connus qui sois 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. 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.

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

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 avec les distances associées qui lui sont les plus proches.
Begin
// Init
$G_{Unknown} \leftarrow$ *{Extract unknown genes}*;
$G_{Known} = M\text{-}G_{Unknown}$;
// Find Closed Genes
Foreach $gp_i \in G_{Unknown}$ **do**
Foreach $gc_i \in G_{Known}$ **do**
$MatDist[gp_i, gc_j] \leftarrow \{gc_i, D(gp_i, gc_j)\}$;
Done
sort $(MatDist[gp_i])$; // sort by distances
$S[gp_i] \leftarrow Seuil(MatDist[gp_i], nbgenesmax)$;
Done
$S=seuil\ M(MatDist, nbgenesmax)$;
Foreach $gp_i \in G_{Unknown}$ **do**
$S[gp_i] \leftarrow min(S, S[gp_i])$;
$Suppress(MatDist[gp_i,i]$; *for each* $i / MatDist[gp_i,i] > S[gp_i]$;
Done
End

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

Il existe de nombreuses possibilités de calculer les distances entre les différentes courbes (par exemple une présentation des différentes distances utilisées pour analyser les données de puces à ADN est proposée dans (Draghici 2003). 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 été obtenu.

4 Expérimentations

Notre approche a été développée en PERL et MatLab, la recherche d'itemsets fréquents utilise l'implémentation d'Apriori développée par (Borgelt 2003). Le 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 est disponible à l'adresse http://www.microbiology.wustl.edu/dept/fac/goldberg/PfArray1.xls.
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 permettraient d'identifier une fonction aux gènes putatifs. 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 (T02580 : Lactate déshydrogénase (LDH), n97850 : Enolase, t18024 : Glycéraldéhyde-3-phosphate déshydrogénase (GAPDH), t18099 : Glucose phosphate isomérase (GPI), t18186 : Aldolase, N97635 : Pyruvate Kinase (PK)). 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 2 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. 2 – Recherche des items au sein des gènes connus proches des gènes putatifs

A ce stade du travail, nous sommes en mesure de proposer une fonction aux gènes putatifs. A plus grande échelle, nous avons pu observer que deux items dont la fonction est antagoniste ne sont jamais associés au même gène putatif. 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 (Thompson 1994) (http://www.infobiogen.fr). Ces gènes ne partageant aucune similitudes de séquences, ils représentent bien des ARN messagers putatifs différents, regroupés par notre approche. Parmi l'ensemble des gènes putatifs, nous en avons découvert 15 associés à la sémantique que nous recherchions. 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 qui est associée à l'oxydation du glucose chez *P. Falciparum* et participe au stockage d'énergie. 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 rôles proches. Afin de vérifier ceci, nous avons effectué une analyse avec BLASTX sur le gène t02499.1 (Gish 1993). Cette analyse utilisant les connaissances du code génétique, nous informe sur les protéines putatives que peut générer un fragment d'ADN. 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 nucléosides dont le rôle est peut être d'économiser de l'énergie en important des 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.

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 *P. Falciparum*. Les résultats obtenus ont permis de faire ressortir des comportements intéressants du transcriptome de *P. Falciparum*.

Références

Agier M., Petit J. M., Chabaud V., Bignon Y-J., et Vidal V., (2004), 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), Biarritz, mai 2004.

Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J., (1990),Basic local alignment search tool, Journal of Molec. Biol, Vol 215, pp. 403-410, 1990.

Borgelt C, (2003), Efficient Implementations of Apriori and Eclat, In Proc. of Workshop of Frequent Item Set Mining Implementations (FIMI), USA, 2003.

Draghici S, (2003), Data Analysis Tools for DNA Microarrays, Chapman & Hall, CRC Mathematical Biology and Medecine Series, 2003.

Gish W., States D.J., (1993), Identification of protein coding regions by database similarity search, Nature Genetics, Vol. 3, pp. 266-272, 1993.

Jiang D., Pei J., Ramanathan M., Tang C., and Zhang A. (2004), Mining Coherent Gene Clusters from Gene-Sample-Time Microarray Data, In Proc. KDD'04, August 2004.

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

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

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