

Mémoire de stage long
d'Anne-Sophie Carpentier

<p>Etude et comparaison de méthodes statistiques pour l'analyse de données de transcriptome</p>
--

DEA de génétique multifactorielle
de l'université d'Orsay Paris XI, cohabilité avec Paris VI, VII et l'INA P-G

Année 2001-2002

Equipe d'accueil : Laboratoire "Génome et Informatique"
CNRS/Université d'Evry
Tour Evry2, 523 Place des Terrasses
91034 EVRY

Responsables de stage: Alain Hénaut et Jean-Loup Risler

Résumé

La plupart des travaux en génomique fonctionnelle utilisent les puces à ADN pour déterminer les variations d'expression des gènes en fonction des conditions expérimentales. Ces puces génèrent des volumes de données importants qui posent des problèmes d'analyse statistique. Aucun consensus n'existe jusqu'à présent sur les méthodes à employer pour analyser les données obtenues. L'objet de ce stage est de rechercher des méthodes d'analyse statistique à la fois robustes et sensibles et de comparer ces différentes méthodes sur deux jeux de données : une expérience sur le métabolisme du soufre de *Bacillus subtilis* et une expérience de transcriptome sur le cycle circadien de la souris.

Cette étude porte sur quatre méthodes : des méthodes exploratoires comme l'analyse en composantes principales (ACP) et l'analyse en composantes indépendantes (ACI) et des analyses plus dirigées comme l'analyse de variance (ANOVA) et la régression par la méthode des moindres carrés partiels (PLS). Par ailleurs, deux méthodes complémentaires de représentation des profils d'expression ont été abordées : la classification hiérarchique et l'ACP.

La comparaison des méthodes se fonde essentiellement sur l'expérience sur *Bacillus subtilis*, car l'organisation en opérons des gènes fournit un témoin interne de la cohérence des résultats. L'ANOVA donne les résultats les plus cohérents dans la mesure où elle classe à peu près de la même façon la plupart des gènes d'un opéron. Elle présente aussi l'avantage de fournir une bonne approximation de la signification statistique des résultats. Cependant, elle nécessite un plan d'expérience complet et le plus régulier possible.

L'ACI et la PLS font ressortir, elles aussi, des opérons. Toutefois, leurs résultats sont, en général, moins cohérents qu'avec l'ANOVA. L'ACI présente néanmoins l'intérêt d'être une technique exploratoire, susceptible de déceler des profils d'expressions inattendus.

L'ACP n'a pas donné de résultats satisfaisants pour la recherche de gènes différentiellement exprimés. En revanche, elle est intéressante pour représenter les profils d'expression obtenus (particulièrement dans le cas des séries temporelles).

Sommaire

1	INTRODUCTION	2
2	PRESENTATION DES DONNEES UTILISEES DANS CE RAPPORT	4
2.1	DONNEES D'UNE ETUDE DU METABOLISME DU SOUFRE DE BACILLUS SUBTILIS.....	4
2.2	DONNEES D'UNE ETUDE DU RYTHME CIRCADIEEN DE LA SOURIS	4
2.3	TRANSFORMATIONS PRELIMINAIRES.....	5
3	METHODES EXPLORATOIRES TYPE ACP, ACI.....	6
3.1	UNE METHODE CLASSIQUE DE VISUALISATION DU NUAGE PAR SA PROJECTION SUR LES AXES D'INERTIE : L'ACP.....	6
3.1.1	<i>Exposé de la méthode</i>	6
3.1.2	<i>Résultats obtenus</i>	7
3.2	UNE METHODE DE RECHERCHE DES AXES MAXIMISANT L'ECART A LA NORMALITE : L'ACI.....	9
3.2.1	<i>Principe de la méthode</i>	9
3.2.2	<i>Réalisation concrète</i>	10
3.2.3	<i>Résultats obtenus</i>	11
4	METHODES DIRIGEES.....	13
4.1	L'ANALYSE DE LA VARIANCE – ANOVA	13
4.1.1	<i>Exposé de la méthode</i>	13
4.1.2	<i>Résultats obtenus</i>	15
4.2	PLS	17
4.2.1	<i>Exposé de la méthode</i>	17
4.2.2	<i>Résultats obtenus</i>	17
5	LA REPRESENTATION DES PROFILS D'EXPRESSION	19
5.1	LA CLASSIFICATION HIERARCHIQUE	19
5.1.1	<i>Exposé de la méthode</i>	19
5.1.2	<i>Résultats obtenus</i>	20
5.2	L'ANALYSE DES SERIES TEMPORELLES PAR ACP.....	20
5.2.1	<i>Exposé de la méthode</i>	20
5.2.2	<i>Résultats obtenus</i>	21
6	SYNTHESE : COMPARAISON DES DIFFERENTES METHODES.....	22
7	AMELIORATIONS ENVISAGEABLES.....	24
8	BIBLIOGRAPHIE	25
ANNEXE : IMPLICATION DE L'ARGININE DANS LES VOIES METABOLIQUE DU SOUFRE DE B. SUBTILIS.....		27

1 Introduction

La dernière décennie du vingtième siècle a vu un changement radical de la recherche en biologie : des analyses de l'expression de gènes isolés, nous sommes passés à l'étude simultanée de nombreux gènes ou protéines. Cette avancée n'a pu se réaliser qu'à partir de progrès technologiques importantes et de la mise en place de séquençages systématiques de génomes. Bien que la majorité des travaux concernent les bactéries (90 % des génomes connus sont des génomes bactériens et 60 % des programmes en cours portent encore sur des bactéries), les observations s'accumulent sur les organismes pluricellulaires, depuis une mouche (*Drosophila melanogaster*) jusqu'à l'Homme en passant par les plantes (*Arabidopsis thaliana*). De nombreuses plantes d'intérêt agronomique sont en cours de séquençage comme le riz (presque achevé), le maïs, le tournesol ou encore le coton.

La généralisation du séquençage systématique a ouvert une nouvelle ère de recherche : la génomique fonctionnelle. L'objectif est maintenant de déterminer le rôle des gènes identifiés par le séquençage et de déceler leurs interactions afin de comprendre le fonctionnement du génome dans son ensemble. Les données expérimentales proviennent principalement de la mesure des niveaux intracellulaires des ARN (le transcriptome), des protéines (le protéome) et des métabolites (le métabolome).

En raison d'une relative facilité, la plupart des travaux portent actuellement sur la mesure des concentrations intracellulaires des ARN. Le principe de la méthode est basé sur l'hybridation spécifique d'un fragment d'un gène et de l'ARN correspondant. Plusieurs techniques coexistent actuellement (dépôts sur des membranes de nylon ou sur des lames de verres, utilisation de la radioactivité ou de la fluorescence, fragments de gènes de petite ou de grande taille), mais ceci a peu d'influence en pratique sur les résultats. Les objectifs des travaux peuvent être classés en ordre de complexité croissante :

1. obtenir des profils d'expression ayant une valeur diagnostique,
2. déterminer les variations d'expression des gènes en fonction des conditions physiologiques,
3. déceler les covariations afin d'en inférer les unités de régulation et leurs réseaux d'interaction.

L'analyse du transcriptome se distingue radicalement des expériences classiques en biologie par le volume de données qu'elle génère. Chaque extraction d'ARN total donne au minimum une mesure pour chaque gène alors qu'il y a de quelques milliers à quelques dizaines de milliers de gènes dans un génome. Les problèmes statistiques varient selon les étapes. Il est possible de distinguer plusieurs étapes en fonction des outils utilisés :

- le traitement des images et la quantification (passage des pixels aux niveaux d'expression),
- la transformation des données (filtres, normalisation, standardisation),
- l'analyse des données dans un espace à de nombreuses dimensions,
- et, enfin, la mise en relation des résultats avec d'autres types de données (bibliographie...).

Actuellement, il n'existe pas de consensus sur les méthodes à employer à ces différentes étapes. La seule chose communément admise est que les règles simplistes (comme « sélectionner les gènes dont le niveau d'expression varie d'un facteur deux ou trois ») (Chen

et al. 1997) sont inappropriées (Pan 2002, Draghici 2002), entre autre parce que l'ampleur des variations spontanées du niveau d'expression dépend des gènes.

De très nombreux travaux portent sur la recherche de méthodes d'analyse statistique à la fois robustes et sensibles qui pourraient être appliquées à l'étude du transcriptome (Chen *et al.* 1997, Efron *et al.* 2000, Ideker *et al.* 2000, Newton *et al.* 2001, Tusher *et al.* 2001, Lin *et al.* 2002, Pan *et al.* 2001). C'est aussi l'objet de ce stage. Plusieurs types de méthodes (classiques ou récentes) ont été étudiés : des méthodes exploratoires comme l'analyse en composantes principales (ACP) et l'analyse en composantes indépendantes (ACI) et des analyses plus dirigées comme l'analyse de variance (ANOVA) et la régression par la méthode des moindres carrés partiels (PLS).

2 Présentation des données utilisées dans ce rapport

Deux expériences d'analyse du transcriptome servent de support à ce rapport. Les données m'ont été aimablement fournies par l'équipe d'Antoine Danchin de l'Institut Pasteur de Hong Kong et par Franck Delaunay du laboratoire CNRS de physiologie des membranes cellulaires de Nice-Sophia-Antipolis.

2.1 Données d'une étude du métabolisme du soufre de *Bacillus subtilis*

Cette étude vise à analyser l'effet du changement de source de soufre sur le transcriptome de la bactérie *Bacillus subtilis* (Sekowska *et al.* 2001). Les cultures ont été menées en présence, soit de méthionine, soit de méthylthioribose. Elles doivent conduire à l'obtention de profils d'expression très semblables car peu de gènes sont supposés impliqués par ce changement de source de soufre. L'espace effectif des gènes doit donc avoir une dimension réduite, ce qui facilitera l'interprétation des résultats.

Les travaux sur *Bacillus subtilis* utilisent des membranes de nylon commerciales portant des sondes pour les 4107 gènes de la bactérie. Chaque gène fait l'objet de deux dépôts (les duplicata). Un écart excessif entre les valeurs des duplicata est le signe d'une anomalie sur la membrane (petit détrit, etc.). Les mesures sont dans ce cas vérifiées par l'expérimentateur. L'expérience montre que les membranes sont fiables, elles peuvent être réutilisées une dizaine de fois sans que cela affecte de façon sensible les résultats.

Pour chaque expérience, les ARN messagers totaux sont extraits d'une culture en mini-fermenteur. En fait, l'ARN n'est pas directement quantifié mais d'abord transcrit en ADNc pour des raisons de stabilité et de résistance à la dégradation. C'est au cours de cette étape que le phosphore radioactif est incorporé dans les ADNc. La préparation est faite en double, en utilisant dans un cas 1 µg et dans l'autre 10 µg d'ARN prélevés dans la même préparation d'ARN. Ceci est une façon d'éviter que des valeurs soient difficiles à apprécier, à cause d'un problème de saturation pour les fortes valeurs ou d'un bruit de fond excessif pour les plus faibles. Ce témoin permet aussi d'apprécier la relation entre la concentration d'ARN et la quantité d'ADNc produite.

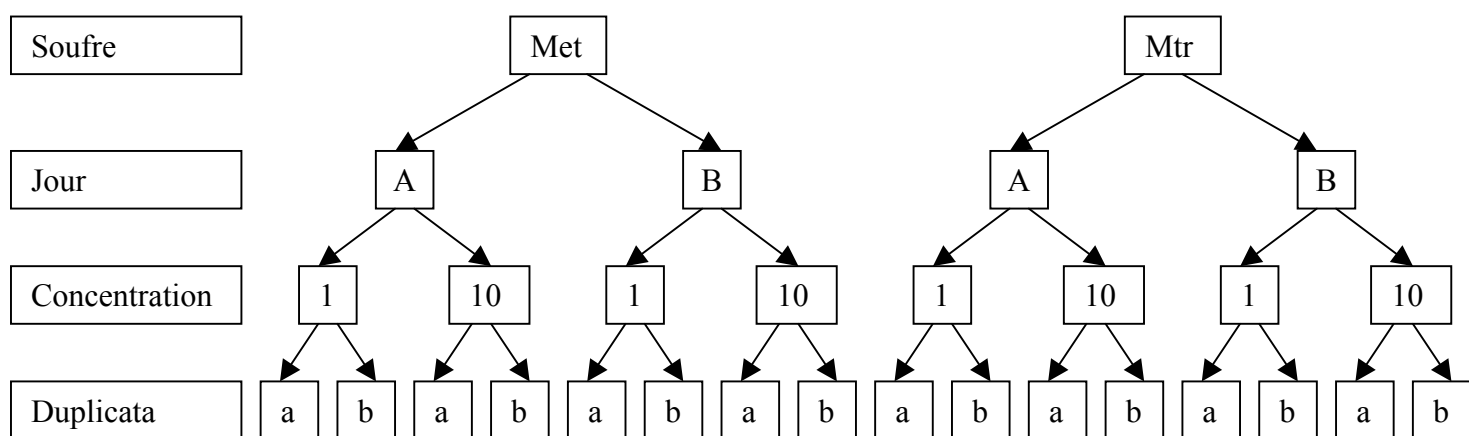
La variabilité introduite par des facteurs non maîtrisés a été estimée en répétant l'ensemble de l'expérience, cultures comprises, à deux jours différents.

Le plan d'expérience et les symboles utilisés sont résumés figure 1.

2.2 Données d'une étude du rythme circadien de la souris

L'étude porte sur l'évolution de l'expression de gènes selon le cycle jour / nuit dans le foie de la souris (cycle de 12 h de jour et 12 h de nuit). Pour cela, l'extraction des ARN messagers est réalisée toutes les 4 heures après l'allumage de la lumière (0, 4, 8, 12, 16 et 20 heures). Pour chaque extraction, la synthèse des ADNc est réalisée deux fois indépendamment (expériences A et B) afin de tenir compte du bruit qui est probablement introduit par l'étape de RT-PCR. Les ADNc sont marqués par des fluorochromes (rouge et vert).

Figure 1 : Plan d'expérience de l'analyse du transcriptome de *Bacillus subtilis*



Les mesures portent sur environ 6 000 gènes identifiés et 6 000 clusters d'EST (12 487 séquences codantes en tout). Elles sont réalisées sur des puces Affymetrix.

2.3 Transformations préliminaires

Le tableau de données (noté X) résumant une analyse du transcriptome est une matrice dont les lignes correspondent aux gènes et les colonnes aux conditions expérimentales. La valeur X_{ij} est la valeur mesurée pour le gène i pour la condition expérimentale j .

Dans toutes les colonnes, la distribution des données brutes est fortement asymétrique et comporte un petit nombre de très grandes valeurs. Ces dernières influencent très fortement l'évaluation de la plupart des grandeurs statistiques, notamment la détermination des axes d'inertie du nuage des gènes dans l'espace des conditions expérimentales. Pour remédier à ce biais, les données brutes sont remplacées par leur logarithme.

L'estimation et la gestion du bruit de fond sont des sources de problèmes pour les expérimentateurs. Une étude approfondie de plusieurs types de protocoles expérimentaux a montré que le mieux est de n'introduire aucune correction (Cette partie ne figure pas dans le mémoire).

3 Méthodes exploratoires type ACP, ACI

Le laboratoire a créé un logiciel dédié à l'analyse des expériences sur le transcriptome, GeneANOVA (Didier *et al.* 2002). Il offre trois grands types de méthodes : l'analyse en composantes principales (ACP), l'analyse de variance (ANOVA) et la régression PLS. Ces méthodes sont complétées par des outils facilitant la manipulation et la représentation des données.

L'analyse en composantes indépendantes (ACI) a été adaptée aux expériences sur le transcriptome par l'équipe de Bruno Torrèsani (Centre de physique théorique et Laboratoire d'analyse, topologie et probabilités – Marseille). Le stage a permis de modifier et d'enrichir l'implémentation initiale.

3.1 Une méthode classique de visualisation du nuage par sa projection sur les axes d'inertie : l'ACP

3.1.1 Exposé de la méthode

L'Analyse en Composantes Principales (ACP), aussi connue sous le nom de PCA (*Principal Component Analysis*), est une méthode descriptive permettant de visualiser la géométrie du nuage des observations quand l'espace a plus de trois dimensions (la dimension de l'espace dépasse dix dans les expériences analysées dans le rapport). Elle procède en deux étapes, la détermination des axes d'inertie du nuage (qui forment un système d'axes orthogonaux), puis sa projection sur les plans déterminés par les axes d'inertie pris deux à deux. Mathématiquement (Bry 1995), ceci revient à déterminer les vecteurs propres de la matrice XX' (avec X , la matrice des observations centrées réduites). Les axes sont numérotés par valeurs propres décroissantes. Les axes d'inertie sont donnés par les vecteurs propres et les valeurs propres correspondent à la part de l'inertie associée à chaque axe. Ils ne dépendent que de la forme du nuage et sont indépendants des coordonnées dans l'espace initial.

Des points voisins dans le nuage sont voisins sur tous les plans. En revanche, deux points peuvent être voisins sur un plan et distants dans l'espace, à cause des déformations dues à la projection. Comme il est impossible d'étudier toutes les projections, l'analyse d'un plan est restreinte aux points qui en sont très proches, la distance étant mesurée par la corrélation au plan.

L'analyse de la sphère des corrélations permet de visualiser les similitudes entre les profils d'expression associés aux différentes conditions expérimentales. La corrélation linéaire entre deux profils est égale au cosinus de l'angle entre les vecteurs correspondants aux conditions expérimentales (un angle de 90° montre une absence totale de corrélation entre les profils d'expression).

En plus de cette représentation graphique, l'ACP fournit également les coefficients de la combinaison linéaire permettant de passer du système d'axe initial au système d'axes final. L'analyse du signe et du poids des conditions expérimentales initiales permet de repérer les nouveaux axes qui représentent le mieux tel ou tel facteur (par exemple, dans le cas de *Bacillus subtilis*, les axes qui opposent le mieux la source de soufre, en mettant les cultures

faites avec de la méthionine d'un côté et celles faites avec le méthylthioribose de l'autre). Les gènes fortement corrélés avec ces axes sont des gènes dont l'expression varie principalement en fonction du facteur étudié.

3.1.2 Résultats obtenus

3.1.2.1 Le métabolisme du soufre chez *Bacillus subtilis*

Le tableau 1 récapitule les résultats obtenus dans l'expérience sur *Bacillus subtilis* :

- Le premier axe correspond au niveau moyen de l'expression des gènes dans l'ensemble des conditions expérimentales. Cette différence de niveau d'expression est toujours la principale source de variation dans les analyses de transcriptome, mais elle n'est pas très intéressante pour les biologistes puisqu'elle revient à dire qu'il y a des gènes qui sont peu ou pas exprimés et d'autres qui le sont fortement.
- Le deuxième axe sépare les deux concentrations d'ARN (1 ou 10 µg). Cette opposition n'est pas triviale. Les données ayant été au préalable centrées réduites, il ne devrait pas y avoir de différences importantes dans les mesures faites en prélevant 1 ou 10 µg de la même préparation d'ARN. Ceci montre que, pour un bon nombre de gènes, les niveaux d'expression prédits sont notablement différents suivant la quantité d'ARN utilisée pour préparer l'ADNc.
- Le troisième axe oppose le jour de culture (A ou B). Là encore, l'importance de ce facteur est une surprise puisque les cultures étaient faites dans des conditions aussi parfaitement contrôlées que possible (culture en phase exponentielle dans un mini-fermenteur).
- Le cinquième montre la variation de l'expression des gènes selon la source de soufre (met et mtr). C'est l'axe le plus intéressant dans le cadre de cette expérience. La faible inertie de cet axe (0,36 % de l'inertie totale) est un résultat attendu. En effet, l'expérience a été conçue pour que le changement de source de soufre affecte peu de gènes. Ce sont ces quelques gènes qui créent l'inertie propre à l'axe *soufre*. Comme les changements de niveau d'expression sont faibles, ces gènes restent près du premier axe d'inertie et la déformation du nuage le long de l'axe *soufre* sort à peine du bruit de fond.
- Le septième axe permet de séparer les duplicata (a et b).
- Le quatrième et le sixième axe ne sont pas directement interprétables en terme de combinaison des facteurs *soufre*, *jour* et mode de préparation de l'ADNc.

L'étude de la projection des gènes sur le plan déterminé par les axes 3 et 5 (le jour de culture et la source de soufre) permet d'identifier graphiquement les gènes dont l'expression varie en fonction de la source de soufre, mais pas de la date (figure 2). Ce sont les gènes dont la coordonnée sur l'axe 3 est proche de 0 (d'un jour sur l'autre, elle reste proche de l'expression moyenne) et qui sont éloignés du centre sur l'axe 5 (il existe deux niveaux d'expression bien distincts en fonction de la source de soufre).

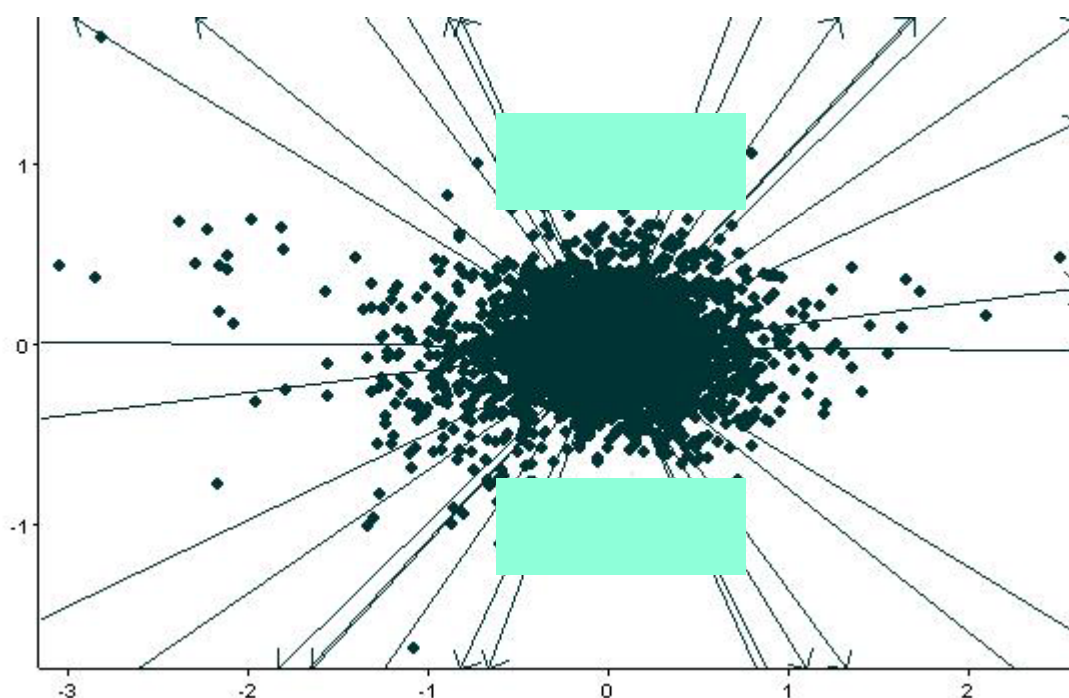
La corrélation exprime mathématiquement cette idée, indépendamment de la dimension du nuage des gènes : la corrélation d'un gène i avec un axe x_l est le rapport de la projection du gène sur cet axe (sa coordonnée $x_{i,l}$) sur d_i , la distance du gène i au centre de gravité du nuage

Tableau 1 : Matrice des vecteurs propres des données du transcriptome de *Bacillus subtilis*

Axes d'inertie	1	2	3	4	5	6	7
% d'inertie	94,64	2,26	1,01	0,56	0,36	0,29	0,26
metA1a	0,25	0,31	-0,16	-0,11	0,34	0,00	-0,31
metA1b	0,25	0,31	-0,17	-0,13	0,37	0,07	0,16
metB1a	0,25	0,18	0,12	0,36	0,12	0,31	-0,30
metB1b	0,25	0,18	0,11	0,37	0,15	0,43	0,21
metA10a	0,25	-0,26	-0,33	0,21	0,20	-0,43	-0,18
metA10b	0,25	-0,26	-0,33	0,19	0,26	-0,30	0,31
metB10a	0,25	-0,25	0,26	-0,23	0,13	0,09	-0,32
metB10b	0,25	-0,25	0,26	-0,24	0,18	0,18	0,17
mtrA1a	0,25	0,33	-0,08	-0,37	-0,22	-0,18	-0,17
mtrA1b	0,25	0,33	-0,09	-0,37	-0,19	-0,09	0,32
mtrB1a	0,25	0,16	0,22	0,31	-0,36	-0,29	-0,16
mtrB1b	0,25	0,15	0,21	0,32	-0,29	-0,21	0,26
mtrA10a	0,25	-0,24	-0,35	-0,03	-0,38	0,27	-0,29
mtrA10b	0,25	-0,24	-0,35	-0,03	-0,34	0,37	0,19
mtrB10a	0,25	-0,23	0,33	-0,13	0,00	-0,15	-0,20
mtrB10b	0,25	-0,23	0,33	-0,14	0,04	-0,06	0,31

Matrice des vecteurs propres obtenus par l'ACP. Ce tableau présente les coefficients des combinaisons linéaires qui permettent d'obtenir les coordonnées des axes d'inertie en fonction des coordonnées initiales. Les valeurs initiales ont été au préalable transformées en logarithmes centrés réduits sur les conditions expérimentales. Les coefficients sont grisés ou blancs pour représenter la signification des axes par rapport aux facteurs expérimentaux. L'axe 5 en bleu, discrimine les expériences selon la source de soufre.

Figure 2 : Projection des gènes de *Bacillus subtilis* sur le plan défini par les axes 3 et 5 de l'ACP



En bleu, figurent les gènes dont les niveaux d'expression varient essentiellement en fonction de la source de soufre et qui sont peu influencés par le jour de culture.

($d_i = \sqrt{x_{i,1}^2 + x_{i,2}^2 + \dots + x_{i,n}^2}$ avec $x_{i,2}, \dots, x_{i,n}$, les coordonnées du gène i sur les autres axes). La corrélation avec l'axe x_1 est proche de 1 si la coordonnée du gène i est proche de 0 sur tous les autres axes. La distance d_i dépend beaucoup de la valeur moyenne de l'expression du gène i , alors que ce paramètre n'a pas d'intérêt biologique ici. Afin de tourner cette difficulté, la distance est mesurée sans tenir compte de la coordonnée sur le premier axe d'inertie.

Les vingt gènes les plus fortement corrélés avec l'axe 5 figurent en gras dans la troisième colonne du tableau 9 (p. 22 et 23 bis), avec les gènes membres des même opérons. Il s'agit d'un témoin interne, l'idée sous-jacente étant que les gènes d'un opéron doivent avoir des variations coordonnées de leur expression et donc qu'ils doivent être corrélés aux même axes.

Certains gènes sur-exprimés dans les cultures en méthionine prennent un sens, quand on sait que l'arginine est nécessaire au catabolisme de la méthionine (cf. annexe) et que la culture est faite en milieu minimum. Il s'agit, bien sûr, du gène *carA* impliqué dans la biosynthèse de l'arginine. Mais aussi, et c'est une des découvertes faites avec cette expérience, de l'opéron *yqiXYZ* repéré grâce au gène *yqiY*. Cet opéron code pour des protéines ressemblant à des transporteurs ABC. Une série d'expériences de biologie moléculaire classique a montré que l'opéron *yqiXYZ* est nécessaire à l'entrée de l'arginine quand l'acide aminé est en très faible concentration (Sekowska *et al.* 2001).

Des gènes impliqués dans l'entrée du sulfate sont également dans les vingt gènes sélectionnés. Mais, globalement, l'absence de cohérence dans plus de 90 % des opérons trouvés (quinze sur seize) montre que la corrélation avec l'axe 5 n'est pas un critère fiable pour identifier les gènes dont l'expression varie en fonction de la source de soufre. En d'autres termes, l'axe 5 n'est pas celui qui discrimine le mieux les sources de soufre. Il n'y avait d'ailleurs pas de raison pour que l'axe discriminant soit parfaitement colinéaire avec un des axes d'inertie.

3.1.2.2 Le cycle circadien de la souris

Les résultats sont présentés dans le tableau 2 :

- Comme toujours, le nuage a une forme très allongée car le niveau d'expression moyen varie considérablement selon les gènes (l'inertie le long du premier axe représente 98,4 % de l'inertie totale). Il n'est pas pertinent pour le problème traité.
- Les axes les plus intéressants sont ceux sur lesquels la projection des répétitions est proche : seuls l'axe 3 et l'axe 6 correspondent à ce critère. L'axe 3 associe des valeurs positives aux expériences situées aux temps 0, 4 et 20 tandis que l'axe 6 associe des valeurs positives aux expériences 0, 12 et 16.

Ainsi, l'analyse de la projection sur le plan défini par les axes 3 et 6 devrait permettre d'identifier les gènes dont l'expression varie de façon reproductible au cours du temps. Ce sont les gènes les mieux corrélés avec le plan.

Tableau 2 : Matrice des vecteurs propres des données du transcriptome sur la souris

Axes d'inertie	1	2	3	4	5	6	7	8	9	10	11	12
% d'inertie	98,39	0,47	0,28	0,22	0,17	0,11	0,1	0,08	0,07	0,05	0,04	0,04
0A	0,29	-0,59	0,14	0,13	0,15	0,56	-0,03	0,02	0,05	-0,22	0,03	-0,38
0B	0,29	-0,01	0,29	0,03	-0,29	0,24	-0,07	-0,13	0,52	0,03	-0,02	0,63
4A	0,29	-0,01	0,30	0,13	-0,31	-0,11	-0,05	0,15	-0,56	-0,14	-0,58	0,09
4B	0,29	0,20	0,36	0,43	-0,10	-0,19	0,25	0,26	-0,07	0,01	0,60	-0,13
8A	0,29	0,25	-0,29	0,35	-0,10	-0,22	-0,50	-0,01	0,40	-0,15	-0,20	-0,36
8B	0,29	-0,45	-0,39	0,35	0,24	-0,29	-0,02	-0,11	-0,16	0,37	0,04	0,34
12A	0,29	0,32	-0,29	0,11	-0,03	0,21	0,60	-0,52	-0,06	-0,10	-0,13	-0,09
12B	0,29	-0,13	-0,50	-0,42	-0,50	0,01	0,09	0,37	-0,05	-0,15	0,23	0,03
16A	0,29	0,36	-0,07	-0,16	0,43	0,24	-0,42	-0,03	-0,35	-0,27	0,26	0,27
16B	0,29	0,23	0,00	-0,16	0,38	0,15	0,22	0,54	0,20	0,43	-0,32	-0,06
20A	0,29	0,02	0,22	-0,38	-0,19	-0,03	-0,25	-0,40	-0,12	0,57	0,14	-0,32
20B	0,29	-0,19	0,23	-0,40	0,31	-0,57	0,18	-0,14	0,21	-0,39	-0,05	-0,03

Matrice des vecteurs propres obtenus par l'ACP. Ce tableau présente les coefficients des combinaisons linéaires qui permettent d'obtenir les coordonnées des axes d'inertie en fonction des coordonnées initiales. Les valeurs initiales ont été au préalable transformées en logarithmes centrés réduits sur les conditions expérimentales.

Les coefficients des axes 3 et 6 sont grisés ou blancs afin de montrer la signification des axes par rapport aux facteurs expérimentaux.

Les chiffres 0, 4, 8, 12, 16 et 20 sont les différents temps de mesure. Celles-ci ont été répétées deux fois (A et B).

3.2 Une méthode de recherche des axes maximisant l'écart à la normalité : l'ACI

3.2.1 Principe de la méthode

Ces dernières années, les méthodes d'analyse du signal ont été bouleversées par l'apparition de l'analyse en composantes indépendantes (ACI) (Comon 1994). En revanche, elle n'est apparemment utilisée jusqu'ici que par deux groupes pour l'analyse du transcriptome (Liebermeister 2001, 2002, Chiapetta *et al.* 2002).

La méthode a été développée pour résoudre le problème dit de *la cocktail party*. Des signaux sont recueillis à partir de plusieurs microphones disposés à différents endroits d'une pièce. Le problème est d'isoler les conversations des différentes personnes présentes. L'analyse en composantes indépendantes recherche les sources (ici les conversations) supposées indépendantes.

La problématique peut très facilement être étendue au transcriptome : les sources sont des profils d'expression spécifiques d'une condition physiologique et les signaux sont les valeurs mesurées à chaque extraction d'ARN. Ils mélangent les profils de plusieurs conditions physiologiques (par exemple le profil « culture en méthionine », le profil « jour A » et le profil « synthèse des ADNc avec 1 µg d'ARN » dans les cas de l'expérience metA1 de *Bacillus subtilis*) et ces profils sont, en première approximation, à peu près indépendants puisqu'ils correspondent à peu près à des axes orthogonaux dans l'ACP.

Cette méthode est essentiellement une technique d'identification de modèle linéaire basée sur des statistiques d'ordre supérieur à deux : alors que l'ACP se base sur l'étude des corrélations (moments d'ordre deux), l'ACI utilise les moments d'ordre trois ou plus. Concrètement, si les données suivent une loi normale conjointe, l'étude par ACP extrait toute l'information contenue dans le nuage de points. Par contre, si tel n'est pas le cas, les statistiques d'ordre supérieur peuvent procurer une information pertinente.

Mathématiquement, l'ACI fait l'hypothèse que les Y_i , vecteurs correspondant aux mesures d'expression des gènes dans la condition i , sont obtenus par une combinaison linéaire de vecteurs sources S^m (qui sont les profils d'expression spécifiques que l'on cherche à identifier) :

$$Y_i = \sum_m A_i^m S^m .$$

Les coefficients de mélange A_i^m (matrice de mélange) sont inconnus. Les sources sont supposées statistiquement indépendantes, ce qui est une hypothèse plus forte qu'une absence de corrélation. Par exemple, les coordonnées de points répartis sur un cercle ne sont pas indépendantes ($x^2 + y^2 = C$) alors que le coefficient de corrélation linéaire entre x et y est nul.

3.2.2 Réalisation concrète

Une première transformation, dite de *whitening*, permet d'obtenir des variables centrées réduites décorrélées. La méthode classiquement utilisée est la décomposition selon les valeurs propres de la matrice des corrélations, exactement comme dans une ACP. Cette transformation facilite les calculs ultérieurs en permettant d'aborder directement les statistiques d'ordre supérieur à deux.

Plusieurs méthodes existent alors pour approcher les vecteurs S^m . Le logiciel utilisé (Hyvärinen *et al.* 1999, 2000) recherche les axes où la distribution s'écarte le plus d'une loi gaussienne en utilisant comme critère :

$$\max \left(E(\Phi(V)) - E(\Phi(\gamma)) \right)^2.$$

Avec V la projection des données sur l'axe recherché, γ une variable aléatoire gaussienne centrée réduite et Φ une fonction non quadratique. La fonction utilisée ici est $\Phi(x) = x^4$. Elle mesure l'aplatissement de la distribution par rapport à une loi gaussienne. Par exemple, une distribution paraîtra aplatie si elle comporte de nombreuses valeurs éloignées de la moyenne (« queues lourdes »).

Le logiciel utilisé repose sur une adaptation, par l'équipe de Bruno Torrèsani, du logiciel libre FastICA (sous Matlab) (Hyvärinen *et al.* 1999, 2000). Il faut fixer le nombre de sources indépendantes recherchées (qui est inférieur ou égal à la dimension de l'espace des expériences).

La méthode est une heuristique. Le risque de tomber sur des minima locaux existe. Afin de le limiter, la recherche des composantes indépendantes est lancée une centaine de fois, puis les axes stables sont identifiés de la façon suivante :

1. Les composantes indépendantes qui sont pratiquement colinéaires (produit scalaire supérieur à 0,9) constituent un faisceau.
2. La liste des p gènes les plus éloignés du centre du nuage est établie pour chaque composante indépendante du faisceau.
3. Un faisceau est stable si la liste des p gènes est à peu près la même dans toutes les composantes indépendantes.

Si la liste des p gènes diffère trop d'une composante indépendante à l'autre au sein d'un faisceau, celui-ci est difficile à interpréter et paraît souvent peu pertinent au biologiste. Ce phénomène peut avoir deux causes : le nombre de gènes dans cette région de l'espace est relativement important ou bien les gènes sont en réalité éloignés du faisceau. En revanche, un faisceau stable est une représentation robuste un profil d'expression type.

La distribution des gènes le long d'un axe est également visualisée par histogramme.

Par ailleurs, la distribution des gènes le long d'un axe s'écarte généralement d'une distribution normale par le biais de queues lourdes. Dans l'hypothèse où la distribution de la majorité des gènes suit une loi normale dans cette direction de l'espace, il est possible

d'associer une *p-value* aux gènes les plus distants, correspondant à la probabilité que le gène se situe à la position observée tout en ayant à la même loi de distribution que les autres.

3.2.3 Résultats obtenus

3.2.3.1 L'expérience sur *Bacillus subtilis*

Le nombre de vecteurs sources S^m est fixé à quatorze.

Les neuf faisceaux analysés sont ceux qui ont été retrouvés dans plus de 90 % des itérations, sept d'entre eux sont stables au sens du critère ci-dessus (tableau 3) parmi lesquels six ont un lien évident avec les facteurs expérimentaux (source de soufre, date de l'expérience, préparation de l'ADNc, duplicata). Il faut cependant noter que dans l'un des deux faisceaux qui séparent les duplicata a et b, la condition mtrB1b se situe au niveau des duplicata a (tableau 4 axe 6).

Les deux faisceaux instables (axes 7 et 8) sont inanalysables : la liste obtenue en prenant les vingt gènes les plus extrêmes de chaque composante indépendante, comporte plus d'une centaine de gènes, la plupart étant retrouvée dans moins de 20 % des cas.

Comme pour l'ACP, la cohérence des opérons apporte un témoin interne : les gènes d'un même opéron doivent être assez systématiquement corrélés à un même axe. En revanche, un axe caractérisé uniquement par des gènes isolés n'a probablement pas de signification biologique.

De fait, les faisceaux polarisés sur les duplicata et sur l'interaction jour / concentration en ARN (axes 1, 6 et 2) ne contiennent que des gènes isolés. C'est aussi le cas de l'axe dépourvu de lien avec les facteurs expérimentaux.

En revanche, le faisceau polarisé par l'interaction jour / soufre (axe 3) est caractérisé par un opéron codant pour des protéines ribosomales, un autre pour des protéines impliquées dans la synthèse de purines et quelques opérons de fonction inconnue. De même le faisceau séparant les expériences selon le jour (axe 4) est associé à des opérons impliqués dans le transport de l'ADN, le phénomène de compétence, la formation des flagelles ainsi que la formation de la paroi des spores. Ces deux exemples montrent qu'entre deux expériences réalisées avec la même souche de bactérie à deux jours différents, il y a eu un changement de physiologie des bactéries (Neidhardt *et al.* 1994). Et, bien que les cultures soient en phase exponentielle, on perçoit déjà si elles doivent évoluer vers la transformation ou vers la sporulation.

Enfin, dans le faisceau discriminant les deux sources de soufre (axe 6), cinq opérons ont été identifiés (tableau 9 quatrième colonne p. 22 et 23 bis). Deux sont impliqués dans le métabolisme de l'arginine, *argGHytzD* et *yqiXYZ* (*yqiX* à la vingt et unième position). Le gène *yqiY* présent dans ce dernier opéron a également été identifié par l'ACP. Cet opéron permet potentiellement l'entrée de l'arginine nécessaire au catabolisme de la méthionine. Un opéron impliqué dans le transport d'un antibiotique et notamment la formation de ponts disulfure, un opéron codant pour des lactate deshydrogénase et perméase et enfin certains gènes d'un opéron codant des protéines inconnues ont été également identifiés.

Tableau 3 : Bilan des différents faisceaux d'axes trouvés par l'ACI

	Faisceau 1	Faisceau 2	Faisceau 3	Faisceau 4	Faisceau 5	Faisceau 6	Faisceau 7	Faisceau 8	Faisceau 9
Stabilité du faisceau	x	x	x	x	x	x			x
Liens avec facteurs	x	x	x	x	x	x			
Opérons			x	x	x				

Ce tableau présente les propriétés des faisceaux d'axes trouvés par l'ACI au moins dans 90% des cas quand le calcul est répété : stabilité lorsque la liste des p gènes trouvés à chaque iteration est à peu près constante, liens avec les facteurs expérimentaux (soufre, jour, concentration et duplicata) et présence d'opérons parmi les vingt premiers gènes détectés.

Tableau 4 : Axes identifiés par l'ACI exprimés en fonction des conditions expérimentales (*Bacillus subtilis*)

	Axe 1 Duplicata	Axe 2 Jour /concentration	Axe 3 Jour /soufre	Axe 4 Jour	Axe 5 Soufre	Axe 6 Duplicata
% retrouvé	100	98	100	100	100	100
metA1a	-0,09	-0,19	0,26	0,07	-0,12	-0,27
metA1b	-0,29	-0,19	0,25	0,07	-0,12	0,26
metB1a	-0,04	0,22	0,27	-0,34	-0,19	-0,11
metB1b	-0,34	0,23	0,28	-0,31	-0,21	0,38
metA10a	0,00	0,32	0,26	0,28	-0,06	-0,06
metA10b	-0,31	0,33	0,27	0,28	-0,07	0,38
metB10a	0,13	0,04	0,29	-0,26	-0,21	-0,07
metB10b	-0,44	0,04	0,30	-0,26	-0,21	0,24
mtrA1a	-0,15	-0,30	-0,01	-0,12	0,07	-0,31
mtrA1b	-0,24	-0,29	0,00	-0,13	0,05	0,45
mtrB1a	-0,05	0,20	0,31	-0,31	0,39	-0,18
mtrB1b	-0,33	0,22	0,30	-0,33	0,39	-0,06
mtrA10a	-0,06	0,42	-0,03	0,19	0,06	-0,09
mtrA10b	-0,32	0,42	-0,02	0,20	0,06	0,14
mtrB10a	0,09	0,04	0,32	-0,29	0,51	0,14
mtrB10b	-0,42	0,02	0,34	-0,29	0,47	0,33

Matrices des coefficients des combinaisons linéaires qui permettent d'obtenir les axes de l'ACI (archétype d'un faisceau) en fonction des vecteurs initiaux. Seuls les faisceaux d'axes pouvant être expliqués par les facteurs expérimentaux ont été décrits ainsi que le nombre de fois où ils ont été retrouvés lorsque le calcul a été répété (exprimé en pourcentage).

3.2.3.2 L'expérience sur la souris

La recherche des composantes indépendantes est faite avec huit vecteurs sources S^m . Elle a été répétée cent fois. Quatre faisceaux ont été retrouvés dans plus de 90 % des cas. Deux sont polarisés sur les temps et regroupent les deux préparations d'ADNc alors que les deux autres faisceaux les séparent (tableau 5).

Les deux faisceaux séparant les préparations d'ADNc ne sont pas stables au sens du critère ci-dessus. Ils ne sont pas non plus interprétables avec les facteurs identifiés dans l'expérience.

Les faisceaux 3 et 4 sont stables. Le faisceau 3 est surtout polarisé par les expériences à $t = 0$ et $t = 20$ tandis que le faisceau 4 sépare les expériences à $t = 8$ et 12 des autres. Ensemble, ils permettent d'identifier des gènes dont l'expression varie au cours du cycle circadien.

Parmi les gènes identifiés, entre un tiers (axe 3) et la moitié (axe 4) varie de la même manière dans les deux répétitions (figure 3a). Pour les autres, les variations diffèrent soit en ordre de grandeur soit de sens entre les deux répétitions (figure 3b).

Après avoir identifié les gènes extrêmes de l'axe, il est donc nécessaire de vérifier que les deux répétitions donnent des résultats cohérents. Cette vérification peut se faire par une simple représentation graphique.

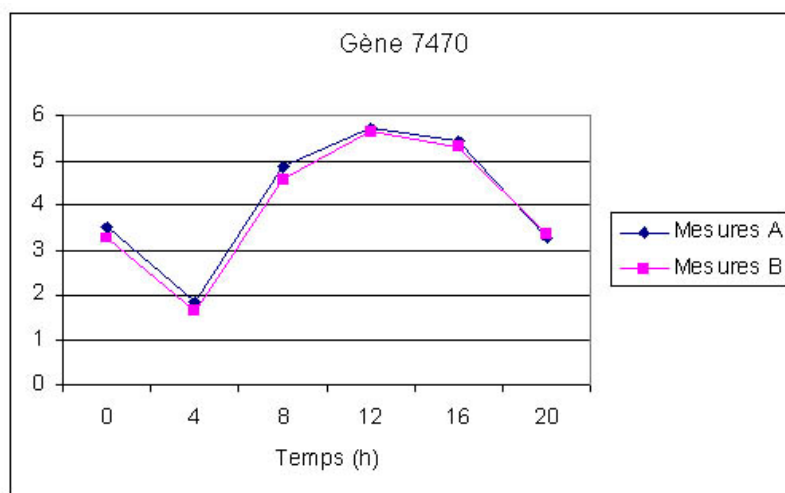
Tableau 5 : Axes identifiés par l'ACI exprimés en fonction des conditions expérimentales (souris)

	Axe 1	Axe 2	Axe 3	Axe 4
% retrouvé	93	93	100	92
0A	-0,285	0,272	-0,611	-0,239
0B	-0,293	0,286	-0,650	-0,200
4A	-0,282	0,324	-0,063	-0,265
4B	-0,303	0,119	-0,053	-0,262
8A	-0,300	0,458	-0,053	0,157
8B	-0,280	0,288	-0,060	0,135
12A	-0,295	0,087	-0,042	0,086
12B	-0,277	0,339	-0,002	0,065
16A	-0,295	0,272	-0,075	-0,334
16B	-0,292	0,287	-0,035	-0,523
20A	-0,288	0,283	-0,318	-0,415
20B	-0,274	0,270	-0,285	-0,390

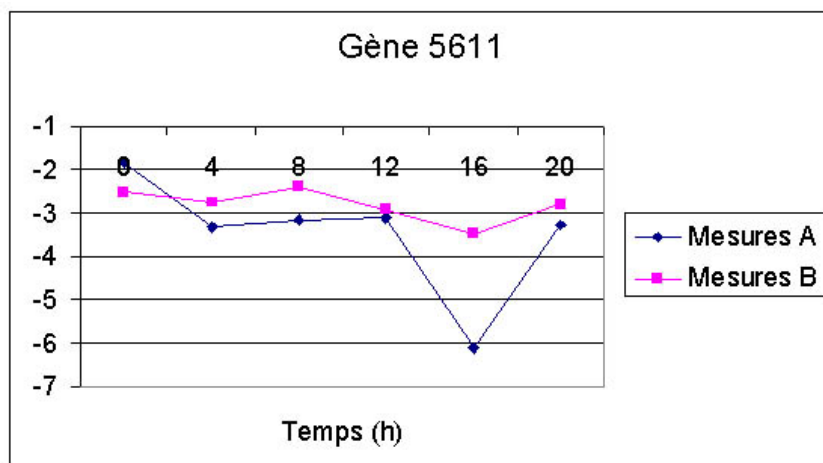
Matrices des coefficients des combinaisons linéaires qui permettent d'obtenir les axes de l'ACI (archétype d'un faisceau) en fonction des vecteurs initiaux et nombre de fois où ces axes ont été retrouvés (exprimé en pourcentage). Les chiffres 0, 4, 8, 12, 16 et 20 sont les différents temps de mesure. Celles-ci ont été répétées deux fois (A et B). Seuls les axes 3 et 4 regroupent les répétitions.

Figure 3 : Exemples de profils d'expression de gènes identifiés par l'ACI

a: Le gène 7470 possède un profil d'expression comparable dans les deux répétitions et varie en fonction du temps (identifié sur l'axe 4)



b: Le gène 5611 également détecté grâce à l'ACI n'a pas le même profil d'expression dans les deux répétitions (identifié sur l'axe 3).



4 Méthodes dirigées

4.1 L'analyse de la variance – ANOVA

Les méthodes utilisées au chapitre précédent donne une image synthétique des proximités des gènes dans l'espace des conditions expérimentales en les projetant sur des axes définis par des propriétés géométriques du nuage (axes d'inertie, axes maximisant des écarts à la distribution gaussienne). Ce sont des méthodes d'analyse multidimensionnelle (*multidimensional scaling*). L'expérimentateur les utilise pour explorer les données et afin d'y trouver, *a posteriori*, une organisation compréhensible.

L'analyse peut aussi partir de facteurs définis *a priori* par l'expérimentateur. Cette approche facilite l'interprétation des observations et permet d'introduire des tests statistiques (et, par là, les *p-value* réclamées par les biologistes).

4.1.1 Exposé de la méthode

L'analyse de la variance (l'ANOVA) (Fisher 1954) permet d'évaluer l'importance relative de différents facteurs sur les fluctuations observées dans un ensemble de conditions expérimentales. Elle permet aussi de déceler les interactions entre les facteurs. Enfin, moyennant certaines hypothèses, il est possible d'attribuer une signification statistique à l'impact des facteurs sur la variable mesurée (la *p-value*).

Dans le cas d'une analyse de transcriptome, le problème se présente schématiquement de la façon suivante. L'expérimentateur identifie les facteurs susceptibles d'influer sur le niveau d'expression des gènes et fixe le nombre de catégories pour chaque facteur. Puis, dans le cas idéal, l'expérience est conçue de telle sorte que toutes les combinaisons de catégories sont réalisées (l'analyse reste possible si certaines combinaisons manquent, mais au prix d'une perte de puissance).

Par exemple, l'expérience réalisée avec *Bacillus subtilis* isole cinq facteurs susceptibles d'influer sur le niveau d'expression Y_{ijklm} mesuré pour un gène donné dans une condition expérimentale donnée : le gène lui-même G_i , la source de soufre S_j , le jour J_k , la préparation de l'ADNc C_l et le duplicata D_m . Le facteur *gène* comprend 4107 catégories (autant que de gènes dans le génome) alors qu'il n'y a que deux catégories dans les autres facteurs. La figure 1 montre que toutes les combinaisons ont été testées et les résultats de l'expérience remplissent un vaste tableau de 4107 lignes et seize colonnes, chaque ligne correspondant à un gène et chaque colonne à une condition expérimentale.

Le tableau général peut être décomposé en sous-tableaux résumant les observations pour chaque facteur. Par exemple, le tableau *gène* est un vecteur de 4107 cases construit en additionnant, pour chaque gène, les valeurs observées dans les seize conditions expérimentales. De la même façon, des regroupements *ad hoc* permettent d'obtenir des tableaux associant plusieurs facteurs pour analyser leurs interactions. Par exemple, le tableau à *gène* \times *soufre* permet d'identifier les gènes dont l'expression varie en fonction de la source de soufre. C'est un tableau à 4107 lignes et deux colonnes. Et le tableau à trois dimensions

gène × soufre × jour permettra de rechercher les gènes dont l'expression est influencée à la fois par le changement de source de soufre et la date de la culture.

Si l'on s'en tient à l'action des facteurs isolés, le niveau d'expression Y_{ijklm} d'un gène dans l'expérience sur *Bacillus subtilis* est donné par l'équation suivante :

$$Y_{ijklm} = \mu + G_i + S_j + J_k + C_l + D_m + \varepsilon_{ijklm}$$

avec ε_{ijklm} , la variation du niveau d'expression non expliquée par les cinq facteurs pris isolément, et μ , le niveau d'expression moyen des gènes sur l'ensemble de l'expérience. ε_{ijklm} est appelé aussi bruit, erreur ou variance résiduelle ; il peut être décomposé à son tour si on introduit les interactions entre les facteurs.

Mathématiquement, la grandeur analysée est la somme des carrés des écarts à la moyenne (la variance) et l'équation ci-dessus signifie que la variance globale de l'expression des gènes est la somme des variances de chaque facteur et d'une variance résiduelle correspondant notamment aux interactions qui ne sont pas prises en compte. On considère qu'un facteur joue un rôle significatif quand sa variance est très supérieure à la variance résiduelle. Si les hypothèses de normalité et d'égalité des variances sont vérifiées, le test de Fisher permet d'associer une probabilité (la *p-value*) à la notion de « significatif ».

L'analyse peut être faite gène par gène. L'équation ci-dessus devient alors :

$$Y_{jklm} = \mu + S_j + J_k + C_l + D_m + \varepsilon_{jklm}$$

Elle est appelée « ANOVA locale » dans le logiciel GeneANOVA (Didier *et al.* 2002).

Du point de vue du biologiste, un bon gène est un gène dont l'expression varie pour le facteur qui l'intéresse (ici la source de soufre), mais qui varie de façon reproductible. Ceci se traduit mathématiquement par une valeur élevée de S_j et une faible valeur de ε_{jklm} , due notamment à l'absence d'interactions entre le facteur intéressant et les autres facteurs.

Afin de faciliter la sélection, GeneANOVA donne une représentation graphique de l'ensemble des gènes, chacun ayant pour abscisse S_j et pour ordonnée la *p-value* associée au rapport S_j / ε_{jklm} . Les meilleurs gènes sont localisés en bas à droite de la figure, dans la zone des fortes valeurs de S_j et des faibles valeurs de la *p-value*.

Le rapport S_j / ε_{jklm} et l'estimation de la *p-value* associée pose plusieurs problèmes :

- L'estimation des variances est imprécise car le nombre d'expériences est réduit. L'imprécision touche tout particulièrement la variance résiduelle ε_{jklm} car elle est calculée avec peu de degrés de liberté. Il s'ensuit d'importantes fluctuations aléatoires de la *p-value*.
- Il est impossible de savoir si les hypothèses qui fondent le calcul de la *p-value* dans le test de Fisher sont vérifiées dans l'analyse du transcriptome.
- La définition d'un seuil de décision est difficile car une expérience génère plusieurs milliers de tests (un par gène). La correction de Bonferroni n'est pas une solution (Nadon *et al.* 2002) car on constate empiriquement qu'elle conduirait à négliger

beaucoup de gènes ayant une réelle signification biologique (résultats ne figurant pas dans le mémoire).

Tout ceci conduit en pratique à considérer que la *p-value* est un indice de qualité plutôt qu'un degré signification statistique.

4.1.2 Résultats obtenus

Aucune transformation autre que le passage aux logarithmes n'est nécessaire avant l'ANOVA. Cependant, les données ont été centrées réduites afin de permettre une comparaison directe avec les analyses multidimensionnelles (ACP, ACI) car celles-ci débutent systématiquement par cette transformation.

4.1.2.1 L'expérience sur *Bacillus subtilis*

Le tableau 6 récapitule l'impact des différents facteurs et leurs interactions :

1. La quantité d'ARN utilisée pour synthétiser les ADNc est le facteur qui modifie le plus la mesure des niveaux d'expression. Il existe une forte interaction avec le facteur *gène*, ce qui signifie que la relation entre la quantité d'ARN et la quantité d'ADNc n'est pas la même pour tous les gènes. La conséquence, en pratique, est qu'il n'est pas possible d'être sûr du niveau d'expression relatif des gènes à partir de l'analyse du transcriptome.
2. L'ANOVA trouve ensuite une forte interaction entre les facteurs *gène* et *jour* qui montre que les bactéries n'étaient pas exactement dans le même état physiologique les deux fois.
3. Comme attendue, l'interaction entre les facteurs *gène* et *soufre* est faible car peu de gènes sont impliqués par le changement de source de soufre.
4. Enfin, on observe une interaction faible, mais significative, entre les facteurs *gène* et *duplicata*. Elle montre que l'imprécision des duplicata est plus importante que les erreurs non identifiées, sans toutefois atteindre le niveau des différences ayant une signification biologique.

Ces conclusions recoupent totalement celles déduites de l'importance relative des axes d'inertie obtenue par l'ACP.

Une ANOVA gène par gène (ou ANOVA locale) permet d'identifier les gènes dont le niveau d'expression change de façon reproductible quand la méthionine est substituée au méthylthioribose. Le tableau 7 donne les seize premiers gènes (en gras) classés par *p-value* croissante (*p-value* inférieure à $4 \cdot 10^{-5}$). On remarque que plusieurs gènes appartiennent aux mêmes opérons. Le fait que plusieurs gènes d'un opéron varient dans le même sens est attendu par le biologiste. C'est, comme on l'a déjà précisé, en quelque sorte un témoin interne de la validité de l'expérience.

Tous les gènes des opérons *yqiXYZ*, *argCJBDcarABargF*, *argGHytzD*, *lctEP*, *ahpCF* et le début de l'opéron *levDE* ont une *p-value* d'au plus 10^{-3} . En général, le facteur *soufre* est responsable d'une part relativement importante de la somme des carrés totale Y_{jklm} (il compte pour plus de la moitié de la somme dans 17 cas sur 21).

Tableau 6 : Résultats de l'ANOVA sur l'expérience sur *B. subtilis*

Facteur	Somme des carrés	DDL	Variance	F	P-value
Gène	62 177	4 106	15,14	648,42	0
Soufre	0	1	0	0	-
Jour	0	1	0	0	-
Concentration	0	1	0	0	-
Duplicata	0	1	0	0	-
Gène Soufre	229	4 106	0,06	2,39	0
Gène Jour	640	4 106	0,16	6,67	0
Gène Concentration	1 427	4 106	0,35	14,88	0
Gène Duplicata	168	4 106	0,04	1,76	0
Soufre Jour	0	1	0	0	-
Soufre Concentration	0	1	0	0	-
Soufre Duplicata	0	1	0	0	-
Jour Concentration	0	1	0	0	-
Jour Duplicata	0	1	0	0	-
Concentration Duplicata	0	1	0	0	-
Résiduel	1 055	45 171	0,02	-	-
Total	65 696	65 711	1	-	-

Somme des carrés, degré de libertés (DDL) et variance due au soufre (variance) des données transformées en logarithme centré réduit des conditions expérimentales.

Tableau 7: Opérons dont au moins un gène a été trouvé par ANOVA

Fonction des gènes de l'opéron	Nom des gènes	Localisation	Sens de lecture	Variance	F	% SC due au soufre	P-value
Similaire à des transporteurs d'acides aminés	yqiZ	2490,5	-	1,77	108,93	67	0
	yqiY	2491	-	0,91	65,99	77	1.10⁻⁵
	yqiX	2492	-	1,70	110,89	77	0
Transformation du glutamate en citrulline	argC	1194.3	+	0,3	11,5	39	6.10 ⁻³
	argJ	1195.4	+	1,73	46,54	57	3.10⁻⁵
	argB	1196.6	+	0,28	52,7	37	2.10⁻⁵
	argD	1197.4	+	0,91	59,72	81	1.10⁻⁵
	carA	1198.6	+	0,1	15,69	53	2.10 ⁻³
	carB	1199.7	+	0,13	20,36	39	9.10 ⁻⁴
	argF	1202.9	+	1,35	33,89	62	1.10 ⁻⁴
Transformation citrulline en arginine	ytzd	3010.8	-	1,91	79,66	69	0
	argH	3012.2	-	2	61,44	69	1.10⁻⁵
	argG	3013.4	-	0,71	77,03	77	0
Phosphotransférase (PTS) spécifique du fructose	levD	2762,1	-	1,13	94,98	84	0
	levE	2761,6	-	0,41	86,24	76	0
	levF	2761,1	-	0,01	1	3	-
	levG	2760,3	-	0,04	4,04	8	7.10 ⁻²
	sacC	2759,3	-	0,17	9,21	30	1.10 ⁻²
Akyl hydroperoxide reductase	ahpC	4118	+	0,65	18,21	57	1.10 ⁻³
	ahpF	4118.6	+	1,4	94,11	69	0
Lactate perméase et deshydrogénase	lctE	329.3	+	5,75	48,59	43	2.10⁻⁵
	lctP	330.3	+	0,9	21,77	48	6.10 ⁻⁴
Diverses enzymes métaboliques	mtrA	2384,6	-	0,02	1,7	2	2.10 ⁻¹
	mtrB	2384	-	0,02	1,84	10	2.10 ⁻¹
Synthèse de la ménaquinone	gerCA	2383,6	-	0,04	6,54	10	3.10 ⁻²
	gerCB	2382,8	-	0,05	98,47	14	0
	gerCC	2382,2	-	0,03	3,61	7	8.10 ⁻²
	ndk	2381	-	0	0	0	-
Gènes isolés							
Fonction inconnue	yycS	4134.4	+	0,14	52,48	64	2.10⁻⁵
Synthèse méthionine	metC	1384,9	-	2,65	46,24	54	3.10⁻⁵

Résultats des ANOVA locales avec la fonction des opérons, le nom des gènes et leur localisation, la variance de leur niveau d'expression due au soufre, le pourcentage des sommes de carrés totale dues au soufre et la p-value. Les données ont été au préalable transformées en logarithme centré réduit selon les conditions expérimentales. En gras figurent les gènes retrouvés parmi les seize premiers.

L'opéron *mtrABgerCABCndk* semble faire exception puisque seul le gène *gerCB* a une *p-value* inférieure à $3 \cdot 10^{-2}$, sans pour autant compter beaucoup dans la somme des carrés totale Y_{jklm} . *gerCB* est probablement l'exemple d'un faux positif dû à l'imprécision de l'estimation de la variance résiduelle ε_{jklm} . L'expression des gènes de la fin de l'opéron *levDEFGsacC* ne semble pas dépendre de la source de soufre.

4.1.2.2 L'expérience sur la souris

L'expérimentateur a isolé trois facteurs, le facteur *gène* (avec 12 487 catégories), le facteur *temps* (avec six catégories : 0, 4, 8, 12, 16 et 20 heures) et le facteur *répétition* de la préparation de l'ADNc (avec deux catégories : A et B).

Les résultats de l'ANOVA (tableau 8) montrent l'importance du facteur *gène*. Elle n'apprend rien au biologiste puisqu'elle résulte de la grande diversité des niveaux d'expression des gènes, indépendamment des conditions expérimentales. Une information plus intéressante est l'interaction, faible mais significative, entre les facteurs *gène* et *temps* car elle montre que l'expression d'un petit nombre de gènes varie au cours du cycle circadien. Enfin, il n'existe pas d'interaction significative entre les facteurs *gène* et *répétition*, ce qui montre que le protocole de préparation de l'ADNc est suffisamment maîtrisé eu égard aux autres sources d'erreur.

L'ANOVA locale permet d'identifier les gènes dont l'expression varie au cours du temps (quatre-vingt dix possèdent une *p-value* inférieure à 10^{-4} et vingt-cinq présentent une *p-value* inférieure à 10^{-5}). Il sont analysés dans la partie 5 du rapport (La représentation des profils d'expression).

Tableau 8 : Résultat de l'ANOVA globale sur les données d'expression sur la souris

Facteur	Somme des carrés	DDL	Variance	F	P-value
Gène	147 418	12 486	11,81	829,65	0
Répétition	0	1	0	0	-
Temps	0	5	0	0	-
Gène Répétition	133	12 486	0,01	0,75	-
Gène Temps	1 393	62 430	0,02	1,57	0
Répétition Temps	0	5	0	0	-
Résiduel	888	62 430	0,01	-	-
Total	149 832	149 843	1	-	-

Somme des carrées, degré de libertés (DDL) et variance due au temps (variance) des données sur l'étude du cycle circadien. Les données ont été transformées en logarithme centré réduit selon les conditions expérimentales.

Figure 4 : Données analysées par régression PLS

Mesures d'expression des gènes de la puce : X		Gènes théoriques : Y	
		Y _{met}	Y _{mtr}
Met		1	0
		1	0
	
		1	0
Mtr		0	1
		0	1
	
		0	1

4.2 PLS

4.2.1 Exposé de la méthode

La PLS (régression par la méthode des moindres carrés partiels ou *Partial Least Square regression*) est une méthode d'analyse multidimensionnelle (Bry X. 1996). Elle est conçue pour analyser des tableaux de données comprenant deux blocs de variables, des variables explicatives X et des variables expliquées Y. L'algorithme cherche à visualiser les relations entre les variables explicatives X et les variables expliquées Y, tout en conservant une bonne représentation des deux groupes de variables. Mathématiquement, ceci revient à rechercher un système d'axes qui maximise tout à la fois la covariance entre les X et les Y et les variances des X et des Y.

Dans le cas de la PLS discriminante utilisée ici, les variables Y sont des variables indicatrices des différentes catégories pour le facteur analysé. On peut les considérer comme des « gènes théoriques » dont l'expression est binaire. Par exemple, le « gène théorique » Y_{met} caractéristique de la méthionine aura pour valeur 1 quand la méthionine a été utilisée comme source de soufre et 0 quand c'est le méthylthioribose, alors que Y_{mtr} aura les valeurs 0 et 1 (figure 4). Le système s'applique à un nombre quelconque de catégories et permet de coder les facteurs directement dans le tableau de données. En ce sens, la PLS discriminante se rapproche de l'ANOVA. D'un point de vue géométrique, l'espace Y est particulièrement simple puisqu'il est polarisé par le facteur étudié. L'interprétation de la projection de X sur Y ne pose donc pas de problème, d'autant que la PLS fait ressortir les corrélations entre les gènes X.

La PLS est utilisée principalement en chimie pour l'analyse des résultats de chromatographie et de spectrographie. Elle a été appliquée également à l'analyse du transcriptome pour des expériences sur des cellules tumorales (Alaiya *et al.* 2000, Musumara *et al.* 2001, Nguyen *et al.* 2002).

4.2.2 Résultats obtenus

4.2.2.1 L'expérience sur *Bacillus subtilis*

L'expérience acquise, au cours du stage, sur la régression PLS discriminante est moins importante que pour les autres méthodes citées. L'évaluation a porté uniquement sur l'expérience sur *Bacillus subtilis*.

La liste des vingt gènes les plus corrélés avec le premier axe (celui qui discrimine la source de soufre) figure tableau 9 dernière colonne (p. 22 et 23) :

- Une partie concorde bien avec les résultats de l'ANOVA (les opérons *yqiXYZ* et *levDE*).

- La correspondance est moins bonne pour les opérons *argGHytzD* et *argCJBDcarABargF*.
- Les autres opérons identifiés ne contiennent qu'un seul gène dans les vingt premiers et sont de fonction inconnue ou éloignée du métabolisme du soufre. L'opéron *ytmAB* est même composé d'un gène plutôt exprimé en présence de méthylthioribose et d'un autre exprimé en présence de méthionine.
- Enfin, certains gènes ne semblent pas faire partie d'un opéron. Ces gènes codent généralement des protéines de fonction inconnue.

Sans être inintéressants, ces premiers résultats ne sont pas totalement satisfaisants. D'autres utilisations possibles de la PLS sont discutées dans la partie 7 (Améliorations envisageables).

5 La représentation des profils d'expression

L'ANOVA identifie les gènes dont l'expression répond de façon reproductible aux modifications d'un facteur donné. Mais elle ne permet pas de savoir s'ils se comportent de la même façon, s'ils ont le même *profil d'expression*.

Cette question du profil d'expression est particulièrement pertinente dans l'étude du cycle circadien chez la souris car le problème biologique dépasse la simple sélection des gènes. L'objectif est de les classer en fonction du nombre de cycles par jour et d'établir la façon dont leur activité s'enchaîne au cours du temps.

Le problème peut être résolu en combinant deux méthodes : une classification hiérarchique, pour regrouper les gènes ayant le même profil d'expression, et une ACP, pour déterminer les périodicités et les décalages de phase.

5.1 La classification hiérarchique

5.1.1 Exposé de la méthode

Des gènes auront le même profil d'expression s'ils ont à peu près le même niveau d'expression dans toutes les conditions expérimentales. Ils auront donc à peu près les mêmes coordonnées et ils seront voisins dans l'espace. Cette notion peut être élargie au cas où les gènes varient de la même façon dans toutes les conditions expérimentales, tout en n'ayant pas le même niveau moyen d'expression. Ils seront voisins dans le sous-espace défini par l'ACP après avoir éliminé le premier axe, puisque celui-ci correspond au niveau moyen d'expression. Il est possible, enfin, de donner la même variance à tous les gènes afin de ne pas tenir compte des différences dans l'ampleur des réponses aux changements de conditions. Dans tous les cas, la recherche de gènes ayant les mêmes profils d'expression se ramène à identifier des gènes voisins dans l'espace.

La question posée est plus simple que la représentation exacte de la géométrie puisqu'elle se limite à la recherche d'amas dans le nuage des gènes. Le travail est encore facilité si on se limite aux gènes pertinents (sélectionnés par ANOVA par exemple) car il y a de bonnes chances qu'ils se regroupent en un petit nombre d'amas.

Les méthodes de classification automatique sont faites pour traiter ce type de problème. Elles sont très nombreuses, chacune ayant sa propre façon de constituer les amas. L'analyse a été faite ici avec une classification hiérarchique en prenant le critère de Ward (les points sont agglomérés un à un et, à chaque étape, le point est associé à l'amas le plus proche, la proximité étant mesurée par la variance de l'ensemble point + amas). Le résultat final est représenté par un dendrogramme, la longueur des branches correspondant à la proximité des amas.

5.1.2 Résultats obtenus

5.1.2.1 L'expérience sur la souris

La similitude des profils d'expression des gènes ayant une p -value inférieure à 10^{-4} dans l'ANOVA a été déterminée en supprimant les différences du niveau moyen d'expression (en éliminant le premier axe de l'ACP) et d'amplitude de variation (en les ramenant tous à la même variance).

La classification automatique fait ressortir plusieurs groupes de profils d'expression (figure 5). Cependant, il est parfois difficile d'en définir clairement le nombre et la limite. Il faut vérifier graphiquement la cohérence du groupe. Il existe aussi des gènes relativement isolés, repérables à la longueur des branches. Finalement, dix gènes sont restés isolés et les autres répartis en dix-sept groupes.

La difficulté suivante est de fixer le profil d'expression typique de chaque groupe. Une solution est de choisir un gène existant réellement, l'autre de créer un archétype en prenant la médiane du groupe. La première solution est plus explicite pour quelqu'un connaissant bien les gènes, la seconde est meilleure aux yeux d'un statisticien car elle diminue le bruit.

L'objectif final est de construire une synthèse claire de l'activité périodique des gènes. L'expérience montre que ceci reste difficile si on ne possède que ce type d'outil.

5.2 L'analyse des séries temporelles par ACP

5.2.1 Exposé de la méthode

L'ACP a été utilisée dans la seconde partie du mémoire pour projeter le nuage des gènes sur les plans définis par les axes d'inertie, la représentation graphique de la matrice des vecteurs propres (la sphère des corrélations) ayant servi principalement à déterminer la signification des axes. Cependant, l'ACP peut apporter une information beaucoup plus riche quand elle est appliquée à l'analyse de séries temporelles (Holter *et al.* 2000). Quelques rappels sur les fonctions périodiques permettent de comprendre pourquoi.

La fonction $\sin(t)$ est la fonction périodique la plus simple, d'ailleurs n'importe quelle fonction périodique se ramène à une somme de fonctions sinusoïdales. La fonction $\sin(t)$ peut être représentée de deux façons, soit déroulée le long de l'axe du temps t , soit par l'ordonnée d'un point parcourant un cercle. Une période correspond à un tour complet du cercle. Deux points séparés dans le temps occupent des positions différentes sur le cercle (sauf si l'écart est un multiple entier de périodes). L'écart est mesuré par l'angle entre les deux points. Deux points diamétralement opposés sont en opposition de phase ($\sin(t_1) = -\sin(t_2)$). Deux points décalés de 90° varient la moitié du temps dans le même sens et en sens opposé l'autre moitié. Dans le vocabulaire de l'ACP, le cosinus de l'angle des deux points est le coefficient de corrélation linéaire et le cercle, le cercle des corrélations. Par exemple, un décalage de 90° correspond à une corrélation nulle. La corrélation peut aussi être nulle car les deux points ne varient pas au même rythme, l'un ayant, par exemple, une période deux fois plus courte que

l'autre. Ces points parcourent des cercles orthogonaux ; la sphère des corrélations est dans un espace à trois dimensions.

En pratique, l'ACP est réalisée sur les gènes les plus caractéristiques afin de limiter autant que possible le bruit car il crée des dimensions sans rapport avec la périodicité du phénomène. Les gènes dont l'expression varie significativement au cours du temps ont été identifiés au préalable par l'ANOVA, et regroupés en classes par classification automatique.

5.2.2 Résultats obtenus

L'ACP a été réalisée sur les vingt-sept gènes sélectionnés à l'issue de l'ANOVA et de la classification hiérarchique (chapitre 5.1.2.1).

Les gènes ayant une période voisine de la durée de l'expérience sont en bordure du cercle des corrélations, sur le premier plan (figure 6). On y retrouve les archétypes des dix-sept groupes définis ci-dessus. La figure permet également de déterminer l'ordre dans lequel s'expriment les gènes (figure 8). Par exemple, les gènes des groupes 2 et 14 ont un décalage de phase d'un peu plus de cinq heures et les gènes des groupes 2 et 16 sont en opposition de phase (figure 6 et 8).

Les gènes bien représentés dans le plan 1-4 ont une période deux fois plus réduite. Parmi eux figurent les gènes du septième groupe, en opposition de phase avec le gène 8340 (figure 7).

Il faut interpréter avec prudence les observations sur l'enchaînement de l'activité des gènes. Ce sont des corrélations, pas des relations causales. Le fait qu'un gène s'exprime avant un autre n'indique pas nécessairement que le premier déclenche l'activité du second.

Figure 5 : Dendrogramme des gènes dont l'expression varie au cours du temps dans l'étude du transcriptome de la souris

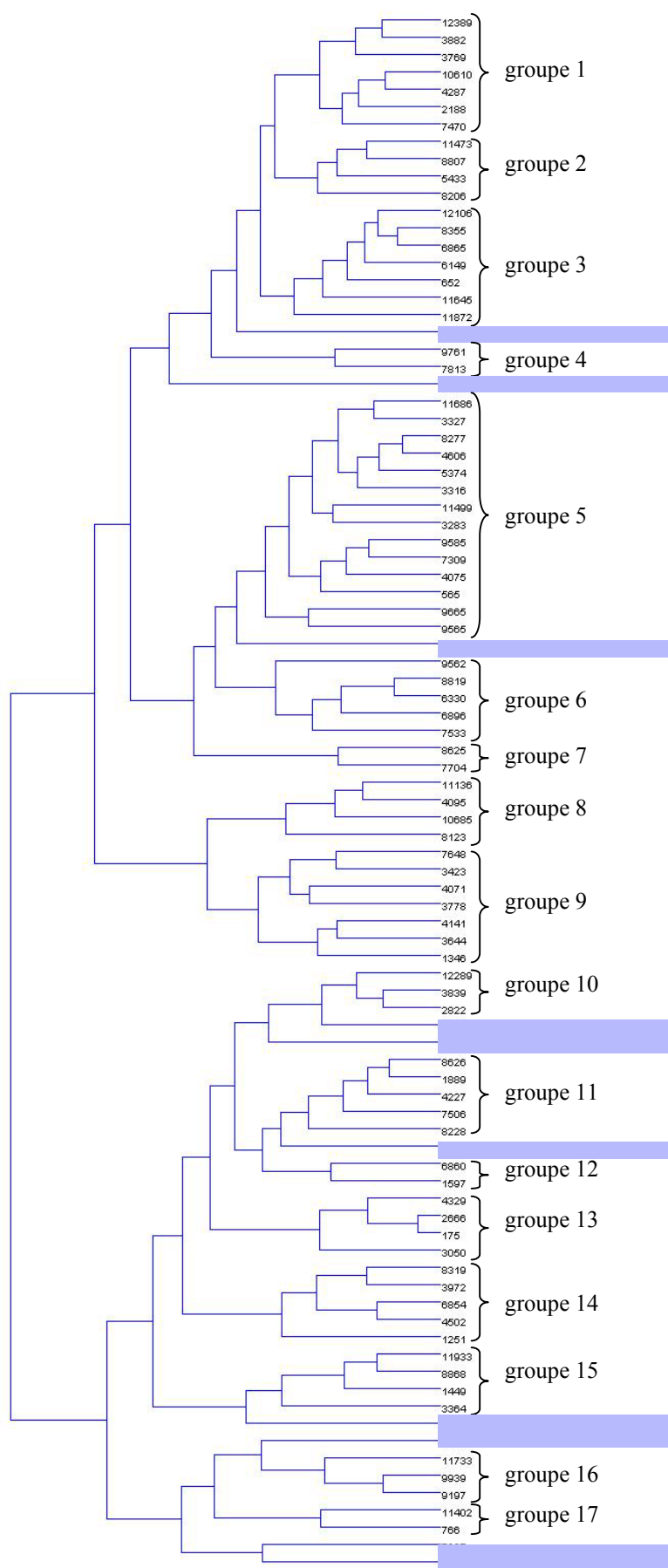


Figure 6 : Projection des gènes représentant les différents groupes sur les deux premiers axes de l'ACP

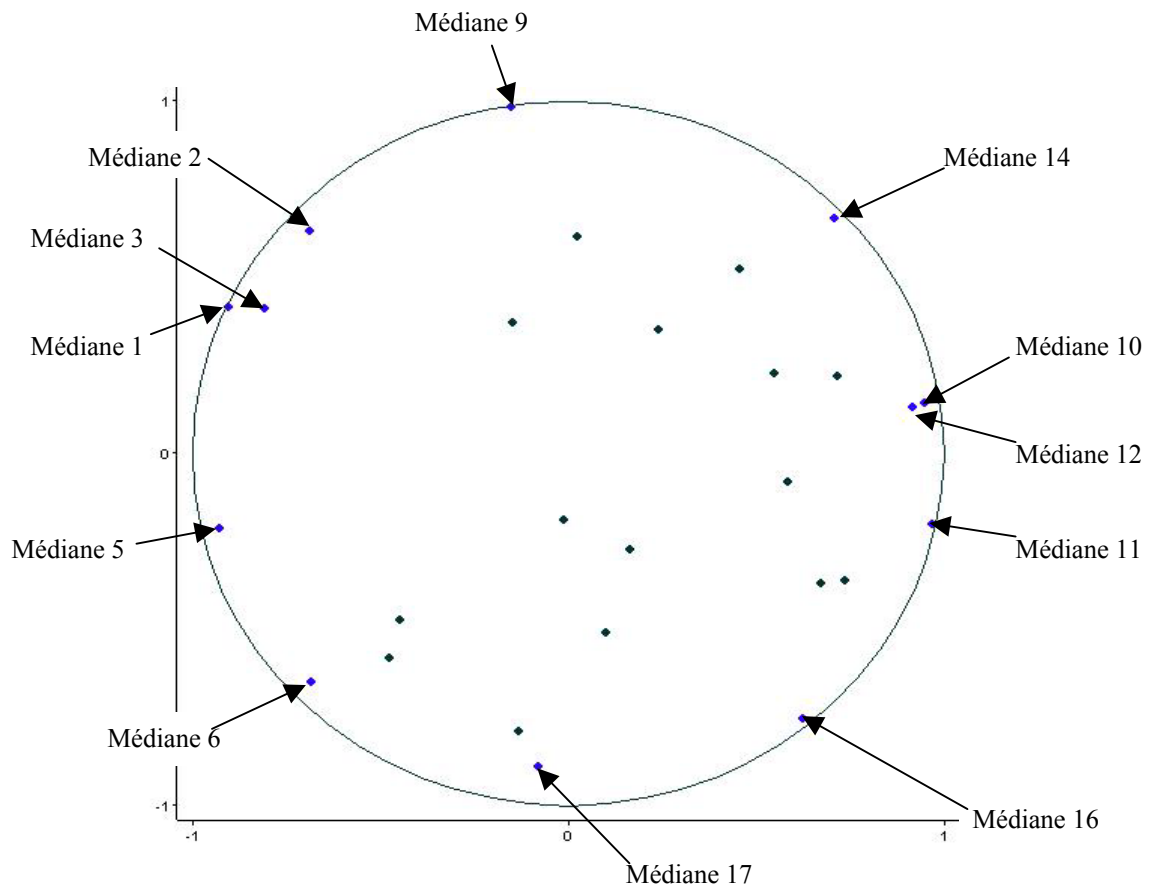


Figure 7: Evolution des niveaux d'expression de gènes dont la période est inférieure à 24H

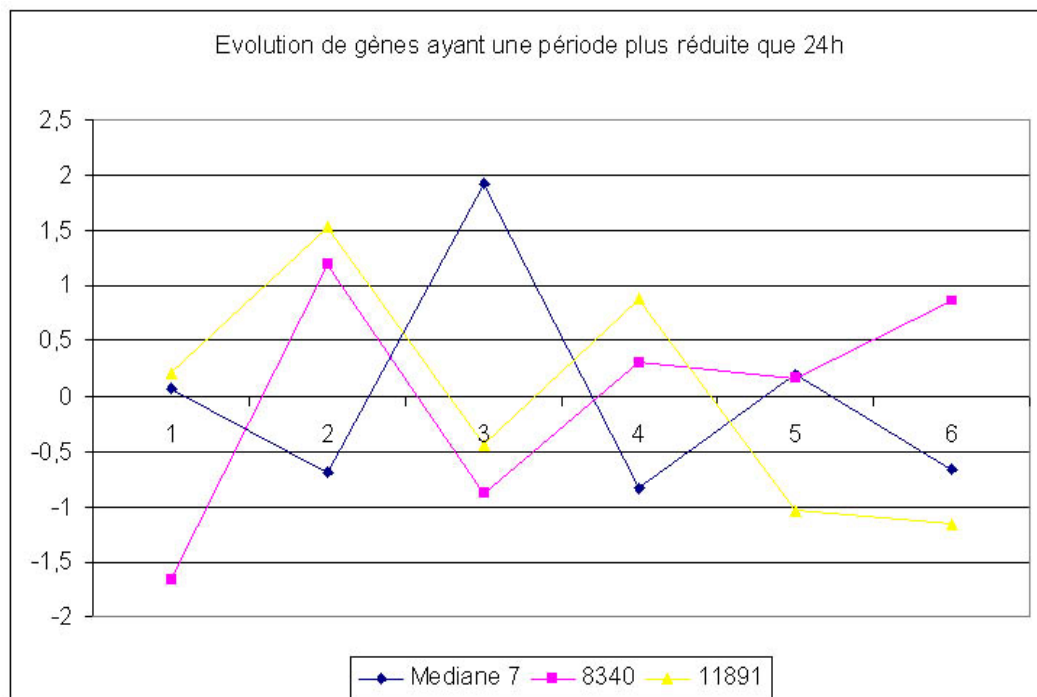
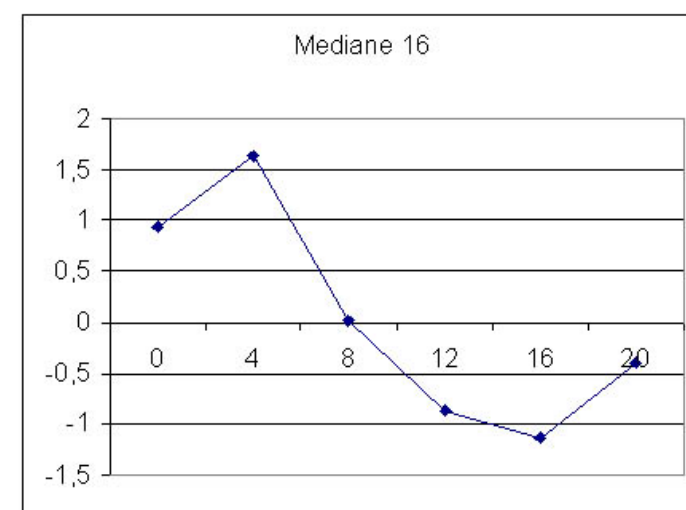
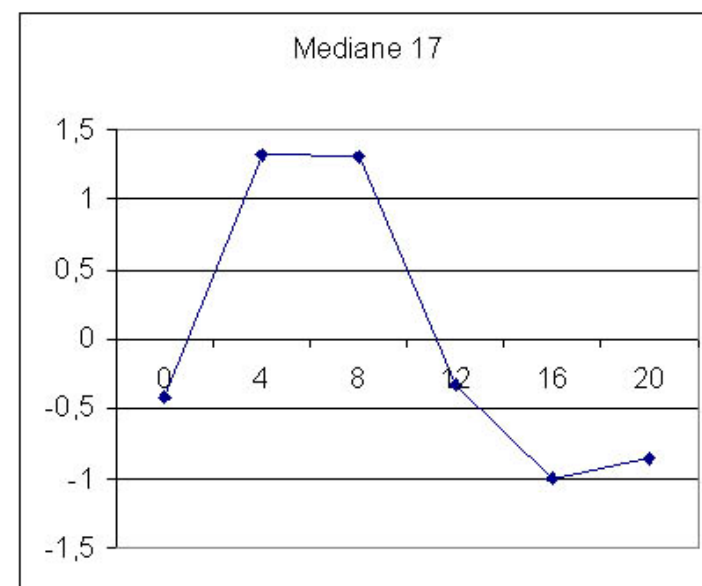
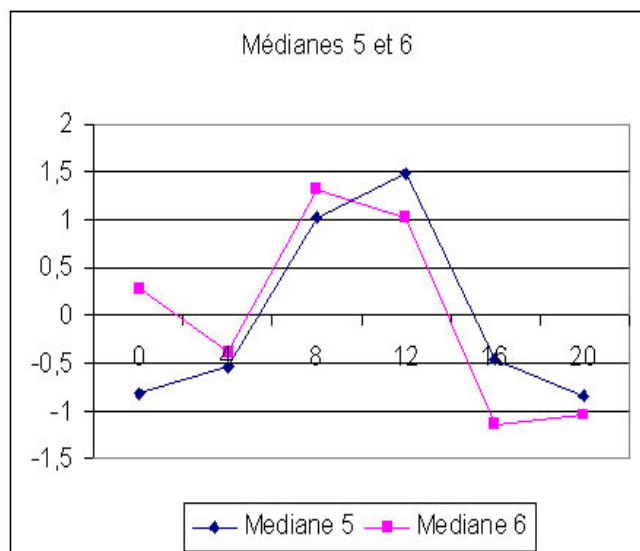
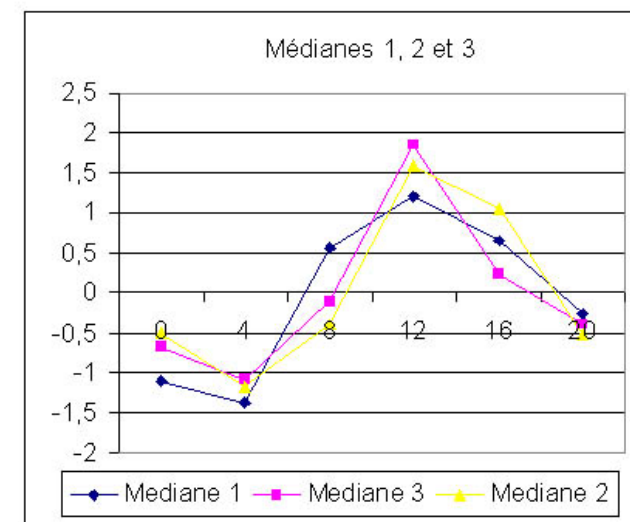
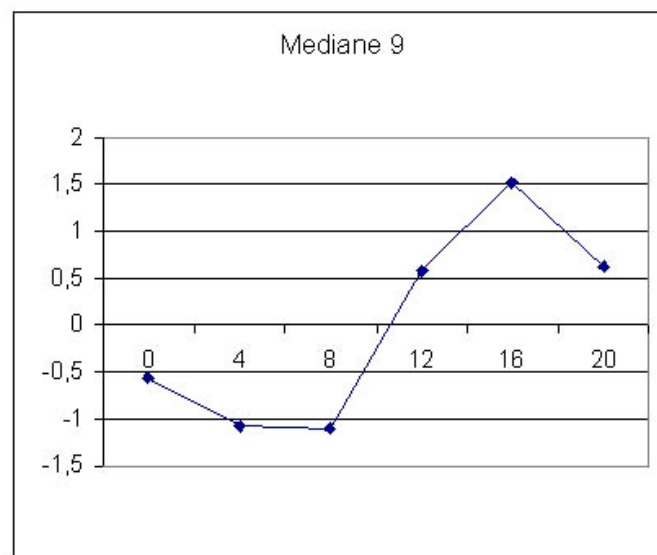
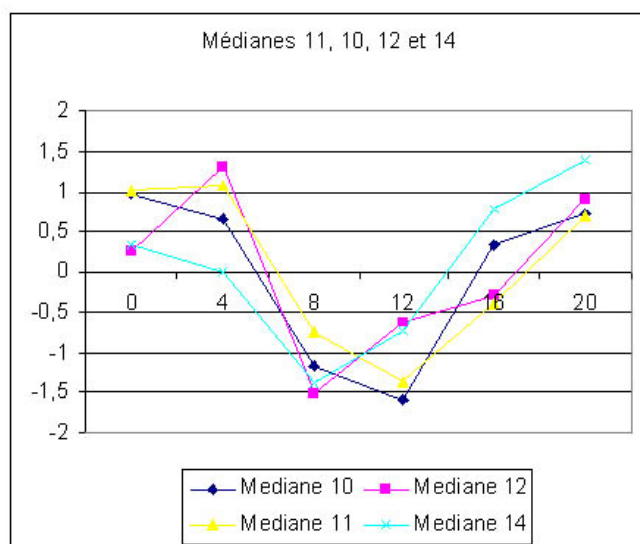


Figure 8 : Profils d'expression des différents archétypes (médiane) des gènes correctement représentés dans l'espace 1-2 de l'ACP



6 Synthèse : comparaison des différentes méthodes

La comparaison des méthodes se basera essentiellement sur l'expérience sur *Bacillus subtilis*, car l'organisation en opéron fournit un témoin interne de la cohérence des résultats. Le tableau 9 résume les résultats des différentes analyses de l'effet de la source de soufre. Les gènes sont classés en ordre de pertinence, le gène le plus « significatif » ayant le rang 1 (sauf pour l'ANOVA, la signification n'est pas réellement un degré de signification statistique).

L'ANOVA donne les résultats les plus cohérents dans la mesure où elle classe à peu près de la même façon la plupart des gènes d'un opéron. Elle présente aussi l'avantage de fournir une bonne approximation de la signification statistique des résultats.

L'ACI et la PLS font ressortir, elles-aussi, des opérons. Cependant, les résultats sont en général moins cohérents qu'avec l'ANOVA. (L'écart type des rangs des gènes d'un même opéron est deux fois plus grand avec la PLS et trois fois plus avec l'ACI). L'ACP n'a pas donné de résultats exploitables (l'écart type des rangs des gènes d'un même opéron est dix fois plus grand qu'avec l'ANOVA).

Les méthodes peuvent aussi être comparées entre elles en comptant le nombre de fois où un gène est classé parmi les vingt premiers par deux méthodes. L'ANOVA a sept gènes en commun avec l'ACI et dix avec la PLS. Ce ne sont d'ailleurs pas tout à fait les mêmes puisque l'ACI et la PLS n'en ont que quatre en commun. Avec ce critère aussi, l'ACP donne de mauvais résultats (le nombre de gènes en commun tombe à cinq).

Si l'on s'en tient aux résultats obtenus avec l'expérience sur *Bacillus subtilis*, l'ANOVA est clairement la meilleure méthode. Cependant, elle nécessite un plan d'expérience complet et le plus régulier possible. Dans les domaines où prédomine l'observation, ce qui est notamment le cas de nombreuses études médicales, il est impossible en pratique de concevoir un plan d'expérience analysable par ANOVA, ou alors, c'est au prix d'un appauvrissement tel que l'ANOVA dégénère en un modèle à un seul facteur. Elle perd alors le plus clair de son intérêt.

L'ACI et la PLS donnent des résultats moins bons, mais encore très utilisables. Les deux méthodes ont l'avantage d'être applicables dans tous les cas, quel que soit le plan d'expérience. L'ACI présente en plus l'intérêt d'être une technique exploratoire, susceptible de déceler des profils d'expressions inattendus. Par exemple, l'ACI a identifié l'ensemble des opérons impliqués dans la synthèse des protéines car une combinaison imprévue de facteurs a fait que, par hasard, la synthèse des protéines était plus faible que d'habitude le jour A, dans la culture avec du méthylthioribose. Le résultat était facile à interpréter car la fonction de ces opérons est connue. Mais on peut envisager aussi de procéder en aveugle et d'enregistrer les informations sur les covariations de l'expression de gènes. C'est une première étape qui devrait mener lentement, au fur et à mesure que les données s'accumuleront, vers l'identification des unités de régulation.

L'ACP ne fournit pas, à elle seule, une solution à l'analyse des expériences sur le transcriptome. En effet, l'ACP est focalisée sur la recherche de corrélations avec les axes d'inertie alors que rien, dans la nature du problème, n'impose qu'ils sont colinéaires aux axes discriminants. En revanche, le recours à l'ACP est pertinent pour l'analyse des séries

temporelles, après avoir utilisé une ANOVA pour sélectionner les gènes ayant une expression variable au cours du temps.

Tableau 9 : Bilan de la comparaison des différentes méthodes

Opérons trouvés avec plusieurs méthodes	Gènes	ACP	ACI	ANOVA	PLS
Similaire à des transporteurs d'acides aminés	yqiZ	62	15	2	8
	yqiY	9	62	9	1
	yqiX	42	21	1	2
Transformation du glutamate en citrulline	argC	310	66	216	386
	argJ	232	10	15	44
	argB	383	166	12	165
	argD	71	27	11	3
	carA	18	962	118	52
	carB	466	201	84	351
	argF	223	43	26	7
Transformation citrulline en arginine	ytzD	88	8	7	13
	argH	177	6	10	6
	argG	282	92	8	298
Lactate perméase et deshydrogénase	lctE	496	1	14	92
	lctP	540	16	65	133
Phosphotransférase (PTS) spécifique du fructose	levD	12	26	4	5
	levE	20	101	6	4
	levF	1 312	3 659	2 141	1 809
	levG	1 970	630	852	1 676
	sacC	296	254	320	221
Akyl hydroperoxide reductase	ahpC	30	45	97	116
	ahpF	17	19	5	64
Fonction inconnue	ycgE	199	1 282	279	110
	ycgF	897	3 748	2 088	2 653
	ycgG	15	978	113	9
Opérons spécifiques à l'ACI					
Fonction inconnue (opéron partiel)	sacV	1 895	119	745	952
	ydcO	2 153	3	789	1 633
	ydcP	1 675	134	1 042	1 942
	ydcQ	2 510	117	1 722	2 369
	ydcR	1 256	14	511	1 283
	ydcS	3 791	1 932	3 584	3 956
	ydcT	507	2	162	464
	yddA	162	57	91	257
	yddB	1 484	4	208	376
	yddC	871	353	886	1 531
	yddD	1 689	17	586	1 499
Transporteur lantibiotique	sunA	2 192	18	351	565
	sunT	355	12	30	51
Formation de ponts disulfure	yolI	176	5	21	37
	yolJ	1 706	3 494	2 726	1 772
	yolK	1 511	1 150	1 283	1 543
Fonction inconnue	yydF	3 582	11	1 885	1 731
	yydG	2 238	2 873	2 336	3 147
	yydH	3 274	2 774	3 003	3 301
	yydI	2 514	1 559	1 312	766
	yydJ	1 197	1 180	1 937	1 882
Protéines du flagelle	fliL	2 293	36	1 204	1 603
	fliM	2 176	41	1 014	1 518
	fliY	925	13	554	800
Opéron spécifique à l'ANOVA					
Diverses enzymes métaboliques	mtrA	2 893	2 160	1 672	2 970
	mtrB	1 364	1 365	1 599	2 057
Synthèse de la ménaquinone	gerCA	910	1 653	487	1 602
	gerCB	1 355	1 101	3	1 336
	gerCC	2 399	2 664	946	1 786

En gras figurent les gènes trouvés dans les vingt premiers par la méthode. Les chiffres correspondent au rang auquel on trouve le gène. Le terme "opéron spécifique à une méthode" signifie qu'il n'a été retrouvé dans les vingt premiers que par cette méthode. La notion d'"opéron partiel" correspond à un opéron de grande taille dont seule une portion a été représentée ici pour des problèmes de présentation.

Opéron spécifique PLS					
Régulation de la respiration	resC	316	697	300	273
(opéron partiel)	resD	2 755	3 187	2 472	2 067
	resE	53	2 080	144	10
Fonction inconnue	yyaE	1 520	2 156	1 103	583
	yyaF	2 472	1 701	194	2 780
Protéines ribosomales	rpsF	964	3 518	995	200
	ssb	2 320	2 059	1 914	634
	rpsR	142	476	104	18
Fonctions diverses (opéron partiel)	yloI	159	2 148	244	17
	priA	1 705	1 821	1 892	3 121
	def	3 584	1 105	1 715	2 569
Fonction inconnue	yisU	1 632	606	846	1 816
	yisT	51	1 134	68	12
Fonction inconnue	ywiE	137	661	83	14
	ywjA	2 668	2 863	1 974	1 896
	ywjB	2 729	234	780	1 079
Fonction inconnue	ytmA	24	1 340	49	20
	ytmB	3 816	3 346	3 114	3 453
Opérons spécifiques à l'ACP					
Protéines interagissant avec l'acétyl-coA	yusM	1 936	2 823	1 985	1 359
	yusL	1 900	2 918	3 531	3 848
	yusK	19	239	227	149
	yusJ	322	1 994	121	145
Perméase au sulfate	ylnA	1 523	458	938	2 391
Sulfate adényltransférase	ylnB	231	256	89	585
Adénysulfate kinase	ylnC	16	577	270	225
	ylnD	21	321	112	164
	ylnE	31	92	37	157
Fonction inconnue	yodL	2	245	96	32
	yodM	3 283	1 501	1 689	1 486
Fonction inconnue (opéron partiel)	dal	1 769	2 376	2 386	1 693
	ydcD	13	2 953	331	45
	ydcE	2 289	995	1 420	582
Fonction inconnue	ypbR	2 883	1 329	350	1 143
	ypbS	11	1 332	191	54
Fonction inconnue	yhdU	10	1 522	292	91
	yhdV	3 311	3 132	3 658	3 774
	yhdW	3 859	3 114	2 637	2 035
Fonction inconnue	ymcB	8	1 203	348	81
	ymcA	887	1 701	2 672	1 618
Exonucléases	sbcD	3 401	1 301	3 869	3 311
	yirY	1 363	2 700	361	910
	yisB	7	2 245	328	95
Opéron partiel	yxnB	36	2 492	470	291
Asparagine synthétase	asnH	6	1 293	313	86
	yxam	1 025	804	2 374	1 484
Fonction inconnue	ydiG	982	2 364	2 567	1 984
	ydiH	1 280	473	249	587
	ydiI	1 751	737	264	2 322
	ydiJ	5	731	178	81
Fonction inconnue	yflD	327	1 721	492	216
	yflC	622	1 889	174	182
	yflB	4	1 465	254	67
Gènes isolés					
Fonction inconnue	yycS	91	368	13	11
Subtilisine E	aprE	3	919	320	36
Protéine de flagelle	hag	1 580	7	1 300	2 013
Module cheA	cheV	768	20	411	960
Synthèse méthionine	metC	26	26	16	93
Fonction inconnue	yoqT	1	2 480	241	24
Fonction inconnue	yppF	14	1 508	384	57
Fonction inconnue	yvzB	1 980	9	1 221	2 139
Fonction inconnue	yvgO	44	674	44	15
Fonction inconnue	yugI	43	1 271	159	16

7 Améliorations envisageables

L'ANOVA ne pose pas de problèmes mathématiques. Il est en revanche nécessaire de réfléchir à l'organisation des expériences sur le transcriptome pour étendre les domaines où elle est applicable. Ceci se fera au hasard des contacts avec des expérimentateurs, chacun d'eux posant son problème, avec ses particularités.

L'ACI et la PLS sont moins connues et leur utilisation pour l'analyse du transcriptome débute seulement. Plusieurs études devraient permettre de mieux cerner leurs capacités et les conditions d'application.

L'ACI pose des problèmes mathématiques :

- L'algorithme de recherche des minima peut-il être amélioré ?
- D'autres fonctions non quadratiques sont-elles plus stable que $\Phi(x) = x^4$?

Ces questions seront abordées avec Bruno Torrèsani et son équipe.

Il faut aussi apprendre à gérer le nombre de sources indépendantes :

- Faut-il partir d'un maximum d'axes et ne garder ensuite que les faisceaux stables ?
- Existe-t-il des paliers dans le nombre de faisceaux stables ?
- Faut-il ajuster le protocole à la fonction $\Phi(x)$?

Avancer sur ces questions nécessite des données expérimentales sur le transcriptome bactérien car, dans ce cas, l'organisation en opéron du génome fournit un témoin interne. Heureusement, le laboratoire a des liens avec plusieurs équipes de biologie moléculaire qui peuvent fournir ce type de données.

Le mémoire ne présente qu'un premier contact avec la PLS. L'application de la PLS discriminante à l'analyse d'expériences sur le transcriptome est nouvelle par rapport à la littérature, mais les résultats sont jusqu'ici peu concluants. On peut attendre *a priori* beaucoup plus de la PLS. Ce nom regroupe en réalité une famille de méthodes qui ont en commun :

- de résoudre très simplement et très efficacement le problème des données manquantes, fréquent dans ce type d'expériences,
- de faire l'hypothèse que les observations résultent d'une combinaison linéaire d'un nombre réduit de variables latentes.

Ce dernier point, partagé avec l'ACI, donne une base naturelle à la modélisation des unités de régulation. Il n'a pas d'équivalent dans l'ANOVA et l'ACP. Il pourrait permettre de traiter des gènes ne variant pas de façon significative au regard de l'ANOVA, mais ayant des profils d'expression semblables. Les variations d'expression de la variable latente correspondant à un tel groupe de gènes devraient être significatives.

Mais, avant d'en arriver là, il faut déjà maîtriser les diverses variantes de la PLS.

8 Bibliographie

- Alaiya A. Franzen B. Hagman A. (2000) Classification of human ovarian tumors using multivariate data analysis of polypeptide expression patterns. *Int. J. Cancer*, 86:731-736
- Bry X. (1995) Analyses factorielles simples. Economica
- Bry X. (1996) Analyses factorielles multiples. Economica
- Chen Y., Dougherty E. R., Bittner M.L. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarrays images. *J. Biomed. Opt.*, 2: 364-367
- Chiappetta P., Roubaud M-C, Torr sani B. (2002) S paration de Sources pour l'analyse de donn es d'expression *JOBIM*.
- Comon P. (1994) Independent component analysis-a new concept? *Signal Processing*, 36: 287-314
- Didier G. (2002) GeneANOVA - gene expression analysis of variance. *Bioinformatics*, 18:490-491
- Draghici S. (2002) Statistical intelligence: effective analysis of high-density microarray data. *DDT*, 7(11): S55-S63
- Efron B., Tibshirani R., Goss V., Chu G. (2000) Microarrays and their use in a comparative experiment.
Manuscript (<http://www-stat.stanford.edu/~tibs/ftp/microarrays.pdf>)
- Fisher R.A. (1954) Statistical Methods for Research Workers. Oliver and Boyd. London
- Holter NS, Madhusmita Mitra, Amos Maritan, Marek Cieplak, Jayanth R Banavar, and Nina V Fedoroff (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. *PNAS*, 97: 8409-8414.
- Hyv rinen A. (1999) Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. *IEEE Transactions*, 10(3): 626-634
- Hyv rinen A., Oja Erkki (2000) ICA: algorithms and applications. *Neural networks*, 13:411-430
- Hyv rinen A., Oja Erkki (2000) Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 13:411-430
- Ideker T., Thorsson V., Siehel A.F., Hood L.E. (2000) Testing for differentially-expressed genes by maximum likelihood analysis of microarray data. *J. Comput. Biol.*, 7, 805-817
- Liebermeister W. (2001) Independent component analysis of gene expression data. Proc. German Conf. Bioinformatics. <http://www.bioinfo.de/isb/gcb01/poster/liebermeister.html>

Liebermeister W. (2002) Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18:51-60

Lin, Y., Nadler, S. T., Lan, H., Attie, A. D., and Yandell, B. S. (2002) Adaptive gene picking with microarray data: detecting important low abundance signals. In *The Analysis of Gene Expression Data: Methods and Software*, ed by G Parmigiani, ES Garrett, RA Irizarry, SL Zeger. Springer-Verlag (to appear), ch. 13.
<http://www.stat.wisc.edu/~yandell/statgen/yandell/pickgene.pdf>

Musumara G. (2001) Shortcuts in genome-scale cancer pharmacology research from multivariate analysis of the National Cancer Institute gene expression database. *Biochemical Pharmacology*, 62:547-553

Nadon R. (2002). Statistical issues with microarrays: processing and analysis. *Trends in Genetics*, 18(5) 265-271

Neidhardt F.C., Ingraham J.L., Schaechter M. *Physiologie de la cellule bactérienne: une approche moléculaire*. Masson Paris 1994

Newton M.A., Kendzierski C.M., Richmond C.S., Blattner F.R., Tsui K.W. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, 8:37-52

Nguyen D. V. Rocke DM (2002) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1): 39-50

Pan W., Lin J., Le C. (2001) A mixture model approach to detecting differentially expressed genes with microarray data. Technical Report. Division of Biostatistics. University of Minnesota (<http://www.biostat.umn.edu/cgi-bin/rrs?print+2001>)

Pan W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18(4) 546:554

Planet P.J. (2001) Systematic analysis of DNA Microarray Data: Ordering and Interpreting Patterns of Gene expression. *Genome Research*, 11:1149-1155

Quackenbush J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.*, 2(6): 418-427

Sekowska A., Robin S., Daudin J.-J., Hénaut A et Danchin A. : Extracting biological information from DNA arrays: an unexpected link between arginine and methionine metabolism in *Bacillus subtilis*. *Genome research*, 2001, 2(6)

Tusher V.G., Tibshirani R. Chu G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.*, 98: 5116-5121

Wold S. Trygg J., Berglund A. (2001) Some recent developments in PLS modeling. *Chemometrics and intelligent laboratory system*, 58:131-150

ANNEXE : implication de l'arginine dans les voies métabolique du soufre de *B. subtilis*

