Analyse du transcriptome de Buchnera aphidicola, la bactérie symbiotique du puceron Acyrthosiphon pisum

Hubert Charles

N° d'ordre HDR-2006-00003 Année 2006

INSA de Lyon - Université Claude Bernard Lyon 1

Analyse du transcriptome de *Buchnera aphidicola*, la bactérie symbiotique du puceron *Acyrthosiphon pisum*

Mémoire présenté pour l'obtention de l'Habilitation à Diriger des Recherches

Hubert CHARLES

Le 12 avril 2006

Membres du jury :
M. Stéphane AYMERICH (Rapporteur)
M. Christian BOUCHER (Rapporteur)
M. Stéphane ROBIN (Rapporteur)
M. Antoine DANCHIN
M. Jean-Michel FAYARD
M. Gérard FEBVAY
M. Christian GAUTIER
M. Hans GEISELMAN

Remerciements

Je remercie Messieurs Stéphane Aymerich, Christian Boucher et Stéphane Robin d'avoir accepté d'être rapporteurs de ce mémoire. Je remercie également Messieurs Antoine Danchin, Christian Gautier et Hans Geiselman, membres du Jury, pour le temps qu'ils ont consacré à mon travail.

Le Laboratoire de Biologie Appliquée (devenu maintenant BF2I) m'a accueilli grâce au Professeur Paul Nardon dès 1993 à l'occasion de mon DEA. Depuis, j'ai toujours travaillé avec un grand plaisir dans ses murs grâce à l'aide, la convivialité, la solidarité et la confiance de chacun de ses membres que je remercie très chaleureusement. Mes remerciements se portent tout particulièrement à Gabrielle Duport, Gérard Febvay et Yvan Rahbé mes « puceroneux » préférés, Jean Michel Fayard et Simon Grenier mes « guides spirituels » vénérés et Federica Calevro ma complice bien-aimée. J'associe à ces mercis un « very big thanks » à Angela Douglas pour sa rigueur et sa précision scientifique qui a su canaliser notre fougue latine.

Merci à Frédérique Laforest, Anne Tchounikine, Marie-France Sagot, Guillaume Beslon, Laurent Duret, Stéphane Génieys, Jean-Pierre Cloarec, Jean-François Boulicaut, Olivier Mazet, Olivier Gandrillon, Jacques Bernillon et Jean-Jacques Madjar pour ces prises de têtes (de becs) toujours enrichissantes dans la salle BSMC ou au DTAMB.

Je remercie tout particulièrement les étudiants « travailleurs de la mer » qui n'ont jamais hésité à grimper au sommet du grand mât pour haler les drisses en pleine tempête : Sylvain Blachon, Claire Allen, Laetitia Duby, Caroline Truntzer, Aurélie Laugraud, Céline Keime, Carole Knibbe, Nicolas Morin, Nancie Reymond, José Viñuélas, Ludovic Cottret et Anthony Cros.

Enfin, un grand merci à Sandrine, Sarah et Jéremy pour leur patience et leur amour et sans qui ce travail n'aurait aucun sens.

Résumé

Les progrès fulgurants de ces dix dernières années réalisés dans les domaines de la microinformatique et de la microfluidique associés au génie génétique (PCR et séquençage) ont permis un changement d'échelle dans la quantité des données acquises au cours d'une même expérience. La transcriptomique est directement issue de ces avancées technologiques. Ce mémoire présenté pour l'obtention d'une Habilitation à Diriger des Recherdes porte sur l'analyse du transcriptome de la bactérie intracellulaire obligatoire des pucerons, *Buchnera aphidicola*. Dans la première partie, les principales méthodes d'analyses statistiques différentielles et d'intégration des données transcriptomiques sont présentées sous la forme d'une analyse bibliographique. La deuxième partie est consacrée au développement d'outils bioinformatiques : ROSO, un logiciel d'optimisation des sondes oligonucléotidiques, la puce *Buchnera* et SITRANS, un système d'information pour la gestion et la publication des données d'expression. Enfin, la dernière partie est consacrée à la caractérisation du transcriptome de *Buchnera* en condition de stress trophique de son hôte, le puceron du pois *Acyrthosiphon pisum*. La régulation transcriptionnelle chez les bactéries symbiotiques intracellulaires à génome réduit est encore actuellement très mal connue. Cette question sera abordée chez *Buchnera* tout d'abord au niveau évolutif par l'étude de la relation entre l'expression des gènes et leur organisation dans le génome, puis au niveau fonctionnel, par la caractérisation de la réponse de la bactérie à une diminution de la quantité d'acides aminés essentiels dans le substrat nutritif du puceron, combinée à un stress osmotique.

Summary

These last ten years, the dazzling advances in computing and microfluidics applied to molecular biology (PCR and sequencing) caused a complet rescaling of the data sets coming from a single experiment. This essay presented for the « Habilitation à Diriger des Recherches » is devoted to the transcriptome of the intracellular symbiotic bacteria of aphids, *Buchnera aphidicola*. In a first part, the main methods of differential statistics and data integration will be presented as a bibliographic report. The second part details bioinformatic tools for microarrays: ROSO, a software for oligonucleotide probe optimisation, the *Buchnera* microarray and SITRANS, an information system for the management of microarray data. The last part characterizes the transcriptome of *Buchnera* in trophic stressed conditions of their aphid host. Transcription regulation is yet little-known in the reduced genomes of intracellular bacteria. This question will be first adressed in *Buchnera* at the level of the molecular evolution with the analysis of the relationship between gene expression and genome organisation. Secondly, the functional response of the bacteria will be analysed in an experimental design combining the effect of essential amino acid depletion and osmotic stress in the diet of the aphid host.

Table des matières

Préambule

Je suis arrivé au Laboratoire de « Biologie Appliquée » de l'INSA de Lyon en 1994. Le professeur Paul Nardon dirigeait alors l'UMR INRA-INSA 203, il était aussi mon directeur de DEA. Le sujet qui me fut attribué était la caractérisation, par électrophorèse bidimensionnelle, des protéines d'interactions de la symbiose chez le charançon des céréales *Sitophilus oryzae* (Coléoptère, Curculionidae). En quelque sorte, il s'agissait déjà d'une approche protéomique des relations hôte - symbiote chez le charançon. Nous étions cependant assez loin de l'analyse à haut débit ! Paul Nardon avait imaginé ce sujet à la suite des travaux du Pr. Ishikawa (Université de Tokyo) travaillant sur le puceron du pois *Acyrthosiphon pisum* (Homoptère, Aphididae). L'équipe japonaise avait montré que les endocytobiotes (bactéries symbiotiques intracellulaires du genre *Buchnera*) synthétisaient, de façon presque exclusive, une protéine que les chercheurs avaient appelée « symbionine » (Ishikawa 1982a et b). Mon travail de DEA permit de mettre en évidence une symbionine chez le charançon (Charles *et al.* 1995).

Ma thèse a alors consisté à caractériser cette protéine, par ailleurs bien connue chez *Escherichia coli,* puisqu'il s'agissait de la protéine chaperon moléculaire GroEL ou Hsp60. L'opéron *groES* a été cloné et séquencé chez les endocytobiotes de *Sitophilus oryzae*. Nous avons montré ensuite que l'expression du gène codant la protéine était sélectivement induite lorsque les endocytobiotes étaient dans leurs bactériocytes. Enfin, nous avons réalisé des expériences sur la régulation de cet opéron en cas de choc thermique *in vivo* et *ex vivo*, dans des bactériocytes isolés (Charles *et al.* 1997b).

Dans une deuxième partie de ce travail de thèse et grâce au séquençage du gène *rrs* (ARN ribosomal 16S) des endocytobiotes de *S. oryzae* (SOPE) par A. Heddi au laboratoire, nous avons montré que les bactéries symbiotiques des *Sitophilus*

céréaliers appartenaient à la famille des Enterobacteriaceae (très proches *d'Erwinia herbicola*). Par électrophorèse en champs pulsés, la taille de SOPE fut estimée à 3,0 Mb (Charles *et al.* 1997a). Enfin, le taux de GC de l'ADN des endocytobiotes a été estimé à 54 ± 2 % par chromatographie liquide à haute performance (Charles 1997 ; Heddi *et al.* 1998).

Cette deuxième partie de mon travail de thèse fut l'occasion de me pencher sur des problèmes d'évolution et de phylogénie moléculaire chez les endocytobiotes, et plus généralement chez les bactéries intracellulaires pathogènes et mutualistes. Ainsi, nous avons pu mettre en évidence que toutes les bactéries intracellulaires obligatoires possèdent des génomes riches en bases A et T, et des tailles de génomes très faibles. Une corrélation très significative entre la taille du génome et le taux de GC a été mise en évidence pour les bactéries intra- et extracellulaires. Avec le séquençage complet de très nombreuses bactéries pathogènes ou symbiotiques intracellulaires, ce résultat s'impose maintenant comme une évidence. Mais en 1997, les données récoltées manuellement dans les quatre volumes du Bergey's Manual of Bacteriology avaient été tout juste suffisantes pour convaincre les relecteurs de notre publication (Heddi *et al.* 1998).

Les résultats physiologiques, obtenus jusqu'alors sur la symbiose chez les *Sitophilus,* laissaient supposer que l'association entre la bactérie et le charançon était très ancienne. En effet, la transmission verticale des endocytobiotes, leur intégration et la forte complémentarité des métabolismes de l'hôte et de ses symbiotes étaient des critères révélant l'existence d'une longue coévolution entre les deux partenaires (Nardon et Grenier 1993). Les résultats apportés par mon travail de thèse semblaient indiquer tout à fait le contraire. La grande taille du génome et le fort taux de GC de ces bactéries, ainsi que leur positionnement phylogénétique semblaient les faire correspondrent à des bactéries intracellulaires « modernes ». C'est ce point clé de la compréhension du système symbiotique que je me suis attaché à résoudre dans mon travail PostDoctoral et dans mes premières années de recherche comme Maître de Conférences à l'INSA dans le laboratoire de Biologie Fonctionnelle, Insectes et Interactions (BF2I).

Mon premier stage PostDoctoral s'est déroulé au Japon chez le Pr. Ishikawa. En effet, celui-ci avait estimé la taille du génome de *Buchnera* à environ cinq fois la taille du génome d'*E. coli* (soit environ 23 Mb) par la technique de mesure de densité de l'ADN (Ishikawa 1987). Cette technique, bien que très utilisée en microbiologie, est réputée pour son imprécision. Néanmoins, il était intriguant, au vu du faible taux de GC de la bactérie et de l'âge très ancien de son association (connu grâce à des fossiles du pucerons conservés dans l'ambre), de constater que son génome soit de taille si imposante. J'ai donc proposé au Pr. Ishikawa d'analyser la taille du génome de la bactérie par électrophorèse en champs pulsés. C'est ainsi que nons l'avons estimée à 640 kb, soit environ le septième du génome d'*E. coli*

(Charles et Ishikawa 1999). Ce résultat corroborait du même coup nos hypothèses de corrélation entre taille du génome, âge de la symbiose et taux de GC des génomes. J'ai également dressé la carte physique du génome de la bactérie, ce qui a permis au laboratoire japonais de se lancer directement dans le séquençage complet de la bactérie en collaboration avec le centre de séquençage du « Riken Institute » (Shigenobu *et al*. 2000). Le Pr. Ishikawa est décédé le 22 novembre 2005, la communauté des symbiologistes a perdu ce jour là un des ses plus grands chercheurs.

J'ai effectué mon deuxième stage PostDoctoral dans le laboratoire de Biométrie et Biologie Evolutive (UMR CNRS 5558) de l'Université Claude Bernard Lyon I, dans l'équipe de M. Gouy pour étudier le positionnement phylogénétique des endocytobiotes d'insectes au sein des γ-Protéobactéries. Ce travail a nécessité la mise en œuvre de méthodes d'analyses phylogénétiques basées sur un modèle d'évolution non homogène pour prendre en compte l'hétérogénéité des taux de GC et des vitesses d'évolution. Nous avons ainsi pu établir que la majorité des endocytobiotes d'insectes provenait d'un groupe de bactéries ancêtres des Entérobactéries actuelles. Le groupe des endocytobiotes n'étant pas monophylétique, nous avons émis l'hypothèse d'un « clade souche » ancestral pour une grande majorité des endocytobiotes d'insectes (mais pas SOPE). Ces bactéries ancestrales devaient posséder des propriétés pathogéniques et un métabolisme adéquat leur permettant d'infecter des insectes phytophages et phloémophages et d'établir ensuite des relations mutualistes intégrées (Charles *et al.* 2001).

Après ma nomination au BF2I, mes activités de recherche se sont focalisées plus directement sur l'estimation de l'âge de la symbiose chez les *Sitophilus*. Cette question nécessitait plus largement la caractérisation de la symbiose au niveau de la famille des Dryophthoridae, famille à laquelle appartiennent les *Sitophilus*. Ce projet s'est essentiellement déroulé sur la période 1999-2002 avec l'encadrement de plusieurs étudiants de DEA et d'un doctorant ainsi qu'une collaboration active avec le Pr. Brian Farrell, spécialiste des Coléoptères au Muséum de l'Université d'Harvard (Boston, USA). Au cours de ce travail, nous avons collecté de nombreuses espèces de charançons Dryophthoridae et leur statut symbiotique a été déterminé grâce à des analyses histologiques et moléculaires (Nardon *et al.* 2000 ; 2001 ; 2003). Une étude phylogénétique, basée sur l'ADN 16S des bactéries symbiotiques et utilisant un modèle d'évolution non homogène, a ensuite été réalisée de façon à reconstruire l'histoire évolutive de la symbiose dans ce groupe d'insecte. Nous avons montré que les endocytobiotes des Dryophthoridae se divisent en trois clades appartenant aux Protéobactéries. Le clade ancestral a été nommé *Candidatus* Nardonella (en l'honneur du Pr. Paul Nardon) ; il semble avoir infecté les Dryophthoridae il y a environ 100 M d'années. Dans le groupe des *Sitophilus*, les SOPE sont vraisemblablement issus d'un remplacement compétitif, par des bactéries pathogènes, il y a 10 à 20 M d'années (Lefèvre *et al.* 2004). Ce travail a été présenté

dans la thèse de C. Lefevre (2004) qui retrace l'histoire évolutive de la symbiose chez les Dryophthoridae. On peut noter que l'ordre des Coléoptères est le plus vaste groupe d'animaux de la terre puisqu'il représente près du tiers de toutes les espèces animales recensées. Depuis longtemps déjà les chercheurs ont essayé de comprendre comment et pourquoi ces insectes s'étaient si fortement diversifiés. Ce travail avec C. Lefevre a montré que la symbiose joue un rôle prépondérant dans les mécanismes d'adaptation et de spéciation des insectes. Les symbiotes, véritables « kit de gènes d'adaptation » transmissibles à la descendance, permettent à des lignées d'insectes de coloniser des niches écologiques moins favorables où la concurrence interspécifique est plus faible, favorisant ainsi la spéciation. Il a de plus permis de lever une partie de l'ambiguïté sur la symbiose chez les *Sitophilus*. La symbiose dans ce groupe est très ancienne, probablement aussi ancienne que chez les pucerons. Il est donc normal d'observer chez le charançon des structures d'accueil très évoluées, comme les bactériocytes, ainsi que l'intrication complète et l'interdépendance des métabolismes bactériens et eucaryotes. Néanmoins, dans le groupe des *Sitophilus*, il est vraisemblable que les symbiotes ancestraux ont été remplacés récemment par de nouvelles bactéries (SOPE) ayant colonisées les bactériocytes. Les SOPE présentent donc des caractéristiques de bactéries « modernes » dans une structure d'accueil ancestrale. Ce résultat nous permettait enfin de réconcilier les résultats physiologiques et moléculaires, apparemment contradictoires, obtenus jusqu'à présent au laboratoire.

En 2000, la séquence du génome de *Buchnera aphidicola* était publiée par le groupe du Pr Ishikawa (Shigenobu *et al.* 2000), l'action des génopoles françaises finançait la construction d'une plateforme transcriptomique sur le campus de la Doua, et le Département de Biosciences de l'INSA créait une nouvelle filière de formation d'Ingénieurs en Bioinformatique et Modélisation sous l'impulsion des Pr. Jean-Michel Fayard et Christian Gautier. Ce contexte nous a poussé à développer au laboratoire une activité de recherche en Bioinformatique. La bactérie *Buchnera*, symbiote du puceron du pois *Acyrthosiphon pisum*, paraissait le modèle d'étude le plus adapté pour débuter une analyse transcriptomique, d'autant que ce programme pouvait venir compléter de façon pertinente les travaux déjà réalisés au laboratoire. Actuellement, les projets de séquençage du génome d'*A. pisum* et d'une de ses plantes hôtes *Medicago truncatula*, rendent le programme encore plus pertinent.

Ce mémoire pour l'obtention d'une habilitation à diriger des recherdes ne portera que sur l'analyse du transcriptome de *Buchnera*, projet qui a occupé mes cinq dernières années de recherche. Le mémoire est structuré en trois parties. La première concerne la présentation de mon cursus, des encadrements, des collaborations et des publications que j'ai réalisés depuis ma thèse. Les deux suivantes relatent mes activités de recherche et d'enseignement.

La deuxième partie concernant mes activités de recherche est la plus importante. Elle comporte quatre chapitres rédigés pour être lus indépendamment. Le premier est une introduction succincte sur le modèle d'étude : le puceron *Acyrthosiphon pisum* et sa bactérie symbiotique intracellulaire *Buchnera.* Il se termine par une présentation du contexte scientifique et du programme de recherche. Le deuxième chapitre porte sur les développements technologiques que nous avons réalisés : ROSO, un logiciel d'optimisation des sondes oligonucléotidiques, la conception d'une puce à ADN dédiée à *Buchnera*, et SITRANS, un système d'information pour la gestion et la publication des données d'expression. Ces trois réalisations sont publiées et ne sont que très brièvement décrites. Le troisième chapitre est consacré à l'analyse statistique des données d'expression. Il aborde des aspects de stratégie expérimentale, une revue des méthodes pour l'analyse qualité, la normalisation, l'analyse différentielle, et enfin, une analyse de quelques méthodes d'intégration de résultats dans le contexte biologique. Cette partie qui correspond à un travail plus personnel sera entièrement développée, elle n'est de toute évidence pas exhaustive, mais plutôt focalisée sur notre modèle et les questions que nous lui adressons. Enfin, le quatrième chapitre, le plus biologique, correspond à l'analyse des données issues des expériences transcriptomiques réalisées au laboratoire sur *Buchnera,* en condition de stress trophique de la part de son hôte. Cette partie est le cœur de notre programme de recherche et correspond au travail de toute l'équipe. Elle sera entièrement détaillée dans ce rapport.

PREMIERE PARTIE

Présentation du candidat

État civil

Hubert CHARLES

Date et lieu de naissance : 3 novembre 1969 à Paris XI^{ème}

Situation familiale : marié, 2 enfants

Adresse personnelle :

38, rue Guillaume Apollinaire
69330 Meyzieu
Tel: 04-78-31-40-66

Adresse professionnelle :

Laboratoire de Biologie Fonctionnelle, Insectes et Interactions (BF2I), UMR INRA/INSA de Lyon 203
INSA Bât Louis Pasteur,
69 621 Villeurbanne cedex
Tel : 04-72-43-80-85
Fax : 04-72-43-85-34
Email : hubert.charles@insa-lyon.fr

Grades et Titres

Depuis le **1er octobre 1998 :** Maître de Conférences. Laboratoire de Biologie Fonctionnelle Insectes et Interactions (Directeur, G. Febvay), Titularisation le 01/09/99. Prime d'Encadrement Doctorale et de Recherche (PEDR) depuis 2001

1998 : Stage Post-Doctoral d'avril à septembre. UMR CNRS 5558, Université Claude Bernard Lyon 1 (équipe de M. Gouy). *Sujet* : Méthodologie de reconstruction phylogénétique dans les symbioses intracellulaires (à fort taux de A+T et forte vitesse évolutive).

1997 : Stage Post-Doctoral d'octobre 97 à mars 98. Zoological Institute, Université de Tokyo, (équipe du Pr. H. Ishikawa). *Sujet* : Analyse du génome de *Buchnera aphidicola*, l'endocytobiote du puceron du pois, *Acyrthosiphon pisum*, par électrophorèse en champs pulsés.

1997 : Thèse de Doctorat (Laboratoire de Biologie Appliquée INSA 406). Félicitations du Jury – Prix Malotau de Guerne 1997. Moniteur à Université Claude Bernard Lyon 1 (Biomathématiques).

1994 : DEA : «Analyse et Modélisation des Systèmes Biologiques». Mention Très Bien.

1993 : Professeur du Contingent en Biologie au Lycée militaire d'Autun.

1992 : INSA de Lyon, département de Biochimie. Diplôme d'ingénieur biochimiste (Félicitations du Jury).

1989 : Premier cycle INSA en section Sport-Etudes (Handball).

1987 : Lycée Jean Perrin (69) en section Sport-Etudes (Handball). Baccalauréat C (Mention Bien).

Encadrement

1. Stages de fin d'études

Michel MOIROUD : Etude de la symbiose intracellulaire chez les animaux. Stage Ingénieur du Conservatoire National des Arts et Métiers de Lyon – *Spécialité* : Biologie en vue des applications, (2 mois, 1997).

Roselyne ALAUYAU : Analyse moléculaire des bactéries symbiotiques intracellulaires chez les charaçons du genre *Sitophilus*. D.E.S.T. option Biochimie et Biotechnologies des Industries Agroalimentaires. Conservatoire National des Arts et Métiers de Lyon, (6 mois 1999-2000).

Sylvain BLACHON : Développement d'un logiciel d'aide à la conception de puces à ADN. Stage de fin d'étude Ingénieur, INSA de Lyon, (6 mois, 2001).

Laetitia DUBY : Développement d'une base de données pour la gestion des données de transcriptome. Stage de fin d'étude IUT – IUT Grenoble II, Statistiques et analyse informatique des données, (3 mois, 2002).

Claire ALLEN : PhD université de York, Analyse du transcriptome de *Buchnera* en condition de stress trophique de son hôte, Directeur de Thèse, Pr. A. Douglas, (2 mois, 2002).

2. DEA, Masters

Sophie CHAUDIER (DEA Analyse et Modélisation des Systèmes Biologiques, 1998-1999). *Sujet* : Caractérisation moléculaire de la symbiose chez les charançons du genre *Sitophilus* : recherche d'une plurisymbiose intracellulaire. Codirection (50%).

Cédric LEFEVRE : (DEA Analyse et Modélisation des Systèmes Biologiques, 1999-2000). *Sujet* : Caractérisation moléculaire et phylogénie des bactéries symbiotiques intracellulaires des charançons de la famille des Dryophthoridae. Codirection (50%).

Alexandre MEYRAND : (DEA Analyse et Modélisation des Systèmes Biologiques, 2000-2001). *Sujet* : Etude du polymorphisme des copies de l'ARN 16S chez les endocytobiotes des charançons du genre *Sitophilus*. Codirection (50%).

Nancie REYMOND : (DEA Analyse et Modélisation des Systèmes Biologiques, 2000-2001). *Sujet* : Bioinformatique des puces à ADN et application à l'analyse du transcriptome de *Buchnera aphidicola*. Codirection (50%).

Nicolas MORIN : (DEA Analyse et Modélisation des Systèmes Biologiques, 2003-2004). *Sujet* : Analyse de l'expression des gènes chez *Buchnera*, bactérie symbiotique des pucerons, par l'utilisation d'une puce à ADN dédiée.

José VIÑUELAS : (Master EEME, Ecologie Microbienne, 2004-2005). *Sujet* : Organisation et évolution du génome de *Buchnera aphidicola*. Coencadrement (50%).

Anthony CROS : (Master aMIV, Méthodes en Bioinformatique Moléculaire, 2005-2006). *Sujet* : Développement d'un module de SITRANS pour le transfert de données vers les bases internationales de données d'expression. Codirection (50%).

3. Thèses

Cédric LEFEVRE : (2000-2004). *Sujet* : Caractérisation et phylogénie des bactéries symbiotiques intracellulaires des charançons de la famille des Dryophthoridae. Codirection : A. Heddi (50 %), H. Charles (50 %). Thèse soutenue le 10/02/2004.

Nancie REYMOND (2001-2004). *Sujet* : Bioinformatique des puces à ADN et application à l'analyse du transcriptome de *Buchnera aphidicola*. Codirection : JM. Fayard (50 %), H. Charles (50 %). Thèse soutenue le 16/12/2004.

Ludovic COTTRET (2005-) *Sujet* : Caractérisation et évolution du réseau métabolique de *Buchnera aphidicola*. Codirection : MF. Sagot (50 %, LBBE, UCBL), H. Charles (50 %). Thèse en cours.

José VIÑUELAS (2005-) *Sujet* : Caractérisation du transcriptome de *Buchnera aphidicola* en condition de stress trophique. Coencadrement : F. Calevro (50 %), H. Charles (50 %). Thèse en cours.

4. Comités de pilotage de thèses

Bertrand BONNAUD : Distribution des éléments transposables dans le génome humain, thèse 2001-2004 (ENS Lyon).

Olivier SEMETEY : Biologie et transmission du BLO responsable du syndrome des basses richesses de la betterave sucrière, thèse 2002-2005, UMR INRA 1088 (Dijon).

Caroline TRUNTZER : Analyse de la valeur pronostique et diagnostique des données de puces à ADN, thèse 2003-2006 (UCB Lyon 1).

Carole KNIBBE : Evolution structurelle des génomes bactériens : un modèle de vie artificielle, thèse 2002-2005 (INSA Lyon 1).

Anamaria NECSULEA : Analyse et modélisation des structures génomiques en relation avec les processus évolutifs. Application au cas des génomes de la famille des Trypanosomatidae, Thèse 2005-2008 (UCB Lyon 1).

4. Collaborations scientifiques

Jacques BERNILLON (IE) et **Jean-Jacques MADJARD** (Pr.). DTAMB, IFR 41 (UCBL). *Collaboration* : Développement de la puce à ADN *Buchnera* sur la plateforme transcriptome de la génopole Rhône-Alpes.

Guillaume BESLON (MdC) Laboratoire PRISMa (INSA). *Collaboration* : Modélisation de l'évolution de la structure des génomes bactériens (groupe BSMC).

Jean-François BOULICAUT (MdC). Laboratoire LIRIS (INSA). *Collaboration* : Analyse et intégration des données d'expression (groupe BSMC).

Angela DOUGLAS (Pr), Department of Biology, University of York, PO Box No 373, York. *Collaboration* : Interactions trophiques (métabolisme du sucre) entre *Buchnera* et son hôte.

Brian FARELL (Pr), Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138 – *Collaboration* : Evolution et systématique des Dryophthoridae.

Nicolas GALTIER (CR), Génétique et Génomique Evolutive, Université Montpellier II - *Collaboration* : Phylogénie moléculaire, modèles d'évolution non homogènes.

Olivier GANDRILLON (DR). Laboratoire CGMC, Université Claude Bernard Lyon1. *Collaboration* : Analyse systémique (groupe BSMC).

Stéphane GENIEYS (MdC). Laboratoire C. Jordan, (INSA-UCBL). *Collaboration* : inférence de réseaux de régulation à partir de données d'expression (groupe BSMC).

Frédérique LAFOREST (MdC) et **Anne TCHOUNIKINE** (MdC). Laboratoire LIRIS (INSA). *Collaboration* : Développement de la base de donnée SITRANS pour la gestion des données d'expression (groupe BSMC).

Olivier MAZET (MdC), Laboratoire C. Jordan (INSA-UCBL). *Collaboration* : Evolution des réseaux de régulation transcriptionnelle (groupe BSMC, http://bsmc.insa-lyon.fr).

Anne MORGAT (CR), **François RECHENMANN** (DR) et **Alain VIARI** (DR), INRIA Rhône-Alpes - Projet HELIX, 655, Av. de l'Europe, 38330 Montbonnot Saint Martin. *Collaboration* : Développement d'une base de connaissance sur *Buchnera* incluant les données d'expression (plateforme Genostar).

Marie-France SAGOT (DR), Laboratoire Biométrie et Biologie Evolutive (UMR CNRS 5558), Equipe BAOBAB, 69621 Villeurbanne. *Collaboration* : Caractérisation et évolution du réseau métabolique de *Buchnera*.

Denis TAGU (CR), INRA Rennes, UMR INRA-Agrocampus BiO3P, BP 35327, 35653 Le Rheu Cedex. *Collaboration* : séquençage du génome du puceron, production de puces pucerons.

Tom WILKINSON (Lecturer), Department of Zoology, University College Dublin, Belfield, Ireland. *Collaboration* : Etude de l'influence de *Buchnera* sur le développement du puceron.

5. Contrats de recherche

J'ai participé à la rédaction et aux travaux de différents contrats de recherches :

Bonus Qualité Recherche inter-établissements (INSA, UCBL, ECL) (2001). « Analyse du transcriptome de *Buchnera aphidicola* et développement de puces à ADN ». 50 keuros.

Programme Bioinformatique inter EPST – Action 1 (2001–2002). « Optimisation des séquences sondes déposées sur les puces à ADN et normalisation des résultats de fluorescence ». 15 keuros.

Programme Bioinformatique inter EPST – Axe 2 (2002-2003). « Développement d'un environnement logiciel pour la gestion, l'étude et le partage d'expérimentation sur les puces à ADN ». - 15 keuros.

Action Bioinformatique des Génopoles (2002-2003). « Recrutement d'un ingénieur CDD pour le développement de SITRANS ». 1 an de CDD.

Région Rhône-Alpes Programme Emergence (2003-2004). « Biologie des Systèmes et Modélisation Cellulaire » 35 keuros.

Projet intradépartement SPE INRA (2004-2006) « Analyse quantitative des réseaux de régulation chez la bactérie *Buchnera aphidicola* ». 15 keuros.

6. Animations scientifiques

Je suis membre de conseils de différentes instances :

- Le **conseil de laboratoire** de l'UMR INRA / INSA de Lyon 203 (BF2I) ;
- La **commission de spécialistes** sections 64 à 69 de l'INSA de Lyon ;
- Le **conseil scientifique** de la plateforme transcriptome Génopole du DTAMB (UCBL) où j'anime un groupe de travail sur l'analyse statistique des données d'expression depuis septembre 2002 avec l'organisation de séminaires et de deux ateliers de formation CNRS- INSERM - INRA en 2003 et 2004.
- Le **conseil scientifique** du PRABI (Plateforme Rhône Alpine de Bioinformatique) ;
- Le **comité de pilotage** du groupe BSMC (http://bsmc.insa-lyon.fr/). Ce groupe de travail inter-laboratoires a été créé en 2000 et regroupe une vingtaine de chercheurs issus de 5 laboratoires de l'INSA et de l'UCBL. Nous organisons des séminaires scientifiques et développons une activité de recherche interdisciplinaire ;
- Le **comité d'organisation** du congrès de bioinformatique JOBIM 2005.

7. Publications – Colloques - Posters

7.1. Articles de revues nationales et internationales

Seules les publications avec comité de lecture sont numérotées sur la liste ci-dessous. Le classement est chronologique.

1. Charles H., Ishikawa H., Nardon P., 1995. Presence of a protein specific of endocytobiosis (symbionin) in the weevil *Sitophilus*. *C.R. Acad. Sci. Paris*, **318**: 35-41.
2. Charles H., Condemine G., Nardon C., Nardon P., 1997. Genome size characterization of the principal endocellular symbiotic bacteria of the weevil *Sitophilus oryzae*, using pulsed field gel electrophoresis. *Insect Biochem Molec Biol*, **27**: 345-350.
3. Charles H., Heddi A., Guillaud J., Nardon C., Nardon P., 1997. A molecular aspect of symbiotic interactions between the weevil *Sitophilus oryzae* and its endosymbiotic bacteria: over-expression of a chaperonin. *B B R C*, **239**: 769-774.

Charles H., Nardon P., 1997. La protéine Hsp60 : chaperon moléculaire de la symbiose intracellulaire. *Regard sur la Biochimie*, **3**: 17-23.

Nardon P., Charles H., Heddi A., 1997. Un modèle de symbiose : le charançon *Sitophilus oryzae*, implications evolutives. In : L'évolution Philosophique. Science, Histoire ou Philosophie, Librairie Philosophique, J. Vrin, Exbrayat J.M., Flatin J. (Eds), pp. 297-305.

Charles H., 1997. Aspects moléculaires de la bactérie symbiotique principale du charançon des céréales *Sitophilus oryzae* (Coléoptère, Curculionidae) et étude de ses interactions avec l'hôte. *Thèse de doctorat*, INSA de Lyon, 132 p.

4. Heddi A., Charles H., Khatchadourian C., Bonnot G., Nardon P., 1998. Molecular characterization of the principal symbiotic bacteria of the weevil *Sitophilus oryzae*: a peculiar G + C content of an endocytobiotic DNA. *J Mol Evol*, **47**: 52-61.

5. Charles H., Ishikawa H., 1999. Physical and genetic map of the genome of *Buchnera*, the primary endosymbiont of the pea aphid *Acyrthosiphon pisum*. *J Mol Evol*, **48**: 142-150.

6. Charles H., Mouchiroud D., Lobry J., Gonçalves I., Rahbé Y., 1999. Gene size reduction in the bacterial aphid endosymbiont, *Buchnera*. *Mol Biol Evol*, **16**: 1820-1822.

7. Heddi A., Grenier A.M., Khatchadourian C., Charles H., Nardon P., 1999. Four intracellular genomes direct weevil biology: nuclear, mitochondrial, principal endosymbiont and *Wolbachia*. *Proc Natl Acad Sci USA*, **96**: 6814-6819.

Charles H., Nardon P., 1999. Intracellular symbiotic bacteria within insects. In: Seckbach J. (Ed.), Cellular Origin and Life in Extreme Habitats. "Enigmatic Microorganisms and Life in Extreme Environments", Vol. 1. Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 651-660.

Heddi A., Charles H., Rahbé Y., 1999. Comparative and phylogenetic aspects of two insect endocytobioses: aphids and weevils. In: Boemare N., Richardson P., Coudert F. (Eds.), Taxonomy, phylogeny and gnotobiological studies of entomopathogenic nematode bacterium complexes, pp. 37-44.

8. Grenier A.M., Wajnberg E., Charles H., Nardon P., 2000. Variabilité de facteurs biologiques et comportementaux dans les populations des trois espèces de *Sitophilus* (Coleoptera: Curculionidae) inféodées aux céréales et produits dérivés, et leur importance dans la caractérisation des espèces. *Ann Soc Entomol Fr*, **36**: 223-238.

9. Pintureau B., Chaudier S., Lassablière F., Charles H., Grenier S., 2000. Addition of *wsp* sequences to the *Wolbachia* phylogenetic tree and stability of the classification. *J Mol Evol*, **51**: 374-377.

10. Charles H., Heddi A., Rahbé Y., 2001. A putative insect intracelllar endosymbiont stem clade, whithin the Enterobacteriaceae, infered from

phylogenetic analysis based on a heterogeneous model of DNA evolution. *C R. Acad Sci Paris*, **324**: 489-494.

11. Heddi A., Charles H., Khatchadourian C., 2001. Intracellular bacterial symbiosis in the genus *Sitophilus*: the "Biological Individual" concept revisited. *Research in Microbiology*, **152**: 431-437.

Nardon P., Charles H., 2001. Morphological aspects of symbiosis. In: Seckbach J. (Ed.), Cellular Origin and Life in Extreme Habitats. "Symbiosis, Mechanisms and Model Systems", Vol. 4. Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 13-44.

12. Nardon P., Lefèvre C., Delobel B., Charles H., Heddi A., 2002. Occurence of endosymbiosis in Dryophtoridae weevils: Cytological insights into bacterial symbiotic structures. *Symbiosis*, **33**: 227-241.

13. Pintureau B., Grenier S., Heddi A., Charles H., 2002. Biodiversity of *Wolbachia* and of their effects in Trichogramma. *Ann Soc Entomol Fr*, **38**(4): 333-338.

14. Nardon P., Charles H., Delobel B., Lefèvre C., Heddi A., 2003. Symbiosis in the Dryophthoridae weevils (Coleoptera, Curculionoidea): morphological variability of symbiotic intracellular bacteria. *Symbiosis,* **34**: 231-235.

15. Sauvion N., Charles H., Rahbé Y., 2004. Effects of jackbean lectin (ConA) on the feeding behaviour and kinetics of intoxication of the pea aphid, *Acyrthosiphon pisum*. *Entomol Exp Appl,* **110**: 31-44.

16. Reymond N., Charles H., Duret L., Calevro F., Beslon G. and Fayard J. M., 2004. ROSO: optimizing oligonucleotide probes for microarrays. *Bioinformatics*, **20**: 271-273.

17. Calevro F., Charles H., Reymond N., Dugas V., Cloarec J. P., Bernillon J., Rahbé Y., Febvay G., Fayard J. M., 2004. Assessment of 35mer amino-modified oligonucleotide based microarray with bacterial samples. *J Microbiol Meth,* **57**: 207-218.

18. Lefèvre C., Charles H., Vallier A., Delobel B., Farrell B., Heddi A., 2004. Endosymbiont phylogenesis in the Dryophthoridae weevils: evidence for bacterial replacement. *Mol Biol Evol,* **21**: 965-973.

Reymond N., Charles H., Rome S., Marti J., 2004. Les données d'expression. In Informatique pour l'analyse du transcriptome, O. Gandrillon and J.F. Boulicaut (Eds), Hermes Science Publications - Lavoisier, Paris (France), pp. 45-65.

19. Pascal C., Pintureau B., Charles H., Katchadourian C., Grenier S., Bolland P., Robin C., 2005. Relationship between *Wolbachia* density and sex-ratio in a Trichogramma strain. *Agrociencia*, **8**:11-22.

20. Laforest F., Tchounikine A., Chaari T., Charles H., Calevro F., 2005. SITRANS: a Web Information System for Microarray Experiments. *Stud Health Technol Inform*, **116**: 33-8.

7.2. Congrès nationaux et internationaux

Le classement ci-dessous est chronologique.

Charles, H., Nardon P., 1995. A chaperonin in the endocytobiosis of the weevil *Sitophilus oryzae*. VII^{ème} European Congress on Biotechnology, Nice (France), février 1995. [Poster]

Charles H., Nardon P., 1995. Interactive proteins in the weevil *Sitophilus* endocytobiosis. 6th International Colloquium on Endocytobiology and Symbiosis, Tübingen (Allemagne), septembre 1995. [Exposé oral]

Charles H., Nardon P., 1996. Analysis of specific proteins of symbiosis in the weevil *Sitophilus*. XXth International Congress of Entomology, Firenze (Italie), août 1996. [Exposé oral]

Nardon P., Heddi A., Grenier A.M., Charles H., 1997. Density control of the principal endosymbiote of the weevil *Sitophilus oryzae*. Second International Congress on Symbiosis, Woods Hole (Boston, USA), avril 1997. [Communication orale]

Charles H., Heddi A., Nardon P., 1997. A chaperonin-like in the principal endocytobiotes of the weevil *Sitophilus*. Sixth International Colloquium on Endocytobiology and Symbiosis, Tübingen (Allemagne), septembre 1995. [Communication orale]

Charles H., Heddi A., Nardon P., 1997. Caractérisation moléculaire de la bactérie principale du charançon des céréales *Sitophilus oryzae* (Coleoptera, Curculionidae). XII^{ème} Colloque de Physiologie de l'Insecte, Nice (France), octobre 1997. [Poster]

Heddi A., Charles H., Rahbé Y., 1999. Comparative and phylogenetic aspects of two insect endocytobioses: aphids and weevils. Bacterial symbionts on survival of entomopathogenic nematodes, COST meeting, Wellesbourne (UK), avril 1998. [Communication orale]

Charles H., Rahbé Y., Nardon P., Heddi A., 1999. Les bactéries intracellulaires des insectes: Phylogénie et évolution moléculaire. Rencontre des Microbiologistes de l'INRA, Dourdan (France), avril 1999. [Communication orale]

Heddi A., Charles H., Khatchadourian C., 2000. Coexistence de deux bactéries symbiotiques intracellulaires dans les charançons du genre *Sitophilus* :

aspects physiologiques et évolutifs. Fonctions de la Biodiversité Microbienne : comprendre pour gérer, Aussois (France), mai 2000. [Communication orale]

Pintureau B., Grenier S., Heddi A., Charles H., 2000. Biodiversité des *Wolbachia* et des effets que ces bactéries induisent chez leurs hôtes. Fonctions de la Biodiversité Microbienne : comprendre pour gérer, Aussois (France), mai 2000. [Communication orale]

Nardon P., Charles H., Delobel B., Lambert A., Nardon C., Heddi A., 2000. Biodiversity of symbiosis in the Rhynchophoridae weevils. XVIIIth International Congress of Zoology, Athènes (Grèce), août-septembre 2000. [Poster]

Charles H., Reymond N., Rahbé Y., Heddi A., Febvay G., Fayard J.M., 2001. Etude du transcriptome de *Buchnera aphidicola*. IPG'01, Villeurbanne (France), avril 2001. [Poster]

Lefèvre C., Charles H., Delobel B., Nardon P., Heddi A., 2001. Caractérisation et phylogénie des bactéries symbiotiques intracellulaires des charançons de la famille des Rhynchophoridae. IPG'01, Villeurbanne (France), avril 2001. [Poster]

Reymond N., Charles H., Beslon G., Fayard J.M., 2001. Développement d'un logiciel d'optimisation de sondes oligonucléotidiques destinées aux puces à ADN. JPGD'01, Villeurbanne (France), avril 2001. [Poster]

Heddi A., Lefèvre C., Charles H., Khatchadourian C., 2001. Weevils and intracellular bacterial associations: impact on insect development and evolution. 9th International Congress on Invertebrate Reproduction and Developement, Grahamstown (Afrique du Sud), juillet 2001. [Communication orale]

Lefèvre C., Charles H., Delobel B., Nardon P., Vallier A., Heddi A., 2001. Caractérisation et phylogénie des bactéries symbiotiques intracellulaires des charançons de la famille des Dryophthoridae. Réseau "Biodiversité et Ecologie Microbienne" : adaptation, variabilité, diversité, Aussois (France), septembre 2001. [Communication orale]

Beslon G., Charles H., Fayard J.M., Gandrillon O., Mazet O., Soula H., 2002. Functional space approach for evolutionary algorithms. Les Journées de Post-Génomique de la Doua - JPGD'02, Villeurbanne (France), mars 2002. [Poster]

Beslon G., Charles H., Fayard J.M., Gandrillon O., Mazet O., 2002. Modeling evolution through a genetic regulation network. Les Journées de Post-Génomique de la Doua - JPGD'02, Villeurbanne (France), mars 2002. [Poster]

Calevro F., Charles H., Reymond N., Cloarec J.P., Rahbé Y., Heddi A., Febvay G., Fayard J.M., 2002. Transcriptome analysis in *Buchnera*: development of a first DNA chip. Les Journées de Post-Génomique de la Doua - JPGD'02, Villeurbanne (France), mars 2002. [Poster]

Charles H., Fayard J.M., Laforest F., Tchounikine A., 2002. Transcriptome information system for the Rhône-Alpes genopole platform. Les Journées de Post-Génomique de la Doua - JPGD'02, Villeurbanne (France), mars 2002. [Poster]

Charles H., 2002. Présentation des données biologiques à l'attention des non biologistes. Les Journées de Post-Génomique de la Doua - JPGD'02, Villeurbanne (France), mars 2002. [Communication orale - Atelier]

Reymond N., Charles H., Beslon G., Fayard J.M., 2002. ROSO: a software to search optimized oligonucleotide probes for microarrays. Les Journées de Post-Génomique de la Doua - JPGD'02, Villeurbanne (France), mars 2002. [Poster]

Lefèvre C., Charles H., Delobel B., Nardon P., Vallier A., Heddi A., 2002. Origine des bactéries symbiotiques intracellulaires des Dryophtoridae : analyse phylogénétique basée sur un modèle d'évolution de l'ADN non homogène. Intersciences 2002, Colloque des doctorants scientifiques de Lyon, Villeurbanne (France), avril 2002. [Poster]

Reymond N., Charles H., Beslon G., Fayard J.M., 2002. ROSO : un logiciel de recherche et d'optimisation des sondes oligonucléotidiques destinées aux puces à ADN. Intersciences 2002, Colloque des doctorants scientifiques de Lyon, Villeurbanne (France), avril 2002. [Poster]

Reymond N., Charles H., Beslon G., Fayard J.M., 2002. ROSO : un logiciel de recherche et d'optimisation des sondes oligonucléotidiques destinées aux puces à ADN. Journées Ouvertes Biologie Informatique Mathématiques (JOBIM), Saint-Malo (France), juin 2002. [Poster]

Lefèvre C., Charles H., Delobel B., Nardon P., Vallier A., Heddi A., 2002. Origine des bactéries symbiotiques intracellulaires des Dryophtoridae : analyse phylogénétique basée sur un modèle d'évolution de l'ADN non homogène. V^{ème} Conférence Internationale Francophone d'Entomologie, Montréal (Canada), juillet 2002. [Communication orale]

Reymond N., Charles H., Beslon G., Duret L., Fayard J.M., 2002. ROSO : A Software to Search Optimized Oligonucleotide Probes for Microarrays. The X^{ème} International Conference on Intelligent Systems for Molecular Biology (ISMB), Edmonton (Canada), Aout 2002. [Poster]

Charles H., 2002. ROSO: A software to search optimized oligonucleotide probes for microarrays. COST 853: Agricultural Biomarkers for Array-

Technology, Waedenswill (Suisse), Septembre 2002. [Communication orale (invité)]

Lefèvre C., Charles H., Delobel B., Nardon P., Heddi A., 2002. Histoire évolutive d'une symbiose : du parasitisme au mutualisme chez les charançons Dryophthoridae. 13ème Rencontres Régionales de la Recherche, Grenoble (France), septembre 2002. [Poster]

Heddi A., Lefèvre C., Charles H., 2002. Impact of endosymbiosis on the Dryophthoridae insect diversity. VIIth European Congress of Entomology, Thessaloniki (Grèce), octobre 2002. [Communication orale]

Calevro F., Charles H., Reymond N., Dugas V., Cloarec J.P., Bernillon J., Rahbé Y., Febvay G., Fayard J.M., 2003. Transcriptome analysis of *Buchnera*, the intracellular symbiotic bacteria of aphids: a methodological approach. XIVème Colloque de Physiologie de l'Insecte, Amiens (France), avril 2003. [Communication orale]

Rahbé Y., Al Ayoubi Z., Calevro F., Charles H., Fayard J.M., Febvay G., Heddi A., 2003. *Buchnera aphidicola*, bactérie endosymbiotique des pucerons : génomes et physiologie symbiotique. XIVème Colloque de Physiologie de l'Insecte, Amiens (France), avril 2003. [Communication orale]

Charles H., Calevro F., Reymond N., Laforest F., Tchounikine A., Beslon G., Soula H., Dugas V., Cloarec J.P., Bernillon J., Rahbé Y., Febvay G., Fayard J.M., 2003. Bioinformatique du transcriptome de *Buchnera*, symbiote intracellulaire des pucerons. Rencontres 2003 des Microbiologistes de l'INRA, Dourdan (France), mai 2003. [Poster]

Heddi A., Lefèvre C., Charles H., Vallier A., 2003. La symbiose bactérienne intracellulaire chez les insectes Dryophthoridés : aspects moléculaires et évolutifs. Rencontres 2003 des Microbiologistes de l'INRA, Dourdan (France), mai 2003. [Communication orale]

Charles H., 2003. Choix des séquences sondes à déposer sur les puces. Formation CNRS : Puces à ADN, Villeurbanne (France), Juin 2003. [Communication orale – Atelier]

Calevro F., Charles H., Reymond N., Dugas V., Cloarec J. P,. Bernillon J., Rahbé Y., Febvay G., Fayard J. M., 2003. A preliminary methodological approach to study the transcriptome of *Buchnera*, the intracellular symbiotic bacteria of aphids. Journées de Post-Génomique de la Doua - JPGD'03. Novembre 2003. [Poster]

Reymond N., Charles H., Beslon G., Duret L., Fayard J. M., 2003. ROSO: a software to search optimized oligonucleotide probes for microarrays. 14ème Rencontres Régionales de la Recherche. Saint-Etienne (France), octobre 2003. [Poster]

Lefèvre C., Charles H., Vallier A., Farrell B., Heddi A., 2004. Phylogeny of intracellular symbiotic bacteria in the Dryophthoridae family: evidence for bacterial replacement in insect symbioses. Conférence Jacques Monod "Ecology and Evolution of Host-Parasite Relationships". Roscoff (France), septembre 2004. [Poster]

Rouillat M. H., Dugas V., Calevro F., Cloarec J. P., Charles H., Phaner-Goutorbe M., 2004. Influence of probe sequence and strand length on the morphology and structuration of DNA chip. Nemo 8, 8th International Conference of nanometer scale science and technology.

Calevro F., Reymond N., Morin N., Dumitru A., Rahbé Y., Febvay G., Bernillon J., Laugier C., Fayard J. M., Charles H., 2005. Analyse transcriptomique de la réponse de *Buchnera aphidicola* à un stress métabolique subi par son hôte le puceron du pois, *Acyrthosiphon pisum*. Journées Réseau Ecologie et Interactions Durables (REID). Lyon (France), Janvier 2005. [Communication orale]

Charles H., Calevro F., 2005. Gene expression measurements with microarray technology in bacteria: transcriptome analysis of the symbiotic bacteria, *Buchnera aphidicola*, in nutritional stress conditions of its aphid host *Acyrthosiphon pisum*. Workshop on statistical analysis of postgenomic data. INAPG Paris (France), avril 2005. [Communication orale]

Calevro F., Reymond N., Morin N., Rahbé Y., Bernillon J., Febvay G., Laugier C., Fayard J.M., Charles H., 2005. Global transcriptional responses of *Buchnera aphidicola* to aromatic amino acid limitation in the diet of its symbiotic partner, the phloem sap feeding insect *Acyrthosiphon pisum*. Society of Experimental Biology Annual Main Meeting 2005 (SEB 2005). Comparative Biochemistry and Physiology Part A, Barcelona (Espagne), juillet 2005. [Communication orale]

Rahbé Y., Calevro F., Charles H., Febvay G., Tagu D., 2005. Of lice and plants: metabolite flows in the trophic pathway of phloem feeding insects as analysed through genomic and EST data from the aphid symbiosis. Society of Experimental Biology Annual Main Meeting 2005 (SEB 2005). Comparative Biochemistry and Physiology Part A, Barcelona (Espagne), juillet 2005. [Communication orale]

Charles H., Calevro F., 2005. Gene expression measurements with microarray technology in bacteria: transcriptomic analysis af the symbiotic bacteria, *Buchnera aphidicola*, in nutritional stress conditions of its aphid host *Acyrthosiphon pisum*. Journées Ouvertes Biologie Informatique Mathématiques (JOBIM) - Journée satellite Vicanne. Lyon (France), juillet 2005. [Communication orale]

Charles H., Calevro F., Reymond N., Morin N., Rahbé Y., Bernillon J., Laugier C., Febvay, G., Fayard J.M., 2005. Genomewide transcriptional changes

associated with nutritional alterations affecting phenylalanine and tyrosine metabolism in *Buchnera aphidicola*. Journées Ouvertes Biologie Informatique Mathématiques (JOBIM). Lyon (France), juillet 2005. [Poster]

Calevro F., Reymond N., Vinuelas J., Groppi A., Barre A., Rahbé Y., Febvay G., Douglas A., Fayard J.M., Charles H., 2005. *Buchnera aphidicola* : evidences for transcriptional regulation in a degenerated genome. Integrative Post-Genomics (IPG'05). Villeurbanne (France), novembre 2005. [Poster]

Charles H., Calevro F., Vinuelas J., 2005. Transcriptional responses of *Buchnera aphidicola* to combined amino acid and osmotic stress in the diet of its symbiotic partner, *Acyrthosiphon pisum*. Meeting COST 853, Agricultural Biomarkers for Array-Technology (WG 4 Chip production and analysis and WG 5 Environmental monitoring),. Lyon (France), novembre 2005. [Communication orale]

Vinuelas J., Calevro F., Bernillon J., Rahbé Y., Febvay G., Fayard J.M., Charles H., 2005. Genomic DNA: an attractive candidate for microarray data normalization. Integrative Post-Genomics (IPG'05). Villeurbanne (France), novembre 2005. [Poster]

7.3. Articles en préparation et/ou soumis

Reymond N., Calevro F., Viñuelas J., Morin N., Rahbé Y., Febvay G., Douglas A., Fayard JM., Charles H. Different level of transcriptional regulation to trophic constraints in the reduced genome of *Buchnera aphidicola APS* (soumis).

Charles H.**,** Calevro F., Perrière G., Fayard JM., rahbé Y. Codon usage bias and tRNA over-expression in *Buchnera aphidicola* after aromatic amino acid nutritional stress on its host *Acyrthosiphon pisum*. (soumis).

Crisan D., Chaari T., Laforest F., Tchounikine A., Calevro F., Teletchea F., Bernillon J. Oger C., Fayard J.M. and Charles H., Microarray data management using the Web information system SITRANS. (en préparation)

Charles H. Analyses statistiques et intégration des données d'expression (en préparation)

DEUXIEME PARTIE

Activités de recherche

Chapitre 1

Présentation du modèle et de la problématique biologique

1. *Acyrthosiphon pisum*, le puceron hôte

Le puceron du pois *Acyrthosiphon pisum* (Harris) est un ravageur notable des régions tempérées. Relativement polyphage, *A. pisum* s'attaque à de nombreuses légumineuses d'intérêt agronomique (pois, fèves, lentilles, luzerne). Les dégâts qu'il occasionne aux cultures sont liés à la ponction de sève phloémienne qui affaiblit la plante, cet effet étant surtout sensible sur les très jeunes pousses. Sur des plantes plus âgées, le rejet de miellat favorise la contamination par des champignons, la fumagine. Mais le dommage le plus important est sans doute sa capacité à transmettre de très nombreux virus végétaux, même sur des plantes sur lesquelles il ne se nourrit que très rarement (virus APV et AMV des légumineuses, virus Y de la pomme de terre).

La biologie de *A. pisum* est assez complexe car il existe dans cette espèce plusieurs types de populations, parfois appelées races, caractérisées notamment par des préférences alimentaires différentes (Dixon *et al.* 1987). Schématiquement, le cycle commence au printemps ou une jeune larve émerge d'un œuf d'hiver pondu sur une légumineuse persistante. Cette jeune larve va se transformer en femelle ailée (fondatrice) qui va alors donner naissance par parthénogenèse à des femelles aptères, les fondatrigènes. Lorsque la population deviendra trop dense, des émigrants ailés seront produits et s'envoleront pour rechercher l'hôte primaire, une légumineuse annuelle. Pour certaines races, ce changement d'hôte n'existe pas. Durant la période estivale, des cycles de reproductions parthénogénétiques se succéderont

(virginipares aptères et ailés). Les virginipares aptères possèdent une fécondité maximale, les embryons localisés dans le ventre de leur mère portent déjà en eux les embryons de leur descendance ; on parle de télescopage des générations (figure 1). Ce sont ces populations clonales qu'on entretient au laboratoire. Durant l'automne, lorsque les conditions climatiques deviennent moins favorables et avec la diminution de la photopériode, les virginipares vont donner naissance à des individus sexués. Ceux-ci se croiseront et les femelles sexuées iront chercher leur plante hôte secondaire pour y déposer quelques œufs d'hiver. L'obtention en laboratoire de générations sexuées et d'œufs demande une très grande expertise et n'est pas maîtrisée au Laboratoire BF2I.

Figure 1. *A. pisum*, femelles ailées larves et adultes aptères (à gauche) ; coupe schématique d'un puceron (*Myzus persicae*) avec le chapelet de bactériocytes dans la cavité interne formant le bactériome (à droite) d'après (Minks et Harrewijn 1987).

Les pucerons se nourrissent exclusivement de sève phloémienne. Les principaux composés azotés du phloème sont des acides aminés libres dont la concentration peut atteindre jusqu'à 800 mM. Parmi les acides aminés les plus représentés, on trouve l'aspartate, le glutamate, l'asparagine et la glutamine qui sont des acides aminés non essentiels, c'est-à-dire que le puceron peut les synthétiser (Sandstrom et Moran 2001). A l'inverse, les acides aminés essentiels sont souvent très rares. Ils sont au nombre de neuf : l'histidine, l'isoleucine, la leucine, la lysine, la méthionine, la phénylalanine, la thréonine, le tryptophane et la valine. Le phloème contient également des substances inorganiques (potassium, phosphate et micro-nutriments) en quantité relativement abondante, de l'ordre de 1 à 5 $g.L^{-1}$. En revanche, les lipides et les stérols sont quasiment absents de ce milieu (Sandstrom *et al.* 2000 ; Sandstrom et Moran, 2001).

Pour absorber les nutriments présents en très faible concentration, le puceron est obligé d'absorber des quantités très importantes de sève. Il doit donc évacuer le sucre excédentaire de son tube digestif sous peine de se vider de son eau par effet d'osmolarité (Wilkinson *et al.* 1997). En effet, la sève phloémienne possède une concentration en saccharose très élevée. Les pucerons ingèrent ce sucre à des taux bien supérieurs à leurs propres besoins en carbone et de fortes concentrations de saccharose non assimilé sont éliminées dans le miellat (Rhodes *et al.* 1997 ; Douglas *et al.* 2001). Wilkinson *et al.* (1997) ont montré que l'osmorégulation est contrôlée par le puceron hôte. Plus précisément, ce contrôle passe par l'activité de l'α-glucosidase du tube digestif. A faible concentration de glucose, cette enzyme hydrolyse le saccharose pour former du glucose et du fructose. Lorsque la pression osmotique augmente, l'enzyme polymérise des molécules de sucres (glucose et saccharose) pour former des oligosaccharides. Cette production a pour effet de diminuer la pression osmotique du tube digestif et les oligosaccharides sont ensuite excrétés dans le miellat. L'osmorégulation est très efficace chez les pucerons puisqu'il a été montré que le miellat est excrété à une pression osmotique à peu près équivalente à celle de l'hémolymphe. Cette régulation ne fait pas intervenir la bactérie symbiotique *Buchnera*. Néanmoins, il a été montré une sensibilité un peu plus importante à un choc osmotique pour des pucerons aposymbiotiques par rapport à des pucerons symbiotiques. Le rôle de *Buchnera* dans ce mécanisme de régulation de la pression osmotique serait indirect, il consisterait à augmenter l'assimilation du glucose à cause de la consommation bactérienne.

Enfin, une autre façon pour le puceron d'acquérir des nutriments essentiels est de les « faire » synthétiser par sa bactérie symbiotique intracellulaire, *Buchnera*. L'importance du rôle nutritionnel de la bactérie reflète l'inadéquation du substrat alimentaire de l'insecte hôte. Les deux partenaires symbiotiques sont devenus entièrement dépendants l'un de l'autre au cours de leur coévolution datant de près de 150 millions d'années. *Buchnera* peut survivre quelques heures en dehors de son hôte, mais n'est pas cultivable *in vitro* (Whitehead et Douglas 1993). A l'inverse, il est possible de débarrasser les pucerons de leur bactérie symbiotique, mais ceux-ci se développent beaucoup moins bien et deviennent stériles (Douglas 2003). L'interaction trophique *Buchnera* - puceron est au cœur de notre problématique. Nous avons étudié plus particulièrement l'apport de la bactérie en acides aminés.

2. *Buchnera aphidicola*, la bactérie symbiotique intracellulaire des pucerons

Biologie et histologie

Tous les groupes d'insectes se nourrissant de sève comme seule ou principale source de nourriture sont dépendants de microorganismes symbiotiques pour leur développement et leur reproduction. La bactérie *Buchnera aphidicola* a été

découverte par Buchner en (1965). Elle est localisée dans le cytoplasme de cellules spécialisées, appelées bactériocytes. On trouve des bactériocytes dans la cavité interne du puceron ; ils sont plus ou moins associés entre eux pour former un organe, le bactériome (figure 1). Des bactériocytes sont également présents à l'intérieur des embryons de pucerons. Le passage de *Buchnera* d'une génération à l'autre nécessite une phase de réinfection de l'embryon à un stade très précoce de son développement. Ce phénomène est encore relativement mal connu (Humphreys et Douglas 1997 ; Braendle *et al.* 2003, Wilkinson *et al*. 2003).

Les bactéries ne gisent pas librement dans le cytoplasme hôte, elles sont entourées d'une membrane de type eucaryote appelée symbiosome (figure 2). La caractérisation de cette membrane et de l'espace périplasmique qu'elle délimite est très importante pour le décryptage des interactions symbiotiques, mais peu de travaux ont été consacrés à ce sujet jusqu'à présent (Houk *et al.* 1976 ; 1977).

Buchnera est une bactérie de forme sphérique, d'un diamètre de deux à cinq microns, ce qui fait d'elle une bactérie de grosse taille (figure 2). Elle appartient au groupe des γ3-Protéobactéries très proches des entérobactéries et notamment de *E. coli*. Sa position phylogénétique a été beaucoup discutée dans la littérature (Munson *et al.* 1991 ; Charles *et al.* 2001 ; Herbeck *et al.* 2005). La concentration des *Buchnera* dans un puceron est d'environ 10^7 cellules par mg de puceron frais (Baumann *et al.* 1995).

Buchnera est accompagnée, chez de nombreuses espèces de pucerons, par d'autres bactéries symbiotiques, dites endocytobiotes « secondaires ». Ces dernières sont de plusieurs types et n'infectent pas toutes les populations de pucerons d'une même espèce. Elles appartiennent toutes au groupe des Entérobactériacées, leur rôle physiologique est mal connu et a longtemps été négligé (Oliver *et al.* 2005 ; Scarborough *et al.* 2005). L'équipe japonaise du Dr. Fukatsu a beaucoup travaillé sur ces bactéries secondaires, elle a montré notamment qu'il pouvait être possible d'éliminer *Buchnera*, et que les bactéries secondaires restauraient une part de la fitness (croissance et fertilité) du puceron (Fukatsu et Ishikawa 1998 ; Fukatsu et al. 2000 ; Koga *et al.* 2003 ; Tsuchida *et al*. 2004). Il est à noter que le clone LL01 de *A. pisum* utilisé dans notre travail est dépourvu de symbiotes secondaires. Enfin, des *Wolbachia* et des *Rickettsia* ont été trouvées dans des souches naturelles des pucerons du pois et du cèdre (Gomez-Valero *et al.* 2004 ; Sakurai *et al.* 2005). Ce type de symbiose à α-Protéobactéries semble néanmoins plus rare chez les pucerons, et le clone LL01 n'est pas contaminé.

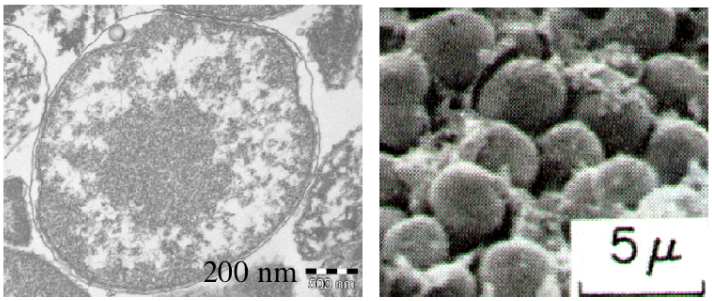

Figure 2. *Buchnera* vue au microscope électronique à transmission (à gauche), le symbiosome est bien visible. Vue en balayage (à droite).

Evolution du génome de Buchnera

L'enfermement de la bactérie dans le bactériocyte a des conséquences importantes sur son évolution. Les populations de *Buchnera* ont des effectifs faibles et sont totalement isolées les unes des autres. Durant toute la phase de reproduction clonale du puceron, il ne peut pas y avoir de recombinaisons génétiques entre les populations de *Buchnera* séquestrées dans leurs hôtes. A chaque génération d'hôte, ces populations subissent des réductions drastiques d'effectifs (goulets d'étranglements), car seuls quelques individus vont contaminer les embryons. Il n'est pas certain qu'un brassage génétique (transmission par les deux sexes) existe au moment de la reproduction sexuée. Cette dynamique de population très particulière, appelée « cliquet de Muller » augmente de façon drastique la vitesse d'évolution du génome de *Buchnera.* Les portions non codantes évoluent rapidement par dérive génétique tandis que les parties codantes accumulent des mutations délétères à des taux relativement élevés. Les biais mutationnels ont ainsi une influence très forte sur la composition en base du génome (Rocha et Danchin, 2002). Le génome de *Buchnera* se caractérise par un taux de GC de l'ordre de 30 %, et des protéines semblent avoir perdu certaines de leurs fonctionnalités. Ces résultats ont été très largement débattus dans la littérature (Moran *et al.* 1995 ; Moran, 1996).

Le génome de *Buchnera* est très réduit (400 à 600 kb) chez tous les pucerons (Charles et Ishikawa 1999 ; Gil *et al.* 2002). Cette réduction s'explique par l'évolution en milieu intracellulaire de *Buchnera* qui a ainsi perdu tous les gènes qui ne lui étaient plus nécessaires dans cet environnement stabilisé. La réduction du génome semble s'être déroulée dans les tous premiers temps de l'association symbiotique. En effet, la remarquable synténie entre les trois génomes des *Buchnera* issues des pucerons du pois, des céréales et du pistachier montre que les génomes n'ont pratiquement plus évolué en taille depuis près de 50 à 70 millions d'années (Moran et Mira 2001b ; Silva *et al.* 2001 ; 2003). Les raisons de cette stase sont liés

à la perte des principaux gènes impliqués dans les phénomènes de recombinaison et de réparation de l'ADN d'une part, et à la perte des séquences répétées d'autre part.

Komaki ei Ishikawa (1999) ont montré chez *A. pisum* une très forte ploïdisation de *Buchnera*, puisqu'ils estiment entre 25 et 100 le nombre de copies du génome de la bactérie. Ces auteurs montrent encore que le nombre de copies du génome de la bactérie varie avec l'âge de l'insecte (Komaki et Ishikawa 2000). Il est possible qu'une certaine régulation de l'expression des gènes puisse dériver de la régulation de la ploïdie chez *Buchnera*, néanmoins aucune preuve n'a pu être apportée jusqu'à présent.

Depuis le séquençage de la bactérie symbiotique de A. *pisum* (Shigenobu *et al.* 2000), trois autres génomes de *Buchnera* ont été séquencés : celui du puceron des céréales *Schizaphis graminum* (Tamas *et al.* 2002), du pistachier *Baizongia pistaciae* (Van Ham *et al.* 2003) et du cèdre *Cinara cedri* (Moya *et al.*, non publié). Chez *A. pisum*, le génome de *Buchnera* comporte 617 gènes dont neuf pseudogènes et dix gènes localisés sur des plasmides (pLeu et pTrp).

Capacités métaboliques de Buchnera

Buchnera dispose d'un métabolisme aérobie fonctionnel. Le bactériocyte est une cellule riche en mitochondries et le bactériome est bien alimenté en oxygène par le système trachéen de l'insecte. *Buchnera* peut synthétiser l'ATP (opéron *atp*), le NADH (opéron *nuo*) et le cytochrome C (opéron *cyo*). La glycolyse et la voie des pentoses sont bien conservées. Il existe un système PTS de transport de glucose qui semble fonctionnel. A l'inverse, le métabolisme anaérobie et de fermentation est très dégradé.

Buchnera est spécialisée dans la biosynthèse des acides aminés. Alors que seulement 2 % (environ une centaine) des gènes de *E. coli* codent des enzymes de la biosynthèse des acides aminés, *Buchnera* en compte 10 % (une soixantaine). Le métabolisme des acides aminés est une bonne illustration de la mise en place d'une interdépendance des partenaires dans l'évolution de l'association symbiotique. En effet, la figure 3 montre chez *Buchnera* la perte des gènes codant les acides aminés non essentiels pour le puceron, alors que les gènes codant les acides aminés essentiels ont été conservés par la bactérie, sauf pour la tyrosine et la cystéine. Le cas de la tyrosine est très particulier, puisqu'en effet, cet acide aminé très important chez les insectes (biosynthèse de la cuticule) est fabriqué par le puceron à partir de la phénylalanine produite par la bactérie. La tyrosine est ensuite exportée par le puceron vers la bactérie. Ce mécanisme a été appelé navette tyrosine phénylalanine dans le système symbiotique, mais les transporteurs permettant ce transfert ne sont pas encore connus.

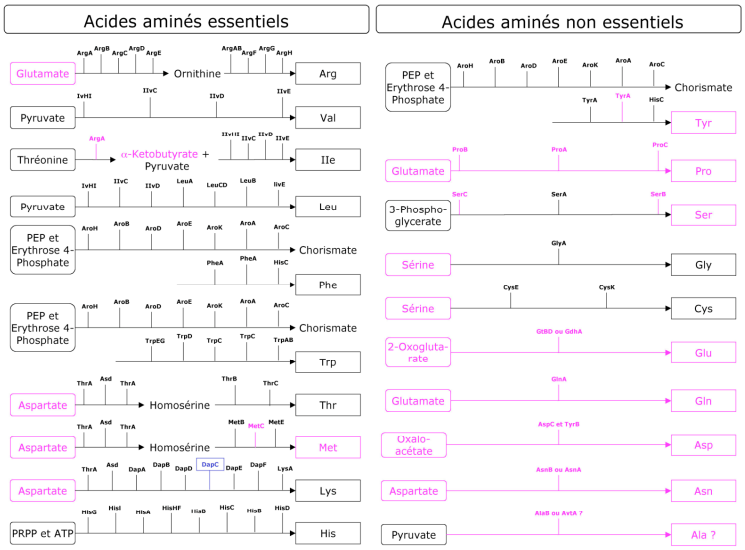

Figure 3 d'après Reymond (2004). Représentation des voies de biosynthèse des acides aminés essentiels (à gauche) et non essentiels (à droite). Les étapes pour lesquelles aucun gène n'a été annoté et les précurseurs non synthétisés *de novo* sont indiqués en rose. L'enzyme *ilvE* est absente des voies de biosynthèse de la valine, de l'isoleucine et de la leucine, mais une autre aminotransférase assure sans doute ces étapes. Dans la voie de biosynthèse de la lysine, le gène *dapC* n'a été identifié chez aucun organisme. Enfin, la dernière étape de biosynthèse de la phénylalanine est catalysée par *TyrB* chez *Escherichia coli*. Cette étape est assurée par une autre aminotransférase non définie chez *Buchnera*.

Le métabolisme des lipides chez *Buchnera* est extrêmement dégradé. La bactérie ne produit pas de lipopolysaccharides, ni de phospholipides. Ces composés sont pourtant essentiels à la constitution de la membrane plasmique cellulaire. Une importation des lipides depuis la cellule bactériocytaire est donc très probable.

Buchnera possède un arsenal d'enzymes relativement faible pour la réparation des dommages de l'ADN et les mécanismes de recombinaison. Ce point a été très discuté dans la littérature (Van Ham *et al.* 2003).

Alors que la bactérie est spécialisée dans l'échange de métabolites, seuls 18 gènes (3 %) du génome de *Buchnera* sont associés à cette fonction. A titre de comparaison, *E. coli* possède 661 gènes (15 %) annotés comme transporteurs. Ce point particulier, élément clé de notre travail, sera discuté à plusieurs reprises dans ce mémoire.

Régulation de la transcription chez Buchnera

La production des acides aminés est régulée chez les bactéries libres par un réseau complexe incluant des régulations transcriptionnelles, des mécanismes d'atténuation et d'inhibition allostérique des enzymes par les produits finaux des réactions (Panina *et al.* 2003). La délocalisation sur deux plasmides des gènes codant les enzymes de la voie terminale de la biosynthèse de la leucine (plasmide pLeu) et du Tryptophane (plasmide pTrp) pourrait correspondre à une adaptation de *Buchnera* pour la régulation transcriptionnelle de l'expression de ces gènes, importants dans la fonction symbiotique. Néanmoins, aucune preuve expérimentale n'a pu être donnée à ce sujet pour le moment.

L'analyse globale du génome de *Buchnera* montre que la bactérie a perdu de très nombreux régulateurs transcriptionnels (Shigenobu *et al.* 2000). Ainsi, les grandes voies de signalisation utilisant l'adénylate cyclase (gène *cyaA*) ou l'AMPc (gène du récepteur *crp*) sont absentes chez *Buchnera*. La bactérie symbiotique ne possède que deux facteurs sigma (σ^{32} et σ^{70}) codés par les gènes *rpoH* et *rpoD*. Dans le métabolisme des acides aminés, aucun des régulateurs transcriptionnels décrits chez *E. coli* n'est retrouvé dans le génome de *Buchnera*. Concernant le métabolisme des sucres, il a été montré chez *E. coli* que la plupart des gènes de la glycolyse étaient exprimés constitutivement ou sous le contôle de régulateurs transcriptionnels globaux. Au contraire, la voie des pentoses phosphates est contrôlée par de nombreux régulateurs transcriptionnels, mais aucun n'est conservé chez *Buchnera*. Les séquences génomiques de régulation (promoteurs et terminateurs) semblent également très altérées chez *Buchnera*. Moran et Mira (2001a) montrent que la perte des promoteurs est souvent associée à la formation d'un groupe de gènes contigus (avec pertes des gènes du brin opposé) favorisant ainsi la formation de nouvelles unités de transcription. Néanmoins, les études réalisées jusqu'à présent ne sont restées que théoriques et toujours assez peu approfondies. Une part des difficultés de ces recherches est liée au biais AT des régions intergéniques, ces séquences régulatrices étant souvent, elles aussi, très riches en bases A et T (de Hoon *et al.* 2005).

L'absence de régulateurs transcriptionnels annotés chez *Buchnera* ne signifie pas forcément que la bactérie ne régule plus la transcription de ces gènes. N. Reymond dans sa thèse de Doctorat (2004) énumère un certain nombre de mécanismes et d'hypothèses alternatives de régulation. Parmi celles-ci, elle évoque la régulation transcriptionnelle par des protéines du métabolisme comme dans la mitochondrie (Hall *et al.* 2004), par des modifications du superenroulement de l'ADN grâce à des polyamines (Feuerstein *et al.* 1991), ou encore par des régulateurs globaux comme les protéines « histone-like » (Dorman et Deighan 2003). Enfin, il n'est pas exclu que des régulateurs eucaryotes importés du bactériocyte puissent également intervenir dans la bactérie.

Des articles récents incluant des analyses de puces à ADN attestent du rôle prépondérant que peuvent avoir certains régulateurs globaux, comme les protéines « histone-like » (Dorman et Deighan 2003). Quatre « histone-like » ont été décrites chez *E. coli* : IHF (Integrating Host Factor, codée par les gènes *himA* et *himD*), FIS (Factor for Inversion Stimulation, gène *fis*), H-NS (Histone-like Nucleoid Structuring, gène *hns*) et HU (heat unstable, gènes *hupA* et *hupB*). Les protéines « histone-like » des bactéries ne possèdent pas d'homologie de séquence avec les histones des eucaryotes, mais elles montrent une forte affinité pour l'ADN. Le facteur IHF montre les plus fortes affinité et spécificité de fixation. FIS possède un site de fixation plus dégénéré et les deux autres facteurs sont relativement aspécifiques. Ces protéines sont responsables de la compaction du chromosome bactérien et interfèrent ainsi avec les mécanismes de la recombinaison, de la réplication et de la transcription. Leur rôle dans la régulation de la transcription est ambivalent (*dual regulator*), c'est-à-dire que les facteurs peuvent être activateur ou répresseur selon leur site de fixation ou la présence d'autres régulateurs. Le mécanisme d'action est toujours lié à la topologie locale de l'ADN au niveau du site promoteur et de l'ARN polymérase. On pourra se référer à la revue de Dorman et Geighan (2003) pour une analyse plus détaillée de ces mécanismes d'action. Chez *E. coli*, il a été montré que près de 5 % des gènes étaient régulés par le facteur IHF (Freundlich *et al.* 1992 ; Lynch *et al.* 2003). Le même résultat a été obtenu pour le facteur FIS chez *Salmonella enterica* (Kelly *et al.* 2004). Chez cette bactérie, les auteurs ont noté que le régulon FIS était responsable du maintien de la bactérie en condition intracellulaire grâce à l'induction de nombreux gènes de virulence, de transport et de métabolisme. Chez *Buchnera*, alors que la majorité des régulateurs transcriptionnels spécifiques ont été perdus, les quatre protéines « histone-like » de *E. coli* ont été conservées, à l'exception du gène *hupB* codant la sous unité β du facteur HU. Des traces de régulation de ces facteurs seront donc recherchées dans nos données d'expression (chapitre 4).

3. Analyse des interactions trophiques hôte - bactéries

La lutte contre les pucerons nécessite une bonne connaissance de leur biologie. Cette connaissance inclut notamment la physiologie de la reproduction, le comportement et la nutrition. Au laboratoire BF2I, nous avons travaillé essentiellement sur la nutrition de ces ravageurs. Les aspects comportementaux, avec la technique d'électropénétrographie (EPG), ont également été abordés, mais toujours dans le contexte des interactions trophiques entre la plante et l'insecte (Rahbé *et al.* 2000). Plus précisément encore, le laboratoire s'est spécialisé sur le rôle de la symbiose dans la biosynthèse des acides aminés. Les expériences de radiotraçage de métabolites conduites par G. Febvay sont une clé de voûte de notre expertise dans le domaine. Ces travaux ont apporté des connaissances sur les interactions trophiques de ce couple modèle, mais ils ont également permis de

développer des milieux artificiels adéquats pour l'élevage des pucerons en laboratoire, ouvrant ainsi la voie à de nouvelles expérimentations (Febvay *et al.* 1988). Ainsi, grâce au marquage successif de tous les acides aminés au ^{14}C dans des milieux artificiels, leur devenir a été analysé dans des pucerons symbiotiques et aposymbiotiques. Il a été montré, de façon très significative, que les pucerons aposymbiotiques ne sont pas capables de réaliser des interconversions entre acides aminés contrairement aux pucerons symbiotiques qui peuvent synthétiser tous les acides aminés à partir de quelques précurseurs non essentiels (Febvay *et al.* 1995 ; Liadouze *et al.* 1995). Le métabolisme du saccharose, composé majoritaire du phloème, a également été analysé. Lorsque du saccharose marqué au ^{14}C est fourni à des pucerons aposymbiotiques, aucune radioactivité n'est fixée dans les protéines. A l'inverse, les pucerons symbiotiques sont capables de métaboliser le saccharose pour l'incorporer dans tous les acides aminés, sauf l'histidine et l'arginine (Febvay *et al.* 1999). D'autres équipes ont travaillé sur ces aspects nutritionnels ; parmi les groupes les plus importants on peut citer les équipes d'A. Douglas (Université de York, Angleterre) (Douglas 1998), H. Ishikawa (Université de Tokyo, Japon) (Wilkinson et Ishikawa 1999), ou encore A. Moya (Université de Valence, Espagne) (Van Ham *et al.* 1997 ; 1999).

La sève phloémienne est un milieu très déséquilibré. C'est également un milieu de composition très variable entre les différentes plantes d'une part, mais également pour une même plante en fonction de son développement, de la luminosité et de la température par exemple. Les plantes développent, de plus, des mécanismes de défense qui interagissent avec le puceron au niveau comportemental (taux d'ingestion) ou physiologique (toxicité). Par ailleurs, les besoins nutritionnels du puceron sont également très variables avec le stade de développement de l'insecte. Si l'on comprend mieux au niveau métabolique l'apport de *Buchnera*, on ne sait pas grand-chose du rôle de la symbiose sur l'adaptabilité du puceron à cette variabilité nutritionnelle. C'est ce dernier point que nous avons analysé au niveau transcriptionnel.

Nous avons conçu deux expériences permettant de comparer des populations de *Buchnera* soumises à des demandes variables vis-à-vis de leur hôte. La première expérience est très ciblée et correspond à l'élimination des deux acides aminés tyrosine et phénylalanine dans l'alimentation du puceron. La deuxième expérience combine une déplétion globale en acides aminés essentiels avec une variation de la pression osmotique du milieu. Ces deux expériences sont décrites et analysées dans le chapitre 4.

Buchnera est décrite par certains auteurs comme une bactérie « dégénérée » (Moran 2003 ; Wilcox *et al.* 2003). On peut comprendre la position de ces auteurs qui ont essentiellement travaillé au niveau de l'évolution moléculaire sur *Buchnera*. Au laboratoire, nous nous positionnons plutôt au niveau de la physiologie, et nous

avons participé au décryptage du rôle nutritionnel de la bactérie dans le système symbiotique. Nous avons tendance à défendre une position plus fonctionnelle de la bactérie, sa « dégénérescence » pouvant être considérée comme une « hyperspécialisation ». On sait qu'une des adaptations les plus robustes pour résister à la variabilité des conditions expérimentales est de conserver une plasticité phénotypique importante. *Buchnera* a vraisemblablement perdu cette variabilité à tous les niveaux. Il n'y a pas de raison objective pour penser que l'apparente dégradation des mécanismes de régulation transcriptionnelle ne corresponde pas également à une spécialisation.

L'interprétation des résultats de nos expériences a donc été réalisée à deux niveaux. Nous nous sommes tout d'abord interrogés sur les capacités globales de régulation transcriptionnelle de la bactérie. Peut-on observer une trace de régulation dans nos données d'expression, et si oui peut-on la corréler à des éléments d'organisation du génome de la bactérie et/ou d'évolution moléculaire ? Cet espect est décrit dans les paragraphes 3.2 et 3.3 du chapitre 4. Nous avons ensuite essayé d'interpréter les résultats au niveau fonctionnel : peut-on interpréter la réponse transcriptionnelle de *Buchnera* vis-à-vis du contexte physiologique de l'expérience ? Ce travail est décrit dans le paragraphe 3.4 du chapitre 4

.

Chapitre 2

Développements technologiques : sélection d'oligonucléotides sondes, conception de la puce *Buchnera* et développement d'un système d'information

1. Introduction

L'avènement de la transcriptomique a fait émerger une classe de données entièrement nouvelle en biologie. Les progrès fulgurants de ces dix dernières années réalisés dans les domaines de la microinformatique et de la microfluidique associés au génie génétique (clonage de l'ADN, PCR et séquençage) ont permis un changement d'échelle dans la quantité de données acquises au cours d'une même expérience, ainsi qu'une miniaturisation de la plupart des appareillages. En effet, il est maintenant possible d'obtenir des estimations simultanées des niveaux d'expression de plusieurs milliers de gènes d'un organisme, d'un tissu ou même d'une seule cellule grâce à une expérience unique. La technologie des puces à ADN est issue du principe de « Southern blotting » (Southern *et al.* 1994). Le « reverse Southern » et les « macroarray » se sont ensuite popularisés grâce à la miniaturisation des procédés donnant naissance aux puces à ADN. Les applications des puces à ADN sont très variées. Le génotypage (hybridation ADN - ADN) est l'application qui intéresse le plus le monde industriel comme outil diagnostique de maladies et de pathogènes (Guschin *et al.* 1997 ; Lindroos *et al.* 2002 ; Huentelman *et al.* 2005) ou encore pour la taxonomie moléculaire (Edwards-Ingram *et al.* 2004).

L'analyse transcriptomique (analyse des ARN) est une technique de recherche exploratoire introduite maintenant dans tous les domaines de la biologie. Le principe de la technique est présenté sur la figure 4. Le nombre de gènes représentés sur la puce (sondes) varie de quelques centaines pour les puces de faible densité (généralement fabriquées dans les laboratoires) à plus de 100 000 gènes pour les puces à haute densité produites par la société Affimetrix®. De très nombreux articles concernant à la fois les aspects technologiques et des applications ont été publiés sur le sujet. On pourra se référer notamment au volume spécial de la revue Nature Genetics (vol 32, 2002), ou aux revues plus récentes de D'Ambrosio *et al.* (2005) ou de Peeters et Van der Spek (2005).

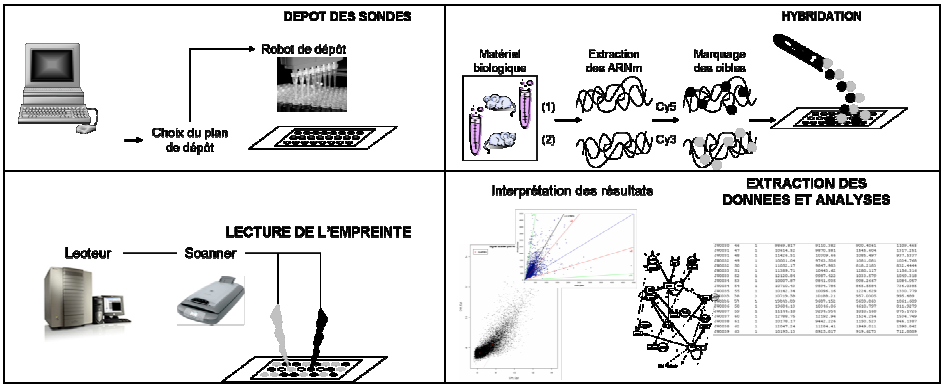

Figure 4 d'après Reymond *et al.* (2004b). Principe de la technique des puces à ADN. Des sondes (fragments d'ADN ou oligonucléotides) sont déposées (spots) sur une surface (verre, nylon, plastique). Ces séquences sont spécifiques des gènes dont on veut étudier l'expression. Les ARN messagers (ARNm) sont extraits des échantillons biologiques, marqués et déposés en solution sur la lame. Après hybridation, le signal est mesuré et donne une estimation de la quantité de chaque ARNm présent dans l'échantillon biologique. Les résultats sont ensuite analysés statistiquement.

Au laboratoire, nous avons développé une puce à ADN dédiée au génome de la bactérie *Buchnera*. Ce développement a été réalisé en parallèle avec la mise en place de la plateforme transcriptome de la génopole sur le site du DTAMB (UCB Lyon 1). Chaque étape du processus a dû être définie, testée puis validée. Lorsque cela a été nécessaire des développements informatiques ont été réalisés. Ce travail a été réalisé en collaboration avec Jacques Bernillon, ingénieur responsable de la plateforme, qui nous a apporté une aide précieuse.

2. Optimisation des sondes : le logiciel ROSO

Nancie Reymond durant sa thèse au laboratoire BF2I, a développé ROSO (Recherche et Optimisation de Sondes Oligonucléotidiques), un logiciel pour la sélection de sondes optimisées pour les puces à ADN. Le logiciel est publié (Reymond *et al.* 2004a) et accessible librement pour les chercheurs universitaires sur le site du PRABI (http://pbil.univ-lyon1.fr/roso). Le logiciel a été validé sur des jeux de données artificiels pour tester sa sensibilité et sa spécificité. Puis il a été testé sur des jeux de données réels. Le logiciel a été utilisé pour sélectionner les sondes de la puce *Buchnera* et de la puce *Ralstonia* développée par C. Boucher (INRA, Toulouse) (Occhialini *et al.* 2005). Entre mars 2003 et septembre 2005, 1448 projets ont été réalisés avec ROSO via le Web et 30 exécutables ont été distribués dans des laboratoires français et internationaux. Une licence a été vendue à un industriel. Je ne présenterai dans ce rapport que le principe général du logiciel, suivi d'une brève analyse critique.

Spécification du logiciel

ROSO est un logiciel pour le choix d'oligonucléotides sondes de 25 à 300 bases pour les puces à ADN. Les sondes optimisées par ROSO sont choisies dans des régions spécifiques de chaque gène de façon à limiter au maximum le problème de l'hybridation croisée. Les sondes sont sélectionnées par le logiciel de façon à homogénéiser les rendements d'hybridation du point de vue thermodynamique.

Chaque sonde possède 100 % de similitude avec sa cible propre. L'aspécificité d'une sonde est alors définie par son pourcentage de similitude maximale avec l'une des autres cibles du jeu de données. Comme les ARN messagers ne sont pas sélectionnés avant le marquage et l'hybridation, les sondes doivent être spécifiques vis-à-vis de tous les gènes de l'organisme d'étude, même si ceux-ci ne sont pas représentés sur la puce. ROSO permet à l'utilisateur de définir un jeu de séquences externe pour effectuer cette analyse d'aspécificité. L'aspécificité des sondes est déterminée en utilisant le logiciel Blast (Altschul *et al.* 1997) paramétré pour détecter des similitudes compatibles avec l'utilisation de la technique. Ce paramétrage est décrit sur le site de ROSO.

Les sondes optimisées par ROSO sont homogènes d'un point de vue thermodynamique. Elles possèdent des températures d'hybridation (Tm) homogènes. Le Tm (*melting temperature*) est la température à laquelle la moitié des molécules de la population de la sonde est fixée à sa cible. Le Tm est calculé en utilisant la méthode thermodynamique du plus proche voisin (SantaLucia 1998). Cette méthode permet de prendre en compte à la fois la nature des bases et leurs positions relatives. La formulation du modèle est donnée sur le site de ROSO. Comme l'hybridation est réalisée à température fixée sur toute la surface de la puce,

si les Tm sont homogènes, les rendements d'hybridation seront comparables. L'homogénéisation des Tm est réalisée par itérations successives dans l'algorithme de ROSO (figure 5). S'il le souhaite, l'utilisateur peut également fixer arbitrairement un intervalle de Tm pour le choix de ses sondes.

Les sondes sélectionnées sont également dépourvues de structures secondaires stables. Les structures secondaires sont des repliements complémentaires de la sonde sur elle-même (épingles à cheveux), ou entre deux sondes associées en quinconce (homoduplex). On suppose que ces structures pourraient géner l'hybridation, même si très peu de travaux sont venus pour le moment confirmer ou infirmer cette hypothèse. La stabilité des structures secondaires est relativement difficile à calculer, car elle doit prendre en compte l'influence des mésappariements. Le logiciel Mfold est le logiciel de référence pour ce type de calcul (Zuker 2003). Mfold est un logiciel libre, néanmoins nous ne l'avons pas utilisé dans ROSO, car il a été très difficile de récupérer le code du logiciel auprès des auteurs. Nous avons utilisé une approximation plus rapide et n'intégrant pas l'influence des mésappariements. Les résultats de ROSO ont été comparés à ceux de Mfold et montrent que notre logiciel surestime assez systématiquement la stabilité des structures secondaires, ce qui nous contraint à une sélection un peu trop drastique, mais nous assure une bonne qualité de sondes.

Enfin, ROSO permet de choisir des sondes dans des régions spécifiques de gènes (régions 3', 5' ou médiane). Il est basé sur un principe d'itérations avec des relaxations successives de paramètres. Un guide d'utilisation est fourni sur le site du logiciel.

Implémentation de ROSO

L'exécutable ROSO a été programmé en langage C, l'interface Web est en PHP. L'algorithme est présenté sur la figure 5.

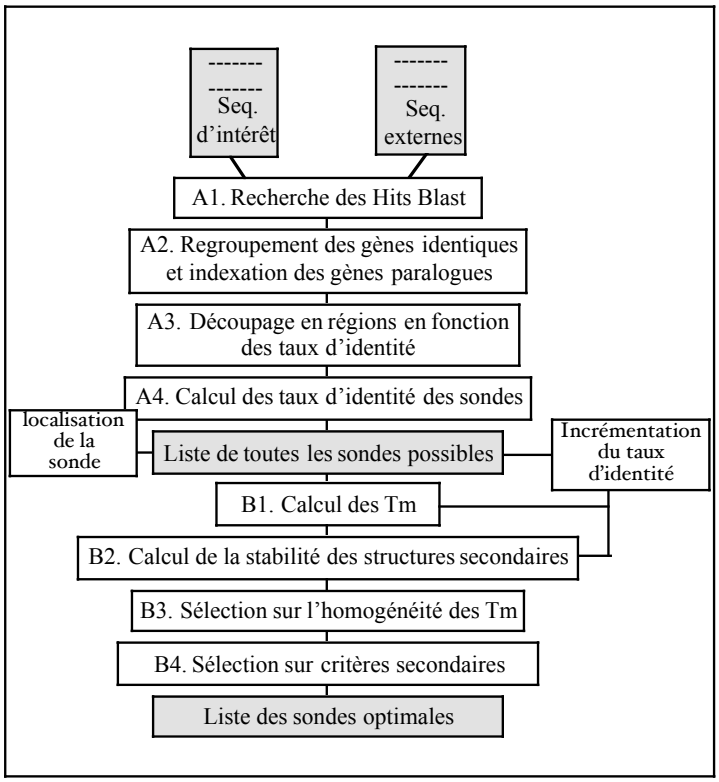

Figure 5. Algorithme ROSO. L'utilisateur fournit deux fichiers de séquences : sequences d'intérêt (séquences représentées sur la puce) et séquences externes (génome entier). Les zones de similarité « Hits Blast » sont recherchées (A1), les gènes identiques et paralogues sont indexés (A2), un coefficient de similarité est affecté à chaque base (A3), permettant de calculer l'aspécificité de chaque sonde (A4). La liste des sondes possibles, selon les critères de localisation sur le gène, est alors passée au crible de l'homogénéisation thermodynamique. Pour une spécificité donnée, les Tm (B1) et la stabilité des structures secondaires sont calculés (B2). Si aucune sonde n'est trouvée pour un gène, la recherche est élargie à des sondes ayant moins de spécificité et/ou localisées ailleurs. Dès qu'au moins une sonde est disponible pour chaque gène, le logiciel sélectionne un jeu de sonde de façon à définir le plus petit intervalle de Tm possible (B3). Le choix final du jeu de sonde est basé sur quatre critères secondaires (taux de GC, énergie libre des pentamères extrêmes…).

Analyse critique

De nombreux logiciels d'optimisation de sondes ont été développés, on pourra se référer à la thèse de N. Reymond (2004) pour une analyse comparative détaillée. Ces logiciels commencent tous en découpant les gènes d'intérêt en sondes de la longueur

souhaitée par l'utilisateur. Ces sondes sont ensuite alignées sur l'ensemble du jeu de gènes (Blast) l'une après l'autre, puis analysées du point de vue thermodynamique. Dès qu'une sonde correcte est obtenue pour un gène, la recherche est lancée sur le gène suivant. L'algorithme ROSO ne fonctionne pas sur ce schéma. Il sépare la recherche de similitude (Blast) de l'analyse thermodynamique. Ce découplage, qui nous a été suggéré par Laurent Duret (LBBE, UCBL), ne nous permet qu'une estimation moyenne de l'aspécificité des sondes, mais il possède deux avantages majeurs. Premièrement, l'étape de recherche de similarité, très coûteuse en temps, peut facilement être parallélisée en découpant le fichier d'intérêt en autant de sous fichiers que de processeurs disponibles. Deuxièmement, la recherche thermodynamique, moins coûteuse en temps de calcul, peut être réitérée pour calculer l'ensemble des sondes disponibles. Ce calcul permet alors de sélectionner les sondes, non pas sur une optimisation *a priori* de la valeur de Tm, mais sur un critère d'homogénéité : choisir l'ensemble des sondes montrant l'étendue la plus faible. Enfin, le découplage permet à l'utilisateur de participer au choix de la relaxation des critères de spécificité, de localisation, de Tm ou de stabilité des structures secondaires.

Le logiciel ROSO pourrait être amélioré en intégrant le code du logiciel Mfold pour augmenter les performances du calcul de la stabilité des structures secondaires. ROSO est optimal pour un choix de sondes de taille comprise en 25 et 70 paires de bases. Il reste fonctionnel pour des sondes longues de type fragment de PCR, mais ne propose pas de choix d'amorce PCR pour l'amplification des fragments choisis. Il n'est pas utilisable pour les sondes courtes (15 à 20 bases) de type « génotypage ». Le choix de ces sondes est, en effet, plus lié au contrôle de l'aspécificité qu'à l'homogénéisation des paramètres thermodynamiques. L'optimisation des sondes de génotypage doit intégrer des relations phylogénétiques et une recherche de signature taxonomique que ROSO n'intègre pas dans sa version actuelle. Pour le moment, ces perspectives de développement ne sont pas envisagées au laboratoire.

3. Conception de la puce *Buchnera*

Ce travail a été réalisé essentiellement sous la responsabilité de F. Calevro (MdC, INSA). Il ne sera donc que très brièvement résumé ici.

Une « mini puce » Buchnera

Le projet a débuté en 2000 et en collaboration avec J.-P. Cloarec (MdC, Ecole Centrale de Lyon, laboratoire IFOS). Une puce à ADN (100 spots) représentant quatre gènes de *Buchnera* (*pheA*, *eno*, *aroH* et *ilvH*) a été conçue sur des lames de type ROSATech et Corning. Grâce à cette « mini puce », un plan d'expérience a été élaboré de façon à analyser l'influence d'un certain nombre de facteurs : le choix du

support, le type de solution de dépôt et la concentration des sondes, la purification des ARN messagers de *Buchnera*, le type de marquage et le choix des amorces, les conditions d'hybridation et de lavage des lames, le choix des témoins positifs et négatifs, et enfin, l'estimation de la spécificité de l'hybridation. Ce travail a été publié (Calevro *et al.* 2004). Le tableau 1 résume les principaux choix technologiques retenus.

Type de support :	RosaTech
Solution de dépôt :	Tampon phosphate (150 mM), Bétaine (1,5 M)
Concentration de dépôt :	120 µM (20 ± 1 °C, 50 ± 3 % HR[1])
Préparation des ARN messagers :	Broyage, filtration sur membrane et extraction au TRIZOL®
Marquage des ARN messagers :	Incorporation indirecte de bases aminoallylés dans les ADNc par transcription reverse
Type de fluorochrome :	Cy3, Cy5
Choix des amorces pour le marquage :	Amorces semi-spécifiques (Talaat *et al.* 2000) et amorces aléatoires
Hybridation :	Chambre d'hybridation (42 °C, 16 h)
Lavage :	2 x 5 min (SSC 2x, 0,1 % SDS, 42 °C) 2 x 10 min (SSC 0,1x, 0,1 % SDS, TA[2]) 4 x 10 min (SSC 0,1x, TA) 1 x 10 s (eau distillée, TA)
Témoins négatifs :	Séquences aléatoires, organismes divers
Témoins positifs :	*pcp*, *pelK* (*B. subtilis*) ; *pelL* (*E. chrysanthemi*)
Spécificité :	5 mésappariements = absence de signal 3 mésappariements = signal très faible

Tableau 1. Paramètres de la procédure complète sélectionnés par le travail sur la mini puce *Buchnera* ([1] : Humidité Relative ; [2] : Température ambiante).

La puce complète Buchnera

Le passage à la puce complète *Buchnera* (6144 spots) a été réalisé au cours de la thèse de N. Reymond. Les oligonucléotides (35 bases, bras C_6 en 5') ont été choisis avec le logiciel ROSO et synthétisés par la société Eurogentec (Belgique). Les conditions expérimentales définies sur la minipuce ont été recalibrées, notamment pour le choix du support (lames QUANTIFOIL) et de l'hybridation automatique en station Ventana. Sur cette puce, deux à trois sondes ont été définies pour chaque gène (sauf pour les plus courts) et chaque sonde est représentée par un quadruplet de points (deux doublets déposés par deux aiguilles différentes). Des témoins négatifs et positifs ont été répartis sur toute la surface de la puce. Pour plus d'informations,

on pourra se reporter à la thèse de N. Reymond (2004). La figure 6 présente une reconstitution à partir des deux images rouge et verte d'une hybridation de l'expérience AAS (cf. Chapitre 4).

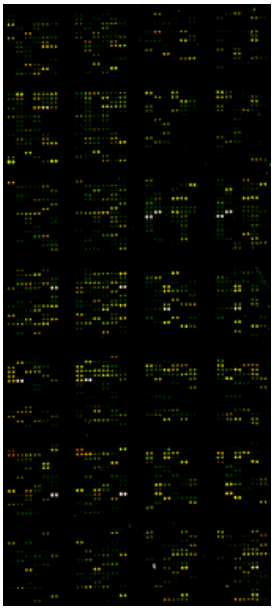

Figure 6. Hybridation en double couleur (superposition des images verte et rouge) de la puce complète *Buchnera* (6144 plots).

Analyse critique

Il faut tout d'abord noter que la mise au point d'une expérience de puce à ADN est toujours très délicate. En effet, à cause du prix de la technique (150 à 200 euros par lames) et du nombre de paramètres à tester, il n'est pas possible de tester toutes les combinaisons possibles de facteurs, ni leurs interactions. De plus, l'estimation même de la qualité d'une hybridation n'est pas aisée (estimation du signal moyen, estimation du bruit de fond, du rapport signal sur bruit, de l'analyse qualité…). L'approche est donc souvent empirique et consiste à tester les paramètres un à un pour ne conserver que le meilleur à chaque étape, ce qui revient à négliger dans la plupart des cas les interactions (Calevro *et al.* 2004).

Les oligonucléotides de la puce complète ont une longueur de 35 bases. Ce choix avait été fait essentiellement pour des raisons de prix et de fiabilité des synthèses de sondes de l'époque. Si nous devions refaire une commande à l'heure actuelle, je pense que nous opterions pour des oligonucléotides plus long (50 bases), conservant une bonne spécificité mais avec une sensibilité plus forte.

Les oligonucléotides ont été positionnés sur la lame dans l'ordre chromosomique (numéro d'accession BU). Ce choix avait été fait avec l'idée qu'un ordre logique sur la puce aurait permis un meilleur contrôle d'erreur par l'observation directe de la puce. Ce point était évidemment une illusion (aucune analyse visuelle n'est réalisable) et la prochaine version de la puce *Buchnera* devra présenter une répartition aléatoire des gènes sur sa surface.

Enfin, tous les gènes ne sont pas représentés sur la puce par le même nombre de répétitions, même si la majorité présente deux sondes (8 spots), certains n'en ont qu'une (4 spots), d'autres trois (12 spots). Comme les données sont systématiquement moyennées par gène pour ne pas mélanger variabilité technique et biologique dans les analyses statistiques, le déséquilibre de la précision sur les mesures crée un biais d'analyse. Il aurait donc été préférable d'équilibrer ce nombre de répétitions, ou d'utiliser des modèles hiérarchisés, intégrant la variabilité technique, pour gérer ce problème.

José Viñuelas (Thèse, BF2I) met actuellement en place une méthode de calibration du signal en utilisant une hybridation avec de l'ADN génomique. Chez *Buchnera*, il n'y a pas de répétitions de gènes, l'hybridation d'une lame avec de l'ADN génomique doit théoriquement donner un signal uniforme sur la lame. Les différences de fluorescence permettent alors de visualiser soit un mauvais dépôt, soit une mauvaise ou une trop forte affinité cible / sonde. Les premiers résultats obtenus jusqu'à présent montrent que l'ADN génomique peut facilement être utilisé pour détecter les aberrations de dépôt et les défections fortes de certaines sondes. L'utilisation de cette méthode pourrait donc au minimum servir de test de la valeur des sondes pour une future version de puce *Buchnera* ne possédant que des sondes montrant des réponses homogènes. Il paraît cependant plus difficile de corriger le signal obtenu pour obtenir une mesure calibrée du signal. Ce travail encore en cours, n'a pour le moment été publié que sous la forme d'un Poster (Vinuelas *et al.* 2005).

4. SITRANS : un Système d'Information pour la gestion, l'échange et la publication des données du TRANScriptome

Le projet de développement d'un système d'information pour l'analyse du transcriptome a été initié par un groupe d'étudiantes (C. Truntzer, A. Laugraud, C. Keime et C. knibbe) de la filière Bioinformatique et Modélisation de l'INSA de

Lyon en 2002. L. Duby (IUT informatique, Grenoble II) a poursuivi le travail au cours d'un stage de trois mois au laboratoire BF2I. Puis, avec l'aide de l'action bioinformatique des génopoles, un ingénieur CDD (D. Crisan) a été recruté pour une période d'un an, ce qui a permis une implémentation du modèle de la base de données. Le projet a ensuite été poursuivi par T. Chaari (Thèse LIRIS), et enfin A. Cros (Master BF2I/LIRIS) est en train de le finaliser. Ces étudiants ont tous été encadrés pour la partie informatique par F. Laforest (MdC, INSA) et A. Tchounikine (MdC, INSA) que je remercie très chaleureusement. La base a été publiée (Laforest *et al.* 2005), seul un résumé succinct suivi d'une analyse critique sera présenté ci-après. SITRANS est consultable à l'URL suivante : http://sitrans.insa-lyon.fr.

Spécification de la base de données

La base SITRANS est conçue pour l'archivage des données d'analyse du transcriptome. Cet archivage doit permettre la consultation, l'échange et la publication de données d'expression (Stoeckert *et al.* 2002). Le groupe international MGED (http://www.mged.org) a défini un standard de description d'une expérience de puce à ADN, le format MIAME (Minimal Information About Microarray Experiment). Ce standard est très peu formalisé et possède de nombreux champs descriptifs en texte libre. Il n'est donc pas très performant pour effectuer des requêtes de consultation. La base SITRANS est construite en suivant le processus expérimental de l'amont vers l'aval, elle respecte le formalisme MIAME, mais possède une description plus précise. Enfin, elle permet une gestion qualitative des ressources de la plateforme (date de réception, type de conditionnement, localisation des produits).

Le schéma de la base a été défini grâce à une modélisation UML permettant une description de l'ensemble des tables et de leurs relations. Ce schéma a été construit à partir du schéma de la base Mediante (http://www.microarray.fr:8080/mediante) développée initialement par H. Prieto et P. Barbry (Génopole Nice). Elle comprend un total de 44 tables. Elle est organisée en cinq grandes parties : (1) description de l'expérience ; (2) description et dépôt des sondes ; (3) description des échantillons et du marquage ; (4) description des hybridations ; et (5) description des images (figure 7). Dans chacune de ces étapes, l'utilisateur renseigne de très nombreux champs comme les protocoles utilisés, les dates des différentes étapes d'expérimentation, et les noms et localisation des tubes et plaques utilisés. Les fichiers de sorties des robots « rearrayer », « spotter », station d'hybridation et scanner sont téléchargeables. Le système assure donc une traçabilité complète de l'analyse. Les règles de cohérence de la base ont été établies, un certain nombre de champs sont obligatoires, d'autres sont facultatifs. Des champs en texte libre ont été conservés pour recevoir les commentaires généraux, mais la majorité des informations ont été structurées.

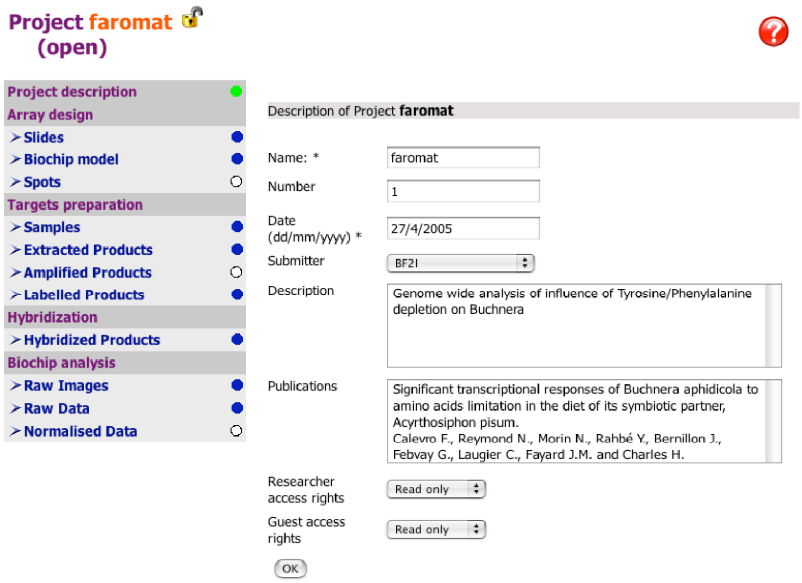

Figure 7. L'interface de saisie de SITRANS présente les différentes étapes du processus dans un menu déroulant situé sur la gauche de l'écran..

En plus des données saisies par l'utilisateur, il existe des « données fournies » définies par l'administrateur de la plateforme. Ces données ne peuvent pas être modifiées par l'utilisateur. Ce sont par exemple les lots de lames stockées, les noms des comptes utilisateurs et les protocoles des appareillages. De même, il existe une hiérarchisation des utilisateurs avec un administrateur (gestion des comptes et des données fournies), des utilisateurs (saisie et consultation) et des visiteurs (consultation uniquement). Enfin, un utilisateur peut interdire la consultation de ces expériences pour préserver des données confidentielles.

Implémentation de la base de données

SITRANS est une application Web. L'architecture du système est de type 4-tier, c'est-à-dire que les différentes parties de la base et les différentes taches sont gérées à quatre niveaux de service différents (figure 8). La couche la plus profonde est la base de données. Le système de gestion de base de données (SGBD) est PostgreSQL. Ce SGBD a été choisi pour sa fiabilité. La base de donnée dialogue avec le serveur d'applications (JBOSS) au moyen d'objets Entreprise JavaBeans de type Entity (EJB). Ces EJB sont des classes java de différents types. Par exemple, les « *entity* » correspondent à la formalisation objet de la base UML (les tables), alors que les « *sessions* » représentent les fonctionnalités de la base. Le serveur JBOSS reçoit des instructions du serveur http Apache. Ce dernier interprète les

requètes émises par le client Web, pour les transmettre à la couche inférieure. Enfin, la couche « présentation » (interprétée par le navigateur Web) permet à l'utilisateur d'interagir avec les fonctionnalités offertes par SITRANS à travers une interface graphique. Cette couche se base sur des Servlets et des JSPs qui sont des technologies utilisées pour développer des pages Web dynamiques. Les Servlets sont des classes Java appelées par des requêtes http, dont le rôle est de générer des pages HTML. Typiquement, les Servlets accèdent au système d'information (par l'intermédiaire des deux serveurs JBOSS et Apache) pour générer le contenu des pages HTML demandées. La technologie JSP est complémentaire aux Servlets. Une page JSP est une page HTML contenant du code Java délimité par des tags XML. Ce code est chargé de générer le contenu dynamique des pages HTML. Dans l'architecture de SITRANS, à chaque étape de description du projet correspond une Servlet. Chaque Servlet gère deux pages JSP : une pour la saisie et l'affichage des informations concernant l'étape courante et la deuxième présentant une synthèse de l'étape en cours.

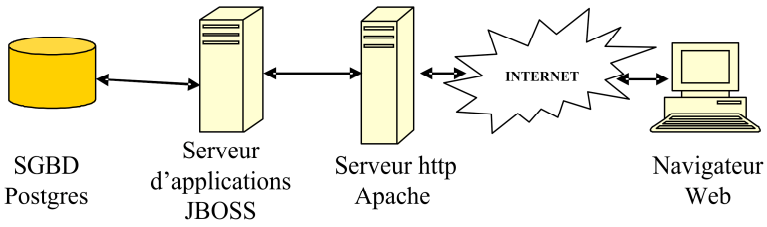

Figure 8. Architecture technique de SITRANS.

Analyse critique

La conception de la base SITRANS a été une expérience de travail interdisciplinaire très enrichissante. Même si le développement n'a pas été réalisé *de novo*, l'adaptation du modèle *Mediante* aux spécificités de la plateforme Lyonnaise a nécessité une interaction très forte entre les informaticiens et les biologistes. Les deux partenaires ont tiré des bénéfices de cette expérience.

Pour les biologistes, le système SITRANS représente un cahier de laboratoire électronique assurant un meilleur contrôle des biais techniques au sein d'un même projet, mais également entre les différents projets de la plateforme. La réflexion sur la modélisation de la base a permis une prise de conscience de la nécessité du développement d'une démarche qualité dans ce type d'expériences coûteuses et

complexes. A. Cros (Master, BF2I) développe actuellement un module permettant l'exportation au format MAGEML (format XML pour la description d'une expérience de transcriptome) de façon à publier automatiquement un projet dans une base de données internationale depuis SITRANS. Enfin, le système peut s'avérer intéressant pour l'aide à la conception de plans expérimentaux. D'une façon très simple, un utilisateur pourra rechercher un projet similaire dans la base de façon à s'en inspirer pour construire sa propre expérience. De façon plus élaborée, quatre étudiantes de la promotion BiM 2005 (V. Fasolo, F. Defay, V. Perie, N. Oudin), travaillent à la mise en place d'un système d'aide pour la construction de plans d'expérience, l'analyse qualité et la normalisation des données d'expression. Ce travail, réalisé dans le cadre de leur projet annuel, devra être raccroché au système SITRANS.

Pour les informaticiens, la conception d'un système d'information de la taille et de la complexité de SITRANS représente un intérêt en soi. Par ailleurs, la thèse de T. Chaari (LIRIS, INSA) est focalisée sur l'adaptabilité des applications au contexte d'utilisation dans les systèmes d'informations. Elle consiste à définir et développer des méthodes de représentations des données qui soient adaptées et adaptables pour chaque utilisateur d'un même système d'information. Par exemple, dans le cas de SITRANS, la consultation linéaire des projets (sur le même modèle que celui de la saisie) ne présente que très peu d'intérêt. Un prototype a été développé permettant d'accéder aux données de la base à partir de n'importe quel objet (par exemple un type de lame). Il est ensuite possible de naviguer dans la base en effectuant des requêtes liées à cet objet, les requêtes pouvant porter sur les liens associés à cet objet ou sur des contraintes de la liste visualisée :

- Quels sont les utilisateurs ayant utilisé ce type de lame ?
- Quels types de protocoles de marquage ont-ils utilisés ?
- Parmi ceux qui ont utilisé le type *Y*, a quelle date ont-ils réalisé les expériences et ont-ils fait référence à des problèmes spécifiques ?

Un système d'onglet permet de retrouver l'historique des requêtes et de se situer dans la base. Une représentation sous forme d'arborescence permet d'avancer ou de reculer dans les données par simples « clic ». Cet exemple de requête simple montre la pertinence d'un tel système permettant à chaque utilisateur de construire une visualisation personnalisée de la base pour répondre à des questions spécifiques.

5. Conclusions

L'analyse du transcriptome a connu un essor extraordinaire dans les années 2000 sous l'impulsion d'une politique gouvernementale encourageant fortement les analyses « post-génomiques » dans le secteur de la santé et de l'environnement. Le

financement des génopoles a ainsi permis l'installation de plateformes transcriptomiques et protéomiques dans de nombreuses villes françaises. Les budgets alloués ont été considérables. Comme trop souvent, ces financements ont été de très courte durée (deux ans) pour les sites, comme celui de Lyon, qui ne présentaient pas une « rentabilité suffisante ». Le travail de gestion remarquable et courageux des responsables de la plateforme transcriptome Lyonnaise a permis de la faire survivre jusqu'à maintenant, mais le futur de la plateforme Lyonnaise est menacé. Lorsqu'on réalise la maîtrise technique et les développements nécessaires pour produire des expériences de transcriptomiques fiables et reproductibles, on peut s'interroger sur le gaspillage engendré par cette politique de financement sur projets à court terme de la recherche publique.

Les réalisations présentées dans ce chapitre ont été pour nous une expérience d'ingénierie très enrichissante qui nous a permis de nous confronter aux problèmes de l'interdisciplinarité (essentiellement informatique et biologie). Le développement technologique est souvent dévalorisé en biologie, néanmoins la satisfaction d'avoir réalisé un logiciel « qui marche », ou une puce « qui s'allume » a été pour toute notre équipe un facteur de motivation bien plus important que n'importe quelle autre forme de reconnaissance.

Chapitre 3

Analyses statistiques et intégration des données d'expression

1. Introduction

La technologie des puces à ADN s'est développée au sein de la communauté des biologistes moléculaires jusqu'alors habitués à de l'analyse « gène à gène ». La gestion de données multiples, la prise en compte de différentes sources de variabilité correspond pour beaucoup à une redécouverte des méthodes statistiques jusqu'alors plutôt utilisées dans les domaines de l'écologie, de l'agronomie ou de l'épidémiologie (Nadon et Shoemaker 2002). Ce chapitre fait le point sur l'analyse de données d'expression issues de la technique des puces à ADN dans le contexte très restreint de l'analyse différentielle utilisant des puces à ADN à double marquage fluorescent. Il se place volontairement à l'interface entre l'expérimentateur et le statisticien proposant une « démarche » pour la conception d'une expérience, de l'analyse statistique, et de l'interprétation. Sa structure suit donc le déroulement logique de l'expérimentation : planification expérimentale, analyse qualité et filtration, normalisation, analyse différentielle et intégration des résultats dans la problématique biologique. Les figures et exemples proposés sont issus de nos expériences sur le modèle *Buchnera* (expériences YF et AAS décrites dans le chapitre 4). Plutôt que d'essayer de proposer une vision exhaustive de toutes les méthodes publiées jusqu'alors dans le domaine, j'essaie ici de détailler plus particulièrement les méthodes que nous avons appliquées sur notre modèle.

Le logiciel R (http://cran.univ-lyon1.fr/) de statistiques et de programmation et plus spécifiquement les bibliothèques de fonctions du projet Bioconductor ont été utilisées (http://www.bioconductor.org/). Ce chapitre est la base d'un didacticiel (projet initié par des étudiantes de la filière BiM de l'INSA) que nous mettrons à la disposition des expérimentateurs désireux de se lancer dans l'aventure transcriptomique (http://prabi.univ-lyon1.fr).

2. Planification expérimentale

Combien de lames dois-je prévoir pour mon expérience ? Combien de répétitions dois-je faire pour chacun des gènes sur la lame ? Dois-je faire des répétitions de mes conditions expérimentales ? Doit-on faire un « dye-swap[1] » systématiquement ? Comment organiser mes traitements et mes témoins par rapport aux hybridations ? Doit-on utiliser des témoins positifs et négatifs et comment ? Ces questions, souvent entendues lors des ateliers de formations que nous avons organisés pour la plateforme transcriptome de la Génopole (DTAMB), illustrent les principaux points à intégrer dans la planification d'une expérience : la puissance d'une analyse (nombre de répétitions), les confusions de facteurs (organisation des hybridations), la spécificité et la sensibilité (faux positifs et faux négatifs).

La corrélation observée entre les rapports de fluorescence issus de répétitions de spots sur la même puce est souvent très forte (supérieure à 95 %). Ceci est souvent interprété comme un signe de reproductibilité. Cependant, si l'échantillon de départ est divisé et hybridé sur différentes puces, dans différents laboratoires et sur des échantillonnages différents, la corrélation peut devenir extrêmement faible (Churchill 2002). La variabilité observée dans les données d'expression provient d'une part du processus technologique (le support, le marquage, l'hybridation, la lecture du signal et l'expérimentateur), c'est la variabilité technique, et d'autre part de l'échantillon, c'est la variabilité biologique. La planification expérimentale a pour objectif de permettre une estimation séparée de ces deux sources de variabilité, la première étant à éliminer et la deuxième à conserver pour réaliser les inférences correspondant aux hypothèses testées. Pour estimer ces deux sources de variabilités, il faudra prévoir des répétitions techniques (issues d'un même individu ou échantillon) et des répétitions biologiques (individus différents analysés dans les mêmes conditions techniques).

D'un point de vue statistique, la prise en compte de ces différentes sources de variabilité nécessite certaines précautions. En effet, il est impossible de comparer

[1] Le dye-swap consiste à répéter systématiquement chaque comparaison en inversant les marqueurs fluorescents de façon à éliminer la confusion de facteurs entre les effets « traitement » et « marquage ».

plus de deux échantillons biologiques sur une même lame (conditions techniques fixées), ou encore la répétition de spots sur une lame ne peut être analysée indépendamment d'un échantillon, d'un colorant ou d'une lame. Ainsi, il existe une hiérarchisation des facteurs dans les plans d'expérience de puces à ADN. De plus, certains facteurs sont fixes, c'est-à-dire qu'on souhaite limiter les conclusions aux seuls niveaux des facteurs testés dans l'expérience (par exemple, les deux fluorochromes ou les deux concentrations en drogue). D'autres facteurs sont aléatoires, c'est-à-dire qu'on a tiré au sort quelques modalités d'un facteur mais qu'on souhaite émettre des conclusions sur l'ensemble des niveaux possibles (la lame est un facteur aléatoire ainsi que l'individu). La prise en compte de ces différents effets nécessitera l'utilisation de modèles mixtes, souvent plus difficiles à mettre en œuvre que les modèles fixes. Le paragraphe 5.2.2 est consacré aux modèles mixtes.

Parmi les travaux consacrés à la détermination de plans expérimentaux, on peut citer Black et Doerge (2002), Pan (2002), Dobbin *et al.* (2003a) et Yang et Speed (2003) qui se sont intéressés à la taille des échantillons dans différents types d'approches comparatives. Les travaux de Dudoit *et al.* (2002 ; 2003a ; 2003b), Ge *et al.* (2003) et van der Laan *et al.* (2003a et b), sont plutôt orientés vers les procédures de contrôle de risque dans les tests multiples. Les chercheurs du groupe de Kerr *et al.* (2000 ; 2001a ; 2001b ; 2002) et Wu *et al.* (2003) ont été les premiers à proposer une stratégie de planification expérimentale en promouvant le concept de plan en boucle s'opposant au plan en référence. Les travaux de Dobbin et Simon ont ensuite comparé différents plans d'expériences en termes de coûts d'échantillons biologiques et/ou techniques, de puissance de détection ou de classification (Dobbin *et al.* 2003a et b ; 2005). Enfin, on peut noter les travaux de Tempelman qui propose une vision très pédagogique et très détaillée de ces problèmes de planification expérimentale utilisant la technologie des puces à ADN avec des modèles mixtes (Tempelman 2005).

2.1. Puissance d'une expérience et nombre de répétitions

Les analyses différentielles produisent généralement une liste de probabilités associée à la liste des gènes présents sur la puce à ADN. Ces probabilités (pvalues) correspondent à la probabilité de réalisation de la statistique du test (différence d'expression entre deux ou plusieurs conditions expérimentales) lorsque l'hypothèse nulle est vraie (H0 : les conditions expérimentales ne diffèrent pas). Ainsi, lorsqu'un gène *g* est associé à une probabilité p très faible, nous sommes en mesure de rejeter l'hypothèse nulle en prenant le risque p de se tromper. Le gène est alors détecté comme gène différentiel. Le choix de la valeur seuil de p pour décider du rejet de H0 est classiquement de 5 %. Cette valeur est appelée risque α ou de première

espèce. Nous verrons plus loin, que dans le cas des puces à ADN qui posent le problème des tests multiples, cette valeur limite mérite d'être réévaluée.

On peut maintenant se poser la question suivante : « sachant qu'il existe une différence Δ entre les deux conditions expérimentales, quelle est la probabilité de détecter cette différence avec l'analyse ? » Cette probabilité correspond à la puissance du test. Nous voyons ici que la puissance d'un test ne peut se définir que par rapport à une valeur fixée de H1. La différence Δ s'exprime différemment selon la statistique utilisée (comparaison de deux ou de plusieurs conditions expérimentales, modèle fixe ou modèle mixte), mais dans tous les cas, on montre que la puissance du test est une fonction de σ (la variance résiduelle) et de n (le nombre de répétitions). Nous verrons enfin que dans le cas de la comparaison de plusieurs conditions expérimentales, l'optimisation du plan d'expérience (l'agencement des différentes hybridations) peut permettre d'augmenter la puissance de l'analyse. Bien souvent, les facteurs déterminant le nombre de répétitions dans un plan d'expérience de puce à ADN ne sont pas liés à l'estimation de la puissance du test mais bien à l'argent nécessaire pour la mise en place de l'expérience. Le calcul de puissance permettra alors de déterminer l'écart minimal détectable par l'expérience pour un nombre fixé *a priori* de répétitions (Pavlidis *et al.* 2003). Dobbin *et al*. (2005) proposent un moyen d'estimer le nombre de lames (n) et de répétitions (m) nécessaires pour assurer une puissance donnée, tout en contrôlant le nombre de faux positifs. Les formules nécessitent d'avoir une connaissance *a priori* sur les ordres de grandeurs des différentes sources de variabilité.

2.1.1. Comparaison de deux conditions expérimentales

Nous ne traitons ici que des lames à double fluorescence (généralement rouge et verte). Nous sommes dans un cas de données appariées car les deux mesures associées à une lame sont faites avec la majorité des conditions techniques fixées (dépôt des sondes, hybridation et lecture au scanner). Le test réalisé est un test de Student (cf. § 5.1) car le facteur aléatoire « individu », confondu dans notre expérience avec le facteur « lame » n'est pas pris en compte. Soient $\bar{y}_{g1}$ et $\bar{y}_{g2}$ les moyennes de fluorescences observées dans l'expérience pour le gène g, le test de Student peut conduire à accepter à tort H0 alors que H1 est vraie (H1 : $\mu_{g2} = \mu_{g1} + \Delta$). Le risque de deuxième espèce β correspond à la probabilité de ne pas détecter la différence réelle Δ. Ce risque est donné par la formule ci-dessous et permet de calculer la puissance du test (1-β) :

$$\beta = prob(t_{\alpha/2} < \frac{\bar{y}_{g1} - \bar{y}_{g2} + \Delta}{\sigma/\sqrt{n}} < t_{1-\alpha/2})$$

La puissance est donc directement proportionnelle au nombre de répétitions n et à la variabilité σ_g. Pour obtenir une bonne puissance, il suffit donc d'augmenter le nombre de lames de l'analyse. La puissance du test de Student est généralement tabulée en fonction de Δ/σ dans tous les bons livres de statistiques. En pratique, on prendra le problème à l'envers en calculant la différence minimale détectée pour un nombre fixé de répétitions.

A titre d'exemple, l'expérience YF avait été initialement conçue avec huit lames organisées en « dye swap » (tableau 2). Après l'étape d'analyse qualité, nous avons éliminé la moitié des lames du jeu de données. Avec quatre répétitions au final, on détectera une différence de 3σ avec une puissance de 90 %. La variabilité étant forte dans cette expérience, la puissance de notre expérience est globalement assez faible. Ce résultat est repris dans la partie analyse de la spécificité (§ 2.3).

Milieu	Lame	Marquage
YF	Q1	Cy5
AP_3	Q1	Cy3
YF	Q3	Cy5
AP_3	Q3	Cy3
AP_3	Q4	Cy5
YF	Q4	Cy3
YF	V45	Cy5
AP_3	V45	Cy3

Tableau 2 Plan expérimental de l'expérience YF organisé en « dye swap » et comportant quatre répétitions de chaque condition expérimentale (YF et AP_3). Chaque lame est hybridée avec un échantillon biologique différent.

2.1.2. Comparaison de plus de deux conditions expérimentales

Lorsque plus de deux conditions expérimentales sont comparées, l'analyse différentielle passe par l'utilisation d'un modèle d'analyse de la variance (§ 5.2). Seul le cas du modèle fixe sera abordé ici. La différence Δ_g s'exprime alors comme le rapport entre les variances inter-conditions et intra-condition pour le gène g. La matrice du plan d'expérience X, décrivant l'agencement des hybridations sur les lames en fonction des fluorochromes et des conditions expérimentales, permet le calcul des variances et des covariances des estimateurs (cf. Eq. A.2 en annexe). Ainsi, il est possible d'optimiser le plan d'expérience par un agencement particulier des hybridations, de façon à minimiser la variance des coefficients d'intérêt du modèle pour augmenter la puissance des tests correspondants (Churchill 2002 ; Speed et Yang 2002). Le déterminant de la matrice X'X est une estimation globale de la variance des estimateurs (critère de D-optimalité), il est d'autant plus grand

que les variances sont faibles. Mais l'optimisation d'un plan d'expérience ne peut pas se limiter à la D-optimalité. D'une part, les effets considérés par l'analyse ne sont pas d'égale importance et il sera par exemple plus intéressant d'avoir un estimateur de l'effet traitement le plus précis possible. D'autre part, l'optimalité peut se poser en termes de robustesse vis-à-vis de l'élimination de lames ou de spots par l'analyse qualité par exemple. Yang et Speed (2002) proposent une analyse détaillée de la précision des différents estimateurs dans des plans très variés. Tempelman (2005) offre également une très bonne revue de ces problèmes incluant l'utilisation de modèles mixtes.

Indépendamment du contexte de l'analyse du transcriptome, on montre qu'un équilibrage des répétitions sur les différentes conditions permet d'optimiser la puissance de comparaisons croisées. A l'inverse, pour comparer *r* conditions à une référence commune, il est préférable d'avoir un nombre de répétitions constant (n_i) pour les différentes conditions et de répéter la référence n_0 fois ($n_0 = n_i\sqrt{r-1}$) afin d'obtenir une puissance maximale (Dagnelie 1980). Dans le cas des données d'expression, ces deux types de comparaisons correspondent aux deux plans les plus classiquement utilisés.

Les plans en référence

Dans ce type de plan, toutes les hybridations sont réalisées en utilisant la même condition comme référence (figure 9). Il faut donc T lames pour comparer T conditions (C_1 à C_T) à une même référence C_0. Habituellement, la référence est systématiquement hybridée avec la même couleur. De cette façon, un marquage différentiel d'un gène ne sera pas confondu avec une différence entre les différents traitements. Dobbin *et al.* (2003a ; 2003b) montrent que lorsque le but du plan est de comparer les différentes conditions à la référence (et non pas les conditions entre elles), il est alors possible d'éliminer l'effet de marquage différentiel en rajoutant un nombre minimal de marquages inverses. Ce cas s'applique par exemple à la comparaison de tissus tumoraux à un tissu sain de référence. Ce point de vue est critiquable dans la mesure où il néglige l'aspect normalisant du « dye swap » complet. Néanmoins il est très pragmatique et à nombre de lames fixées, il sera sans doute préférable de perdre un peu de puissance de détection, ou d'inclure quelques biais, au profit d'un échantillonnage biologique plus large (dans ce cas, les diverses tumeurs).

Dans les plans en référence, les comparaisons ($\Delta_{iO} = C_i - C_0$) sont optimisées, et les comparaisons ($\Delta_{ij} = C_i - C_j$) sont sous optimales. En effet, si la variance associée Δ_{iO} vaut σ^2, alors la variance associée à $\Delta_{ij} = \Delta_{i0} - \Delta_{j0}$ vaut $2\sigma^2$.

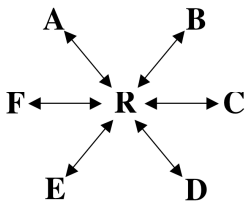

Figure 9. Plan en référence et en « dye swap » à six conditions (12 lames). Les six lames sont systématiquement comparées à la même référence. Les doubles flèches indiquent que chaque hybridation est répétée sur une autre lame en inversant les fluorochromes « dye swap ».

Ce plan a l'inconvénient de mesurer systématiquement l'expression d'une référence qui n'a souvent que peu d'intérêt pour le biologiste et les coûts de marquage sont donc très élevés. Néanmoins, en biologie humaine par exemple, l'échantillon biologique est souvent plus coûteux que la puce. Le plan en référence se révèle alors beaucoup plus robuste, car il permet d'introduire ou d'éliminer une lame de l'analyse sans interférer sur les autres comparaisons. Les plans en boucle, présentés dans le paragraphe suivant, ne tolèrent pas ces modifications.

Les plans en boucle

Les plans en boucle se caractérisent par une association équilibrée des différentes conditions sur chaque lame : A est comparé à B sur une lame, puis B à C sur une autre lame, et ainsi de suite jusqu'à la dernière condition qui sera comparée à A sur la dernière lame (figure 10). Toutes les conditions sont présentes deux fois dans le plan (une fois en rouge et une fois en vert). Comme pour les plans en référence, la comparaison de T conditions nécessite T lames. Ce plan peut être construit en « dye swap » en doublant le nombre de lames. Ici également le « dye swap » peut être complet ou incomplet.

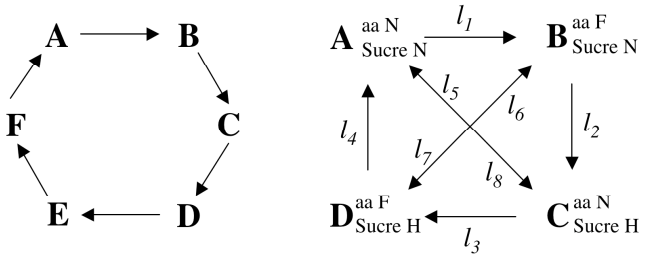

Figure 10. Deux types de plans en boucle. A gauche, un plan en boucle classique avec six conditions et six lames ; à droite, le plan de l'expérience AAS (quatre conditions, huit lames), incluant des croisements supplémentaires entre les quatre conditions qui représentent deux facteurs de l'expérience (acides aminés et sucre) à deux niveaux (N - F et N - H).

Les plans en boucle sont très intéressants pour faire des comparaisons directes entre les conditions. Néanmoins, celles-ci deviennent sous optimales pour des conditions éloignées. Soit *T*, le nombre de conditions du plan et *d,* la distance entre les deux conditions, *i* et *j,* d'une boucle classique. Martin-Magnette et Robin (2004) montrent, en tenant compte des deux chemins pour aller d'une lame à l'autre dans la boucle, que la variance entre deux conditions vaut $\sigma^2 d(T-d)/nT$. Elle est maximale pour la comparaison de deux lames opposées et vaut $\sigma^2 T/4n$. Elle est minimale pour deux conditions adjacentes et vaut $(T-1)\sigma^2/nT$. Il est donc essentiel de faire figurer côte à côte sur le plan (de les croiser sur la même lame) les conditions pour lesquelles les comparaisons sont les plus importantes. Il est également possible d'introduire des comparaisons croisées dans la boucle, comme c'est le cas du plan de l'expérience AAS (figure 10, à droite).

Le plan AAS est équilibré : toutes les conditions apparaissent quatre fois, deux fois rouges et deux fois vertes. Cependant, les quatre conditions représentent deux facteurs, acides aminés essentiels (*aa*) et sucre (*Sucre*) à deux niveaux chacun : normal (*N*) et faible (*F*) pour les acides aminés, et normal (*N*) et hyperosmotique (*H*) pour le sucre. Le plan n'est pas équilibré pour ce qui concerne les facteurs pris séparément. Comme le plan est répété dans l'expérience, huit lames comparent directement les conditions « *sucreN* » et « *sucreH* », et quatre lames comparent les conditions « *aaN* » et « *aaF* ». Le plan est donc plus puissant pour détecter les gènes différentiels induits par le stress osmotique, que par la déplétion en acide aminés. Ce résultat est directement visible sur la matrice des variances covariances $\left(X'X\right)^{-1}$ du plan de l'expérience fourni en annexe (tableau A.2). Les variances des estimateurs des effets A, S et AS sont les suivantes :

$$\sigma_A^2 = 0{,}31\sigma^2,\ \sigma_S^2 = 0{,}29\sigma^2 \ \text{ et } \ \sigma_{AS}^2 = 0{,}17\sigma^2$$

On remarque également des covariances non nulles pour tous les estimateurs (sauf pour le fluorochrome). Une étude de l'optimalité du plan AAS n'a pas été réalisée. Nous avons simplement vérifié *a posteriori* qu'un plan à 12 lames (correspondant au « dye swap » complet) permet d'obtenir des variances équilibrées des trois estimateurs ($\sigma_A^2 = \sigma_S^2 = \sigma_{AS}^2 = 0,32\sigma^2$). Le plan a 24 lames (répétition du plan à 12) permet d'obtenir des variances $\sigma_A^2 = \sigma_S^2 = 0,28\sigma^2$ et $\sigma_{AS}^2 = 0,17\sigma^2$. Nous avons préféré réaliser deux séries parallèles à huit lames, plutôt qu'un seul plan à 16 lames pour des raisons de coût. En effet, un dysfonctionnement d'une étape comme le marquage, l'hybridation ou le lavage des lames aurait pu anéantir l'intégralité de notre expérience sans possibilité de la renouveler. Le découpage en deux séries de lames permettait de se retrancher sur un seul plan à huit lames en cas d'échec d'une série. Il est notable que dans notre étude les « dye swap » ont été réalisés sur les mêmes échantillons biologiques, alors que les différentes conditions et les répétitions du plan ont été faites sur des échantillons différents. Ce facteur aléatoire « échantillon » aurait nécessité la mise en place d'un modèle mixte, et n'a pour le moment pas été introduit. De même, l'effet série de lames n'est pas intégré dans notre modèle.

2.2. Plans incomplets, confusions de facteurs et saturation des modèles (organisation des hybridations)

Quand seulement deux conditions expérimentales sont comparées, on peut faire une analyse complète en incluant un « dye swap ». Mais dès qu'il y a plus de deux comparaisons, le plan incomplet est incontournable. L'équilibrage du plan devient alors véritablement crucial.

La confusion de facteurs est un problème récurrent en analyse du transcriptome. Si l'on compare deux conditions expérimentales avec n répétitions par condition (type YF), et si chaque condition est systématiquement associée à la même couleur, il sera alors impossible d'interpréter une différence de moyenne significative pour un gène comme l'effet d'une expression différentielle entre les conditions de celui d'un marquage différentiel des fluorochromes. On dit alors que les effets « condition » et « marquage » sont confondus. Une façon très simple d'éviter ce problème consiste à répartir les différentes conditions de façon équilibrée sur les deux colorants. Ce type de plan est alors appelé « dye swap » ou « flip flop ». Ce dispositif doit être appliqué également lorsque l'on compare plus de deux conditions pour les plans en référence ou en boucle. Si les « dye swap » sont réalisés systématiquement sur les mêmes échantillons biologiques alors le modèle peut servir pour normaliser les effets d'un marquage différentiel des deux fluorochromes. Néanmoins, il peut devenir très coûteux car il multiplie par deux le nombre de lames de l'expérience.

Classiquement, les modèles d'analyse de la variance utilisés pour les données de puces à ADN prennent en compte les effets lame (L), fluorochrome (D), gènes (G) et conditions expérimentales (A et S dans l'expérience AAS) (cf. § 5.2.1 pour la construction des modèles). Les interactions entre ces différents facteurs devraient en toute rigueur également être considérées. A titre d'exemple, dans l'analyse AAS, on devrait théoriquement estimer les effets de 10 interactions doubles, huit interactions triples, cinq interactions quadruples et une interaction sextuple. Le modèle complet possède ainsi 31 termes et s'écrit de la façon suivante :

$$\begin{aligned} Y_{ijgas} &= \mu + L_i + D_j + G_g + A_a + S_s + LD_{ij} + LG_{ig} + LA_{ia} + LS_{is} + DG_{jg} \\ &+ DA_{ja} + DS_{js} + GA_{ga} + GS_{gs} + AS_{as} + LDG_{ijg} + LDA_{ija} + LDS_{ijs} \\ &+ LGA_{iga} + LGS_{igs} + DGA_{jga} + DGS_{jgs} + GAS_{gas} + LDGA_{ijga} + LDGS_{ijgs} \\ &+ DGAS_{jgas} + LGAS_{igas} + DGAS_{jgas} + LDGAS_{ijgas} + \varepsilon_{ijgas} \end{aligned}$$

Les facteurs L, D, G, A et S possèdent respectivement i = 16, j = 2, *g* = 617, a = 2 et s = 2 niveaux. Un tel modèle demande l'estimation de plus de 90 000 coefficients, alors que le nombre total d'observations est de 19 744 : le modèle est donc très largement saturé. Ainsi, à partir d'un certain nombre de facteurs, l'introduction d'un nouveau facteur correspond à une combinaison linéaire des facteurs déjà introduits. Par exemple, l'interaction LD est confondue avec les effets A+S, puisqu'à chaque combinaison de L et D correspond une combinaison de A et S. Kerr *et al.* (2000) proposent une analyse détaillée des confusions dans un plan d'expérience simple en carré latin. L'ajustement du modèle complet va demander des simplifications. Les interactions d'ordre supérieur à deux (sauf GAS qui est d'intérêt) sont tout d'abord éliminées. En considérant la normalisation des données et l'équilibrage des fluorochromes sur les conditions nous éliminons ensuite les interactions LD, LG, LA, LS, DG, DA et DS. Le modèle simplifié s'écrit alors :

$$Y_{ijgas} = \mu + L_i + D_j + G_g + A_a + S_s + GA_{ga} + GS_{gs} + GAS_{gas} + \varepsilon_{ijgas}$$

La simplification du modèle conduit à une bonne estimation de la variance résiduelle qui permet de tester les facteurs d'intérêt. Nous verrons plus loin les problèmes posés par cette étape. La variance résiduelle contient beaucoup des termes supprimés dans la simplification, mais il est clair que les termes conservés sont eux aussi confondus avec certains des termes d'interactions éliminés.

2.3. Spécificité et sensibilité (faux positifs et faux négatifs)

Dans les expériences de puces à ADN le nombre de gènes analysés est grand, entre plusieurs centaines et plusieurs milliers. Nous avons vu dans le paragraphe précédent comment interpréter les probabilités associées à chaque gène par l'analyse différentielle, en se fixant un seuil α pour remplir la liste des gènes différentiels. Réaliser le test sur un seul ou sur 1 000 gènes ne pose pas de problème spécifique. Néanmoins, on peut se poser deux types de questions à l'issue de la sélection :

- Combien de fois a-t-on choisi un gène à tort ?
- Quelle est la probabilité pour que figure sur la liste au moins un gène invariant ?

Ces deux questions illustrent ce que l'on appelle le contrôle du risque multiple. La première est appelée FWER (Family Wise Error Rate) et définit la probabilité de détecter au moins un faux positif dans la liste des gènes sélectionnés. La deuxième est appelée FDR (False Discovery Rate), elle indique la proportion de faux positifs dans la liste de gènes sélectionnés. Il est clair que si l'on n'a pas pris de précautions particulières sur le choix du seuil α, la réponse à ces deux questions est : « Beaucoup ! » A l'inverse, le contrôle de α ne pourra pas être fait sans altérer la puissance du test.

Pour répondre à la première question, Bonféronni a proposé la seule bonne façon de régler le problème en suggérant de diviser le seuil de risque par le nombre de tests que l'on souhaite réaliser (Dagnelie 1980). Le FWER devient alors α. Mais cette méthode est beaucoup trop conservative et n'a pratiquement aucun intérêt dans le cas de l'analyse du transcriptome. En effet, prendre un risque de $5\ 10^{-5}$ pour sélectionner un gène afin d'assurer un risque global de 5 % pour un jeu de 1 000 gènes revient généralement à ne sélectionner que quelques très rares gènes extrêmement différentiels pour lesquels bien souvent l'expérimentateur connaissait déjà la réponse. Le groupe de S. Dudoit (Université de Californie) propose de nombreux articles sur le contrôle du risque (Dudoit *et al.* 2002 ; 2003a et b). Toutes ces méthodes sont dérivées de la méthode de Bonféronni, et visent à recalculer une valeur de probabilité pour chaque gène sachant que l'on souhaite contrôler le risque global d'erreur. La méthode FWER la plus intéressante est sans doute celle utilisant la méthode de Westfall et Young (1995, citée par Dudoit (2002)) qui permet la prise en compte de la structure de dépendance entre les différents tests (les corégulations de gènes). Cette méthode est basée sur le rééchantillonnage des conditions expérimentales et nécessite un nombre important de répétitions. Ces méthodes sont très utiles par exemple pour la recherche de gènes marqueurs dans le diagnostic d'une maladie, car un très petit nombre de gènes est recherché et il est important que ces gènes ne soient pas de faux positifs. Par contre, le FWER n'a pratiquement aucun intérêt dans une approche systémique, où une vision globale du système est

recherchée. Ce point sera plus largement discuté dans le paragraphe correspondant à l'intégration des résultats statistiques dans le contexte biologique (§ 6).

Une autre façon d'aborder le problème consiste à essayer de répondre à la seconde question pour contrôler le nombre de faux positifs dans la liste de gènes différentiels. Benjamini et Hochberg (1995) ont proposé une procédure de contrôle du FDR sous l'hypothèse d'indépendance des différents tests réalisés pour le calcul des probabilités p_i associées à chacun des N gènes. Selon la procédure, les probabilités (p_i) sont tout d'abord classées par ordre croissant : $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(N)}$. Puis on recherche la plus grande probabilité $p_{(j\max)}$ telle que : $j_{\max} = \max\left\{ j : p_j \leq \frac{j}{N}\alpha \right\}$. Les gènes sélectionnés correspondent alors à tous les gènes dont la probabilité est inférieure à $p_{(j\max)}$. On voit très simplement que si la dernière probabilité $p_{(N)}$ (la plus grande) vérifie encore l'inéquation, c'est-à-dire si elle est inférieure ou égale à α, alors le taux de faux positifs sur les N tests est égal à $\sum_1^N (1/N)\alpha = \alpha$.

La procédure a été reprise ensuite par Storey et Tibshirani (2003) qui définissent la notion de q-value. La q-value d'un gène *g* (ayant une p-value p) est la proportion de faux positifs dans l'échantillon si on considère significatifs tous les gènes associés à une p-value inférieure à p_g. La q-value permet ainsi de garantir un taux de faux positifs non plus global et *a priori* comme dans la procédure initiale de Benjamini et Hochberg, mais bien parmi les gènes sélectionnés, et *a posteriori*. La procédure reste cependant assez conservative en pratique, puisque le nombre de gènes différentiels d'une analyse de transcriptome est toujours très important. Reiner *et al.* (2003) en utilisant des méthodes de simulation, puis Aubert *et al.* (2004) en utilisant des méthodes d'approximation, ont proposé un contrôle de FDR local. Ces méthodes ne considèrent plus l'ensemble des tests de probabilités inférieures à un seuil, mais seulement les tests réalisés au voisinage d'une certaine probabilité ce qui permet alors d'estimer pour chaque gène la probabilité d'être un faux positif.

Sur nos données d'expression, nous avons développé une approche de type FDR par simulation de données en utilisant un jeu de lames témoins. Pour cela, les données normalisées de l'expérience AAS (16 lames, soit 32 colonnes de fluorescence) ont tout d'abord été centrées par colonne autour de la moyenne générale (normalisation inter lame). Les variances n'ont pas été calibrées entre les différentes lames car l'analyse des distributions montre peu de variation (figure 11, ligne du haut).

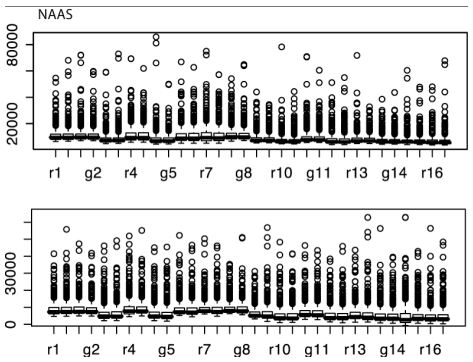

Figure 11. Distribution des 32 colonnes de fluorescence du jeu de données AAS. En haut : lames réelles ; en bas : lames simulées.

Les huit colonnes correspondant aux quatre lames témoins du jeu de données ont été récupérées, et pour chaque gène, la variance et l'écart type ont été calculés. Vingt-quatre colonnes ont alors été générées, gène par gène, par tirage aléatoire dans une loi normale : $N(\mu_g,\sigma_g)$. Cette simulation fait l'hypothèse de variances intra lames homogènes et d'un effet fluorochrome nul. Cette dernière hypothèse est peu contraignante puisque les données sont normalisées. Par contre, la simulation intègre l'hétérogénéité des variances des gènes. Les 12 lames simulées ont ensuite été ajoutées aux quatre lames témoins pour former un jeu de 16 lames, ce jeu a été permuté aléatoirement par colonne de façon à mélanger lames réelles et simulées. Puis les lames ont été décentrées selon le modèle des données réelles. L'analyse a alors été lancée sur le jeu de données simulées. Cette analyse se déroule donc avec l'hypothèse nulle vraie : « il n'y a pas de gènes différentiels. » Le calcul de la probabilité et la sélection sont réalisés exactement comme sur des données réelles, c'est-à-dire que les gènes différentiels doivent montrer une probabilité inférieure à 0,05 et un rapport d'expression extérieur à l'intervalle 0,83 - 1,2. Un nombre de cinq faux positifs est obtenu sur le jeu de données AAS. La liste réelle contenant 92 gènes (cf. tableaux A.5 et A.6 en annexe), on peut donc considérer un taux d'environ 5 % de faux positifs dans notre analyse. Pour étudier l'évolution du FDR en fonction des paramètres de sélection, les courbes ont été tracées pour différentes valeurs de α et différentes valeurs de seuils de rapports d'expression (figure 12). Ces courbes montrent que le niveau de signification du test a peu d'effet sur le nombre de faux positifs (peu de différence entre α = 0,1 et α =0,05), contrairement au seuil du rapport d'expression qui est un paramètre très sensible. Il est notable que nous avions déterminé le rapport de 1,2 de manière complètement empirique. Notre choix avait été guidé par la cohérence de l'analyse fonctionnelle réalisée : les gènes significatifs et au-dessus du seuil semblaient « avoir du sens ». L'analyse théorique réalisée ici confirme la pertinence de ce choix.

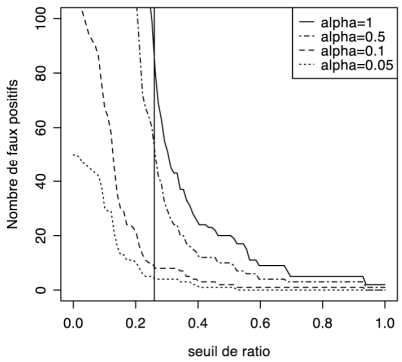

Figure 12. Nombres de faux positifs de l'expérience AAS pour différentes valeurs de α et différentes valeurs de seuil de rapports d'expression (exprimé en valeur absolue du $\log_2$).

Une procédure similaire a été réalisée sur le jeu de données YF, et un taux de 25 % de faux positifs a été déterminé (13/53, figure 13), ce qui confirme la moins bonne qualité des lames YF et la puissance moindre du plan d'expérience évoqué précédemment. Pour diminuer cette valeur, nous pourrions envisager d'augmenter le seuil du rapport, ou de diminuer le risque individuel pour chaque test. Néanmoins l'analyse fonctionnelle montre que le risque est grand d'augmenter parallèlement le nombre de faux négatifs, c'est-à-dire de perdre beaucoup de gènes intéressants pour la discussion.

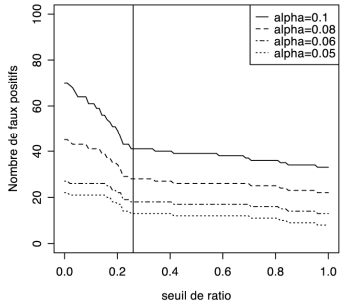

Figure 13. Nombres de faux positifs de l'expérience YF pour différentes valeurs de α et différentes valeurs de seuil de rapport d'expression (exprimé en valeur absolue du $\log_2$).

Enfin, l'estimation du nombre de faux négatifs pourrait être envisagée par simulation. Il faudrait introduire dans les lames simulées sous H0, un nombre fixé de gènes activés dans des rapports bien déterminés (par exemple de 2 à 1.1). On pourrait alors déterminer la sensibilité et la spécificité de la méthode. Néanmoins ce travail n'a pas été réalisé. Il me semble que les hypothèses introduites par une telle simulation seraient trop incertaines puisqu'on ne connaît pas la distribution de l'expression des gènes différentiels, ni la structure (corrélation) de la population de gènes différentiels.

Le calcul du FDR par simulation présente plusieurs avantages, il permet à la fois d'incorporer des hypothèses sur les différents tests (hétéroscédasticité par exemple), mais également d'incorporer des contraintes empiriques de sélection comme la notion de seuil de rapport d'expression. Néanmoins, la procédure est très sensible au biais d'échantillonnage et ne sera véritablement pertinente que si on dispose d'un nombre important de lames témoins (Xie *et al.* 2005).

2.4. Séries temporelles et analyse différentielle

Lorsque l'on souhaite étudier un phénomène biologique dans le temps, ou l'influence d'une drogue selon des doses croissantes, le plan d'expérience n'est plus le même car les différents échantillons ne sont plus indépendants, ils sont ordonnés. Le plan en référence peut être utilisé en prenant, par exemple, le premier temps comme référence. Des plans en boucle peuvent également êtres développés en respectant l'ordre chronologique dans la boucle. De nombreux articles ont été publiés montrant de telles analyses par exemple sur le cycle cellulaire chez la levure (Spellman *et al.* 1998 ; Luan et Li 2004 ; Peng *et al.* 2005), le cycle parasitaire chez *Plasmodium falciparum* (Bozdech *et al.* 2003) ou encore le cycle circadien chez la souris (Beau *et al.* 2005). La plupart de ces travaux utilisent des analyses de profils d'expression et font appel à des techniques de classification. Des tests de profils différentiels sont réalisés grâce à des techniques de permutations notamment.

Lorsque l'on veut, en plus d'une analyse en série temporelle, réaliser une analyse différentielle, le plan d'expérience devient beaucoup plus difficile à mettre en œuvre. Guo *et al.* (2003) proposent une méthode non paramétrique pour détecter une autocorrélation (test de Wald) dans une série temporelle et pour chaque gène. Les probabilités sont calculées en utilisant une méthode de permutation similaire à celle décrite dans l'algorithme SAM (Tusher *et al.* 2001) (cf. 5.1.). De Gregorio *et al.* (2001) ont tenté de détecter les gènes induits ou réprimés dans la réaction immunitaire chez la drosophile. Dans leur approche, la variabilité n'est absolument pas intégrée puisque les données sont systématiquement moyennées pour tous les types de répétitions. Des critères empiriques sont alors utilisés sur les logarithmes des rapports d'expression entre mouches infectées et non infectées à chaque pas de

temps pour sélectionner un jeu de 400 gènes différentiels parmi les 13 197 gènes que contient la puce Affimetrix DrosGenome1. Une classification est finalement réalisée sur les 400 gènes sélectionnés. Le travail de Lin *et al.* (2004) recherchant les gènes cibles potentiels du récepteur activé α de l'œstrogène est plus abouti, mais n'intègre pas encore une planification expérimentale bien définie et les données sont largement sous-exploitées. Yang *et al.* (2003) et Glonek et Solomon (2004) ont proposé des stratégies expérimentales pour effectuer ce type d'analyse. Enfin, Liang *et al.* (2005) proposent une analyse comparative de diverses méthodes paramétriques et non paramétriques permettant la recherche de gènes différentiels dans des séries temporelles.

Ce paragraphe dépasse le cadre de nos expériences sur *Buchnera*. Néanmoins cette réflexion s'insère dans les perspectives du projet. Pour le moment, les pucerons sont laissés de 2 à 4 jours sur le milieu artificiel avant l'extraction des ARNm. La réponse transcriptionnelle de *Buchnera* est ainsi une réponse « à l'équilibre », lorsque les flux sont stabilisés. Il serait très intéressant de pouvoir observer une réponse cinétique du transcriptome de la bactérie. Un plan adéquat devra alors être défini. Pour le moment, ce travail n'a pas été réalisé car la détermination de l'échelle de temps de l'analyse reste très délicate. Un autre aspect concerne l'inférence de réseaux de régulation à partir de données d'expression, ce type d'analyse nécessite également la prise en compte d'un facteur temporel. Ce point sera repris dans la partie perspective du rapport.

3. Analyse qualité et filtration

L'analyse qualité et la filtration sont des étapes cruciales du processus d'analyse des données d'expression. Ce sont elles qui permettront l'obtention de données de qualité et augmenteront la puissance, la sensibilité et la spécificité des analyses statistiques. C'est aussi au cours de ces étapes que le choix de la méthode de normalisation sera effectuté. Nous avons réalisé un protocole détaillé décrivant les paramétrages du scanner et du logiciel GenePix, ainsi que la démarche générale et les commandes R nécessaires à la visualisation des nuages de points. Ce protocole sera prochainement disponible sur le site du PRABI. Dans ce chapitre, une brève analyse critique de ces différentes étapes et les figures correspondant à nos deux expériences YF et AAS seront proposées.

3.1. Détection du signal (lecture au scanner)

La première étape du processus consiste à mesurer la fluorescence des spots grâce à un scanner. Les problèmes de quenching des fluorochromes et de non linéarité du signal sont abordés par Ramdas *et al.* (2001) et Taylor *et al.* (2001).

Différentes options permettent d'améliorer la qualité et la reproductibilité du signal détecté. L'ajustement le plus couramment utilisé consiste à augmenter la puissance des photomultiplicateurs (PMT) jusqu'à obtenir une image à la limite de la saturation afin de conserver un maximum de linéarité entre le signal et l'abondance des transcrits. Smyth *et al.* (2002) prétendent que, contrairement à une idée souvent préconçue, il paraît plus important de se placer pour chaque lame à la saturation limite, plutôt que d'assurer des valeurs de PMT identiques pour toutes les lames. Notre expérience nous a conduit à suivre cette recommandation. Néanmoins, il sera toujours préférable de ne pas travailler avec des lames présentant de trop fortes disparités d'intensité de signal pour obtenir une variabilité acceptable. Pour tenter de remédier à ce problème, le logiciel Masliner (Dudley *et al.* 2002) a été développé ; il facilite l'analyse d'une expérimentation faite avec plusieurs images réalisées avec plusieurs valeurs de PMT. Masliner permet de corriger les problèmes de saturation et de non linéarité. Ce type de méthodes n'a pas été utilisé jusqu'à présent sur la plateforme du DTAMB.

3.2. Détection du signal spécifique et du bruit de fond (analyse de l'image)

La seconde étape consiste à déterminer l'intensité spécifique du spot par rapport à l'intensité du bruit de fond. Il s'agit ici d'un problème d'analyse d'image géré de façons différentes par les divers logiciels commerciaux à un niveau local dans l'environnement direct du spot (Smyth et Speed 2003). Cette étape se déroule en trois phases, la localisation du spot, la segmentation (définition des contours) et l'intégration de la fluorescence. Différentes méthodes de segmentation ont été développées, mais ces aspects ne seront pas détaillés ici. De bonnes revues sur ces problèmes ont été publiées par Yang *et al.* (2001b) et Kafadar et Phang (2003). Wu *et al*. (2001) proposent une méthode d'estimation du bruit de fond basée sur un jeu de spots témoins (tampon de dépôt) répartis sur toute la surface de la lame.

Du point de vue de l'expérimentateur, l'élimination du bruit de fond paraît toujours une étape nécessaire. Néanmoins, du point de vue de l'analyste, cette soustraction est souvent discutable et même néfaste. En effet, si l'estimation du bruit de fond est effectuée avec la même précision que celle du signal, la variabilité sur la mesure sera fortement augmentée par la soustraction, et les analyses statistiques perdront de la puissance. Cette soustraction n'a de sens que lorsque des biais systématiques importants sont présents sur les lames. Notre expérience nous conduit à conseiller de soustraire le bruit de fond seulement lorsque des voiles importants sont visibles sur les lames d'un jeu de données. La fonction « maImage » de la bibliothèque de fonctions R marray est très efficace pour visualiser de tels artéfacts (figure 14).

Un autre problème lié a la soustraction du bruit de fond est l'obtention de valeurs négatives lorsque le signal est très faible (gène faiblement exprimé), la fluorescence extérieure pouvant alors dépasser l'intensité spécifique du plot. La transformation logarithmique des données est alors impossible sans l'élimination de ces valeurs négatives ou nulles. Différentes méthodes ont été proposées pour résoudre ce problème. La plus simple est celle développée par Edwards (2003). Elle consiste à transformer les valeurs négatives par une fonction affine pour les ramener à une valeur comprise entre 0 et 1. La transformation d'Edwards permet de ne pas générer de données manquantes. Elle est facile à mettre en place, néanmoins la distribution des points transformés est très différente de celles des points non transformés. L'inclusion de ces valeurs dans les modèles statistiques peut biaiser fortement les analyses. Kooperberg *et al.* (2002) ont proposé une approche Bayesienne pour reconstituer la valeur des spots négatifs beaucoup plus performante car basée sur des hypothèses de distributions et intégrant la variabilité des pixels associée au signal spécifique ou au bruit de fond.

La figure 14 illustre deux cas extrêmes représentés par les expériences YF et AAS. Dans la première expérience le biais systématique lié au bruit de fond est très fort et nécessite une soustraction, dans la deuxième, il est faible et les données sont analysées sans soustraction.

3.3. Elimination des mauvais spots (filtration)

Les logiciels d'analyse d'image proposent différents critères pour évaluer la qualité des spots, comme la variance du signal ou du bruit de fond, l'aire (A) ou le diamètre (d) du spot. A partir de ces variables, on peut calculer des indices permettant une sélection plus efficace comme la circularité (A/d^2) ou le coefficient de variation CV (écart-type / moyenne). D'un point de vue général, les « mauvais » spots sont caractérisés par une taille anormale (trop grosse ou trop petite), une mauvaise circularité, un fort CV dans le signal ou dans le bruit de fond, ou enfin, un faible rapport signal / bruit. Dès qu'une puce possède plus de 1000 spots, la détection des spots anormaux, pour être correcte, doit impérativement être automatisée en suivant des critères objectifs. Le logiciel GenePix propose différentes implémentations de procédures très intéressantes et faciles à réaliser (Fielden *et al.* 2002). La filtration des données avec le logiciel va consister à associer une valeur de qualité (un flag) à chacun des spots de chaque lame. GenePix propose différents niveaux de qualité : (-100) mauvais spot ; (-75) spot vide ; (-50) spot non détecté ; (0) spot non repéré et (100) bon spot. La détection du signal n'est pas un problème complètement résolu, Ekstrom *et al.* (2004) proposent une étude de différentes méthodes de transformation très intéressante.

Une littérature très abondante est disponible sur le sujet et beaucoup d'auteurs ont proposé des méthodes de filtration, empiriques pour certaines (Wu *et al.* 2001), statistiques pour d'autres (Tseng *et al.* 2001 ; Finkelstein *et al.* 2002a ; Jensen *et al.* 2002, Chen *et al*. 2002). Smyth et Speed (2003) proposent une revue très complète et détaillée de ces différents travaux. Nous avons effectué cette analyse qualité avec le logiciel GenePix selon la démarche suivante, comprenant sept étapes :

(1) détection manuelle des spots situés sous des rayures ou des taches ;

(2) détection semi-automatique des spots de tailles anormales ;

(3) détection automatique des spots saturés ;

(4) détection semi-automatique des spots montrant une mauvaise corrélation entre moyenne et médiane des pixels ;

(5) détection semi-automatique des spots montrant une forte variabilité du bruit de fond ;

(6) détection semi-automatique des spots montrant une forte variabilité du signal ;

(7) détection automatique des spots à fort rapport signal sur bruit.

Les six premières étapes visent à détecter les « mauvais » spots et la dernière sélectionne un jeu de « bons » spots. Le paramétrage du logiciel pour réaliser cette détection sera disponible prochainement sur le site du PRABI. La figure 14B illustre quelques-unes des figures anormales de plots classiquement rencontrées.

L'analyse qualité aboutit à l'association d'un indice qualité à chacun des spots des lames du jeu de données. Il reste à utiliser cette valeur d'indice dans la suite du processus d'analyse. Cet indexage de la qualité sera très utile pour la visualisation graphique des données permettant de décider de l'élimination d'une lame par exemple ou de prévoir la démarche de normalisation qui va suivre. Les bibliothèques de fonctions « marray » proposent un ensemble de représentations graphiques pertinentes présentées sur la figure 14. L'indexage de la qualité permet également de moyenner les répétitions techniques par niveau de qualité. A l'issue de cette analyse visuelle, nous suggérons de regrouper les index de qualité en deux groupes : les bons spots à conserver pour l'analyse (de –50 à 100), et les mauvais à rejeter (-100), les valeurs des spots vides (-75) ayant été éliminées physiquement. Les moyennes des répétitions techniques peuvent alors être envisagées en éliminant les mauvais spots. Si pour un gène, aucun spot de qualité n'est disponible, alors, la moyenne est calculée sur l'ensemble des mauvais spots et l'indice (-100) est associé à la valeur moyenne pour ne pas générer de trous dans le tableau de données. Des fonctions R ont été développées au laboratoire pour générer ces moyennes (Morin 2004).

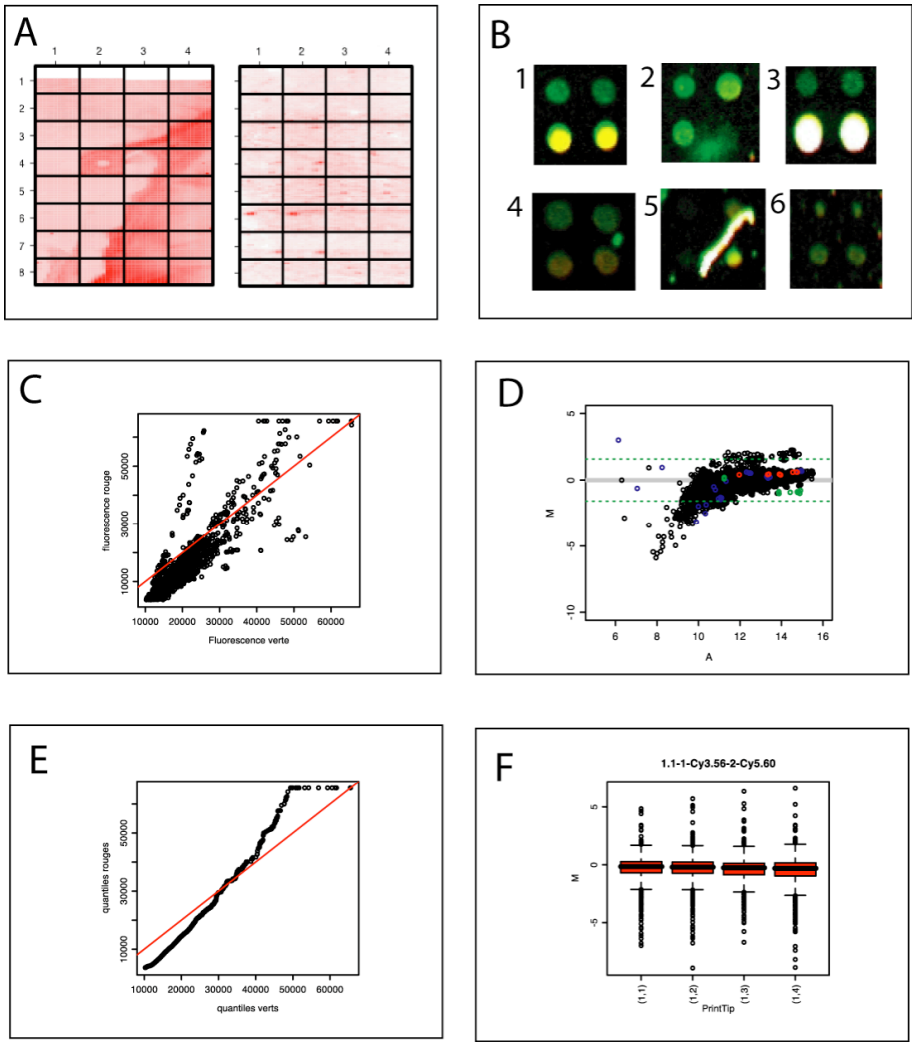

Figure 14. Quelques graphes utiles pour l'analyse qualité. A : la fonction « maImage » permet de représenter une variable sur la puce, ici le bruit de fond rouge sur une mauvaise lame à droite et sur une bonne à gauche. B : Différentes figures de spots, normaux (1), masqués par une tache ou une rayure (2, 5), saturé (3), grains dans le bruit de fond (4), taille anormalement faible (6). C : graphe R=f(G). D : graphe MA, $\log_2(R/G)=f(1/2\log_2(RG))$. E : graphe quantile – quantile. F : boîte à moustaches permettant la comparaison de la distribution des groupes de spots correspondant aux quatre aiguilles de dépôt d'une même lame.

4. Normalisation des données d'expression

La normalisation des données d'expression consiste, dans le cas des lames à double couleur, à calibrer les niveaux des deux fluorescences pour estimer des rapports d'expression corrects entre les deux conditions expérimentales comparées. Dans la mesure où la fluorescence est détectée par des amplificateurs électroniques (PMT) aucune mesure absolue de fluorescence ne peut être réalisée. Il est donc nécessaire de normaliser.

La normalisation des données de fluorescence doit également permettre la distinction entre les sources de variations aléatoires (biologiques et expérimentales) et les sources de variations systématiques. Parmi les biais systématiques, on peut noter : (1) le biais de dépôt lié à l'hétérogénéité des aiguilles du robot de dépôt ; (2) le rendement de marquage et la stabilité différentielle des fluorochromes ; (3) la non linéarité de la détection, la saturation du signal et le « quenching » de la fluorescence. Il est très important à ce sujet que les erreurs techniques et systématiques ne soient pas confondues avec les effets biologiques dans le plan d'expérience. Sinon, il ne sera pas possible de tirer de conclusion biologique de l'expérience.

Enfin, il faut noter que l'étape de normalisation est souvent intégrée à l'analyse statistique différentielle, notamment lorsque des modèles d'analyse de la variance sont utilisés. Le découpage entre l'étape de normalisation et l'analyse statistique n'est ainsi très souvent que théorique. Parmi les bonnes revues sur les méthodes de normalisation des puces double couleur, on peut citer Schuchhardt *et al.* (2000) ; Bilban *et al.* (2002) ; Smyth et Speed (2003).

4.1. Choix des gènes de normalisation

La plupart des méthodes de normalisation sont basées sur l'hypothèse suivante : l'expression de la majorité des gènes analysés ne varie pas entre les différents traitements comparés, ou alors, ces variations sont symétriques entre les gènes réprimés et les gènes activés. La normalisation va alors consister à calibrer les signaux de fluorescence en considérant cette égalité moyenne pour l'ensemble du nuage de points. Cette hypothèse est, dans de nombreux cas, tout à fait acceptable. Elle est utilisée avec succès pour les puces pangénomiques (très grand nombre de gènes analysés) ou bien lorsque les conditions analysées sont très modérées. Par contre, lorsque les conditions d'études font que les hypothèses d'homogénéité et de symétrie des rapports de fluorescence ne sont plus respectées, soit parce que le nombre total de gènes est trop faible, soit parce que le traitement testé est

susceptible d'induire des changements d'expression majeurs et non symétriques, il est nécessaire de normaliser sur la base d'un jeu de gènes invariants dans l'expérience.

L'utilisation d'un jeu de « gènes de ménage » exprimés constitutivement est problématique car l'expression constitutive a été remise en cause pour de très nombreux gènes de ménage. De plus, ces gènes sont généralement tous très fortement exprimés et ne conviennent pas pour réaliser une normalisation dépendante de l'intensité (Eickhoff *et al.* 1999 ; Smyth et Speed 2003).

L'utilisation d'étalons internes, « *spikes* », n'est pas non plus une technique idéale car l'estimation des concentrations d'ARNm ou d'ADNc marqués est assez délicate et coûteuse (van de Peppel *et al.* 2003). Chez les levures et les bactéries, de l'ADN génomique (ADNg) est parfois utilisé car il contient de quoi permettre une hybridation pour toutes les cibles potentielles. Chez les eucaryotes supérieurs, l'importance de l'ADN non codant et la présence d'exons rendent l'utilisation de l'ADNg très problématique à cause de la faiblesse du signal obtenu. Yang *et al.* (2002) ont développé, chez la souris, un système basé sur l'utilisation d'un mélange d'ADN issus de banques d'expression (EST).

Une dernière approche consiste à sélectionner statistiquement les gènes invariants dans une expérimentation donnée. Kepler *et al.* (2002) proposent de réaliser cette sélection en même temps que la normalisation et l'analyse statistique différentielle. La méthode consiste à introduire, dans leur modèle, une pondération du gène liée à sa réponse au traitement pour décider s'il participera à la normalisation. Comme il n'est pas possible de savoir *a priori* si un gène est invariant, une procédure itérative est développée. Une autre approche basée sur la méthode des rangs a été développée par Tseng *et al.* (2001). Dans une expérience comparant deux traitements, les intensités de fluorescence *Rg* et *Gg* sont ordonnées. Un gène *g* est considéré comme invariant si le rang *Rg* ne diffère pas du rang *Gg* à plus d'une différence de *d*, et si l'intensité moyenne n'est pas dans les *l* plus fortes ou les *l* plus faibles. La détermination du paramètre *d* est réalisée par une méthode itérative (Schadt *et al.* 2001). Le seuil *d* est calculé sur la base du nombre de gènes sélectionnés à l'étape antérieure et multiplié par un pourcentage *p*. La limite *l* est appliquée seulement pour la première itération. Les valeurs des paramètres (*d*, *l* et *p*) sont définies empiriquement. Cette procédure a été implémentée sous R par Nicolas Morin (2004), elle est illustrée sur la figure 15.

Une fois le jeu de gènes invariants sélectionné, la normalisation est réalisée sur cet ensemble de points selon l'une des méthodes décrites dans le paragraphe suivant, puis la fonction de normalisation est appliquée aux autres points du nuage.

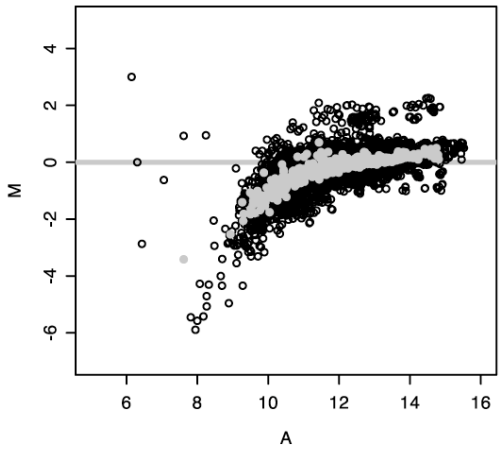

Figure 15. Détection des gènes invariants par la méthode de Tseng *et al.* (2001). Lame 1 du jeu de données AAS. Les gènes invariants sont en gris.

4.2. Les différentes méthodes de normalisation

4.2.1. Notations

Les données d'intensité sont rarement manipulées sans transformation. La transformation la plus utilisée est la transformation logarithmique à base 2 ($\log_2$). La transformation $\log_2$ permet de transformer les rapports de fluorescence en différence, de façon à rétablir une symétrie entre sur- et sous-expression, de stabiliser la variance et de transformer les erreurs multiplicatives en erreurs additives. En effet, si les erreurs sont proportionnelles à la valeur du signal, alors après transformation, elles deviennent constantes sur l'échelle de variation. L'utilisation d'un log à base 2 correspond également au format des images 16-bit fourni par les scanners avec des intensités variant de 1 à 2^{16}, ce qui donne une échelle de 1 à 16 pour les valeurs transformées.

Soit *xg* l'abondance de l'ARNm correspondant au gène *g*. I_{fg} est l'intensité de fluorescence (signal brut) correspondant au fluorochrome *f* pour le gène *g*. Pour limiter le nombre d'indices, on utilisera souvent les notations *Rg* et *Gg* pour désigner *I1g* et *I2g*, les intensités de fluorescence correspondant aux fluorochromes Cy5 (rouge) et Cy3 (vert) pour le gène *g*.

$Y_{fg} = \log_2 I_{fg}$ est la transformée logarithmique de l'intensité,

$M_g = \log_2 \frac{R_g}{G_g}$ est le logarithme du rapport de fluorescence rouge/vert,

$A_g = \log_2 \sqrt{R_g G_g}$ est l'intensité moyenne des deux couleurs du spot.

Toutes les méthodes développées dans cette partie concernent, sauf mention contraire, la normalisation des deux fluorescences sur une même lame.

4.2.2. Différents modèles reliant l'intensité du signal de fluorescence et l'abondance de l'ARNm

Dans l'approximation la plus simple, on peut considérer la relation suivante pour un gène *g* donné et *ε*, un terme d'erreur additif indépendant de l'intensité du signal mesuré :

$$I_g = bx_g + \varepsilon_g \qquad \text{Eq. 1.}$$

Ce modèle permet de normaliser « par la moyenne » les deux fluorochromes en prenant comme hypothèse que les abondances moyennes du jeu de gènes de normalisation sont identiques dans les deux conditions, qu'il n'existe pas de bruit de fond et aucun effet d'intensité. De nombreuses analyses ont montré que cette relation simple ne peut pas être appliquée dans la plupart des cas (Kepler *et al.* 2002). Une amélioration du modèle consiste à introduire un terme de bruit de fond (*a*) dépendant du fluorochrome :

$$I_{fg} = a_{fg} + b_f x_g + \varepsilon_g \qquad \text{Eq. 2.}$$

Ce modèle va correspondre à l'ajustement d'un modèle linéaire de type anova. Les hypothèses associées à ce modèle sont la linéarité entre abondance des transcrits et intensité de fluorescence, l'absence d'effet de l'intensité et une erreur additive.

Des modèles plus complets ont été proposés par Kepler *et al.* (2002) et Rocke et Durbin (2001) intégrant soit la non linéarité du signal, soit une modélisation plus complète du terme d'erreur. Le modèle de Rocke et Durbin (2001) se formule de cette façon :

$$I_{fg} = a_{fg} + b_f x_{fg} e^{\eta_g + \zeta_{fg}} + \varepsilon_g + \delta_{fg} \qquad \text{Eq. 3.}$$

Le terme d'erreur est composé d'une erreur multiplicative ($\eta_g + \zeta_{fg}$) et d'une erreur additive ($\varepsilon_g + \delta_{fg}$), ces deux termes étant décomposés en une partie

indépendante et une partie dépendante du fluorochrome. Tous les termes d'erreurs ont pour moyenne 0 et une distribution symétrique. Sur la base de ce modèle Cui *et al.* (2002) ont réalisé des simulations de données d'expression et ont reproduit, en faisant varier l'importance relative de chaque type d'erreur, des profils caractéristiques illustrant les phénomènes de saturation, d'hétéroscédasticité pour les faibles ou les fortes valeurs d'intensités ou de mauvaise extraction du bruit de fond. Les méthodes de normalisation associées à ces modèles utilisent des transformations particulières permettant de stabiliser les variances sur toute la gamme d'intensités (Cui *et al.* 2002 ; Huber *et al.* 2002).

Tous ces modèles visent à expliquer la réaction d'hybridation entre une cible (l'ARNm) et sa sonde. Pourtant, aucun d'eux n'intègre les propriétés de la réaction d'hybridation elle-même, et notamment le fait que la température de fusion (Tm) est inversement proportionnelle à la concentration en ARNm (Freier *et al.* 1986). Autrement dit, plus un ARNm est rare, moins il s'hybridera efficacement. Ce biais augmente le signal des gènes abondants et diminue le signal des gènes rares. De même, il est théoriquement possible d'aborder le problème de l'hybridation croisée (contamination de la fluorescence d'un spot par l'hybridation aspécifique d'ARNm de séquence proche) par une modélisation de la réaction d'hybridation. En effet, des modèles sont disponibles pour calculer l'affinité de deux sondes imparfaitement appariées (Zuker 2003). Même si des approches empiriques ont été développées pour mesurer l'impact de l'hybridation croisée (Finkelstein *et al.* 2002a), aucune méthode de normalisation intégrant ces notions n'a été développée à ce jour.

4.2.3. Les différentes méthodes de normalisation

Normalisation globale « par la moyenne »

Selon le modèle le plus simple (Eq. 1), la méthode de normalisation dite « par la moyenne » consiste à calculer le rapport des moyennes des signaux de fluorescence sur les deux canaux ($T = \sum_{g=1}^{N} R_g / \sum_{g=1}^{N} G_g$), et de le rapporter à 1. Cette méthode fait l'hypothèse de la proportionnalité entre la fluorescence mesurée et l'abondance des ARNm. La transformation correspondante s'écrit :

$$\begin{cases} R'_g = R_g \\ G'_g = TG_g \end{cases} \qquad \text{Eq. 4.}$$

Si l'on travaille directement sur les logarithmes des rapports de fluorescence, la transformation s'écrit alors :

$$\left(\log_2 \frac{R}{G}\right)' = \log_2 \frac{R}{G} - c = \log_2 \frac{R}{kG} \qquad \text{Eq. 5.}$$

Avec c = $\log_2$ k la moyenne des $\log_2$(R/G) du jeu de données.

Cette méthode de normalisation s'applique lorsque le graphe MA[2] ne montre pas de courbure en fonction de l'intensité, et lorsque la variabilité des M est constante sur toute la gamme d'intensité (figure 16). La médiane, plus robuste vis-à-vis des points extrêmes, peut être utilisée comme estimateur central de la distribution de points (figure 16B).

Normalisation globale par régression linéaire

Si on considère un modèle intégrant une valeur de bruit de fond (Eq. 2), on peut alors écrire la relation suivante pour un gène *g* non différentiel :

$$\begin{cases} R_g = a_R + b_R x_g \\ G_g = a_G + b_G x_g \end{cases}$$
$$\Rightarrow R_g = a_R - \frac{b_R}{b_G} a_G + \frac{b_R}{b_G} G_g \qquad \text{Eq. 6.}$$

La relation *Rg =f(Gg)* est linéaire. La normalisation va consister à ramener à 0 la valeur de l'ordonnée à l'origine $a_R - \left(b_R / b_G\right)$ et à 1 celle du rapport (*Rg/Gg*).

Finkelstein *et al.* (2002b) proposent, pour réaliser la normalisation, une régression de la forme $\log R = a + b \log G$ estimée pour tous les plots de la lame. Afin de ne travailler que sur des gènes invariants, les points correspondant aux résidus *e* de la régression tels que $|e| > 2\sigma_{res}$ sont éliminés, et le modèle est réajusté. La régression est réitérée jusqu'à ce que l'écart entre deux R^2 successifs devienne inférieur à un seuil fixé. Les coefficients de la régression sont alors utilisés pour normaliser *logR* par rapport à *logG*.

[2] graphe du rapport d'expression différentiel, *M=log₂(R/G)*, en fonction de l'intensité du spot *A=O,5log₂(RG)*.

$$\begin{cases} \log_2 R' = \log_2 R - a = \log_2 (R/\alpha) \\ \log_2 G' = b \log_2 G \end{cases} \Rightarrow \begin{cases} R' = R/\alpha \\ G' = G^b \end{cases} \qquad \text{Eq. 7.}$$

Sur la base de ce même modèle, Newton *et al.* (2001) et Kerr *et al.* (2002) proposent deux autres transformations moins fortes appelées « shift-log » où une constante *C* est ajoutée (ou retranchée) à la valeur de fluorescence avant de calculer le logarithme. Ces transformations effectuent une translation de l'ordonnée à l'origine du graphe *R=f(G)* de façon à annuler le terme constant, mais elles n'ajustent pas la pente comme précédemment (Eq. 6). La constante *C* pourrait être estimée par régression aux moindres carrés, mais les auteurs affirment que la méthode n'est pas très robuste vis-à-vis de l'existence de gènes différentiels. Elle est donc estimée de façon à minimiser l'écart absolu de chaque *logR* autour de la médiane des *logR* (c'est le critère MAD, *Median Absolute Deviation*) :

$$\sum_{g=1}^{N} \left| \log_2 \left(\frac{R_g + C}{G_g - C} \right) - Médiane \frac{R_g}{G_g} \right|$$

Une autre façon d'effectuer cette normalisation consiste à utiliser un modèle d'analyse de variance et d'inclure dans l'analyse les termes de bruit de fond et d'influence du colorant. Ces modèles seront développés dans le paragraphe suivant. Ces transformations font l'hypothèse que, sur la gamme de fluorescence observée, la relation *R = f(G)* est linéaire. Néanmoins, on montre que dans la plupart des cas, cette linéarité n'est pas respectée.

Normalisation globale par régression linéaire locale

La méthode de normalisation par régression linéaire locale est basée sur le même principe que précédemment, la seule différence résidant dans l'estimation des paramètres de la régression qui utilise une fonction « *loess* ». L'estimation loess est une régression quadratique locale dont les paramètres sont estimés par la méthode des moindres carrés pondérés. Le poids des observations est proportionnel à la distance au centre de voisinage. Le paramètre « span » (fonction *loess* sous R) défini la taille du voisinage, c'est-à-dire le pourcentage de points du nuage utilisés pour l'ajustement des paramètres locaux. La transformation qui en découle peut s'écrire de la façon suivante (Yang *et al.* 2002) :

$$\begin{cases} \log_2 R' = \log_2 R_g + C_g/2 \\ \log_2 G' = \log_2 G_g - C_g/2 \end{cases} \quad \text{Eq. 8.}$$

En travaillant directement sur les rapports d'expression, la transformation peut s'écrire simplement :

$$M'_g = M_g - C_g(A) \quad \text{Eq. 9.}$$

$C_g(A)$ est la fonction *loess* estimée sur le graphe MA. Ce type de normalisation est le plus utilisé, il est implémenté dans les bibliothèques de fonctions marrays du logiciel R (figure 16C).

Workman *et al.* (2002) ont proposé une autre méthode de normalisation non linéaire appelée « qspline » utilisant une approximation spline cubique des distributions des intensités de fluorescence. Cette méthode a été implémentée pour la normalisation des puces Affimetrix dans la bibliothèque de fonctions R « *affy* ». Selon les auteurs, cette méthode présente l'avantage de ne pas faire l'hypothèse de l'existence d'un faible nombre de gènes différentiels, et ne nécessite donc pas de choix de gènes invariants. Elle présente de plus l'avantage de calibrer les distributions de toutes les lames d'un jeu de données relativement à une distribution de référence qui est constituée par la moyenne géométrique, v_g, des *m* sondes du jeu de données :

$$v_g = \sqrt[m]{\prod_{i=1}^{m} I_{gi}} \quad \text{Eq. 10}$$

La transformation peut être appliquée au cas des lames en double marquage. Pour une lame donnée, les percentiles q_R, q_G et q_v sont calculés à partir des vecteurs R_g, G_g et du vecteur v de référence. Un graphe Quantile-Quantile (*qqplot*) permet d'observer la corrélation entre les couples de percentiles (q_R, q_v) et (q_G, q_v). Chaque couple permet d'ajuster une fonction spline cubique, de la forme : $S_{f=R,G} = f(q_v, q_f)$ avec f la fonction génératrice de type spline (*splinefun* sous R) qui ajuste les paramètres d'une fonction naturelle cubique (B-spline). La fonction S_f est utilisée pour interpoler la courbe pour chaque intervalle entre deux couples de percentiles (q_f, q_v). La normalisation est alors réalisée de façon à ramener la courbe vers la diagonale, selon la transformation suivante :

$$\begin{cases} R' = S_R(q_v) \\ G' = S_G(q_v) \end{cases} \quad \text{Eq. 11}$$

Normalisation globale intégrant l'hétéroscédasticité

Toutes les méthodes présentées jusqu'ici font l'hypothèse d'une erreur additive constante entre le logarithme de la fluorescence et l'abondance des ARNm. L'observation des graphes MA montre souvent une variabilité forte pour les faibles valeurs de A. Pour pallier ce problème, certains auteurs ont tout simplement suggéré d'éliminer les données en-dessous d'une certaine valeur de signal (Yang *et al.* 2001a). Ce qui revient à réaliser une filtration plus drastique jusqu'à obtenir l'homoscédasticité.

Néanmoins, d'autres transformations ont été proposées visant à modéliser la relation entre la variance et l'intensité pour l'intégrer dans un test statistique. Ces méthodes sont présentées dans le paragraphe 5.1 (*modélisation de la variance*).

Normalisation spatiale

Au sein même d'une lame, une hétérogénéité spatiale est parfois observée. Cette hétérogénéité peut être liée à une hétérogénéité du support, du dépôt, de l'hybridation ou du rinçage de la lame. Dans de nombreux cas, cette hétérogénéité est confondue avec le facteur aiguille (car une seule aiguille du robot dépose les plots d'une zone donnée). Yang *et al.* (2002) proposent d'éliminer l'hétérogénéité spatiale en intégrant une normalisation aiguille dépendante « *printTip Loess* ». Cette normalisation peut s'écrire de la façon suivante :

$$M'_g = M_g - c_{ga}(A) \qquad \text{Eq. 12}$$

avec c_{ga} la fonction « lowess » estimée pour la a^{ème} aiguille du robot de dépôt.

Si on considère que l'élimination de l'effet aiguille ne permet pas d'améliorer l'homogénéité spatiale, il est possible d'ajuster une correction sur la surface de la lame à partir des coordonnées (x, y) des spots. De la même façon, on peut utiliser une fonction loess qui estime un paramètre $c_{x,y}$ (Colantuoni *et al.* 2002 ; Cui *et al.* 2002). Wernisch *et al.* (2003) ajustent une loess bidimensionnelle sur leurs données et définissent des zones d'intensités, sur lesquelles ils appliquent une normalisation de type « *PrintTip Loess* ». Enfin, Mary-Huard *et al.* (2004) ont montré sur leurs lames, ainsi que sur diverses lames déjà publiées, un biais de dépôt très important en utilisant une méthode de détection issue de la géostatistique, le variogramme. Les auteurs ne proposent pas de méthode statistique particulière pour résoudre le problème, mais ils insistent sur l'importance d'une répartition aléatoire des sondes sur la puce. En effet, si cette répartition est, par exemple, une répartition alphabétique suivant le nom des gènes, les gènes d'un même opéron seront systématiquement consécutifs et la normalisation spatiale risque d'effacer systématiquement toutes les régulations à l'échelle des opérons.

Normalisation inter-lames

Les méthodes ci-dessus sont adaptées pour normaliser les deux fluorescences d'une même lame l'une par rapport à l'autre. Une calibration inter-lames peut ensuite être envisagée. Yang *et al.* (2002) proposent une méthode de calibration inter-groupe de dépôt qui peut être appliquée à l'inter-lame. Cette calibration stipule simplement que la variance des rapports *R/G* de la lame *i* s'écrit sous la forme $a_i^2\sigma^2$. Les a_i sont estimés en utilisant le critère MAD (déviation médiane absolue). Cette fonction est disponible dans la bibliothèque de fonctions R *marray*.

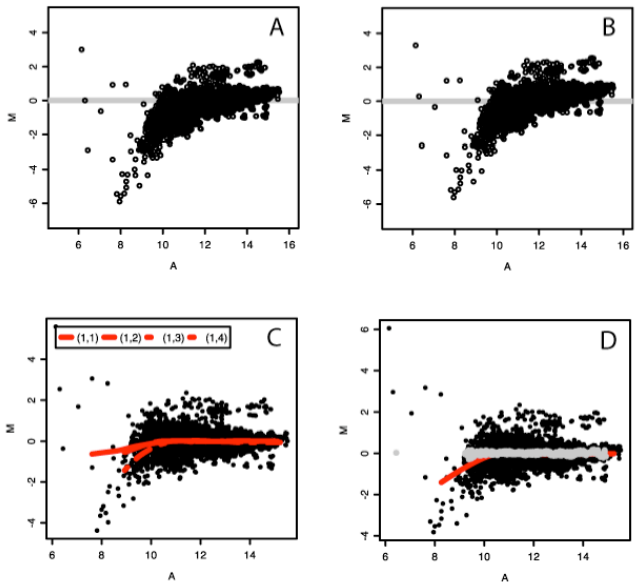

Figure 16. Illustration de différentes méthodes de normalisation. **A** : données non normalisées (lame 1 de l'expérience AAS), **B** : normalisation par la médiane, **C** : normalisation loess par aiguille, **D** : normalisation loess sur un jeu de gènes invariants (en gris).

5. Sélection de gènes différentiels

L'analyse du transcriptome génère des tableaux de données de grande taille et très dissymétriques, avec peu de variables (colonnes) et beaucoup de gènes (lignes). Ces caractéristiques vont avoir des conséquences importantes sur les analyses statistiques. Historiquement, la détection des gènes différentiels a débuté par la

simple observation des valeurs des rapports d'expression, les jeux de données étant souvent constitués d'une seule ou d'un très faible nombre de lames. Des seuils fixés entre trois et cinq étaient considérés comme significatifs, par comparaison à des étalons internes ou à des jeux de gènes de ménage déposés sur la puce (Schena *et al.* 1995 ; DeRisi *et al.* 1996 ; 1996). La bibliothèque de fonctions R « sma » regroupe un certain nombre de méthodes « unilames » pour l'observation des rapports d'expressions notamment les méthodes de Chen et Zhu (1997), Yue *et al*. (2001) et Newton *et al.* (2001) basées sur la relation de proportionnalité entre variance et moyenne des niveaux de fluorescence des gènes d'une même lame. Avec l'amélioration des processus, il a été possible d'augmenter la qualité et le nombre de lames des jeux de données laissant la place à des analyses statistiques plus performantes. Cui et Churchill (2003) proposent une très bonne revue des principales méthodes de base.

5.1. Comparaison de deux conditions expérimentales : tests multiples de Student

On se place dans les deux cas suivants : (1) deux traitements A et B sont comparés sur une seule lame et chaque gène est présent n fois sur la lame, (2) n lames sont hybridées avec les deux traitements, et chaque gène est présent une seule fois sur chaque lame (sinon, on calculera les moyennes par gène des répétitions techniques sur chaque lame). Il est alors possible de calculer la statistique de Student suivante :

$$t_g = \frac{\overline{M}_g}{s_g / \sqrt{n}}, \text{ avec } \overline{M}_g = \frac{1}{n}\sum_{i=1}^{n} M_{gi} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{gi}^A - Y_{gi}^B\right),$$

s_g est l'écart type estimé du gène *g* sur les n répétitions.

L'estimation de s_g est problématique dans la mesure où le nombre de répétitions est souvent faible (<10). Plusieurs méthodes ont alors été proposées pour améliorer cette comparaison.

Estimation spécifique de la variance

Il s'agit d'estimer s_g de façon indépendante pour chacun des gènes, c'est l'hypothèse d'hétéroscédasticité :

$$s_g^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(M_{ig} - \overline{M}_g\right)^2 \qquad \text{Eq. 13}$$

Cette formulation n'est pas satisfaisante. Lorsque le nombre de gènes est important avec un faible nombre de répétitions, les fluctuations aléatoires de la variance vont rendre le test peu puissant et peu spécifique. Lorsque la variance est forte, un gène, bien que différentiel, ne sera pas sélectionné alors que le gène, même s'il n'est pas différentiel, sera sélectionné quand sa variance sera faible.

Estimation globale de la variance

Il s'agit de calculer une variance commune à l'ensemble des gènes de l'analyse, s^2_{com}, c'est l'hypothèse d'homoscédasticité. Pour obtenir cette variance, il suffit de calculer la moyenne des variances individuelles puisque l'on a postulé n constant pour chaque gène. L'estimation de la variance est ainsi réalisée sur un grand nombre de gènes, le test est donc très puissant, mais reste très peu spécifique (beaucoup de faux positifs).

Une façon de pallier le problème consiste à ne tester que les gènes dont la variance individuelle est du même ordre de grandeur que celle de la variance commune (Martin-Magnette et Robin 2004). Les gènes testés devront vérifier la condition suivante :

$$\chi^2_{(\alpha'/2;n-1)} < (n-1)s_g^2 / s^2_{com} < \chi^2_{(1-\alpha'/2;n-1)} \qquad \text{Eq. 14}$$

avec $\alpha' \approx 1/G$ (G étant le nombre de gènes analysés).

Estimation combinée de la variance

Il s'agit de calculer localement une variance stabilisée indépendante de l'intensité d'expression. La méthode la plus célèbre pour réaliser ce test est la méthode SAM (Significance Analysis of Microarray) développée par Tusher *et al.* (2001). Cette méthode consiste à réaliser des tests individuels par gène, en ajoutant une constante à l'estimation de la variance individuelle s_g, pour éliminer la dépendance entre la variance et la mesure de l'expression des gènes :

$$t_g = \frac{\overline{M}_g}{(s_g + a)/\sqrt{n}} \qquad \text{Eq. 15}$$

Pour estimer *a*, Tusher *et al*. (2001) calculent localement différentes valeurs de t_g au moyen d'une fenêtre glissante sur l'ensemble du jeu de données. Ils choisissent alors la valeur de *a* qui minimise le coefficient de variation de la valeur absolue de la statistique t_g. Efron *et al.* (2001) développent une approche similaire mais utilisent le 90ème percentile de la distribution des s_g pour estimer *a*. Dans la méthode SAM, les

probabilités associées à chaque gène sont calculées par permutation des conditions expérimentales. Diverses méthodes non paramétriques ont été développées pour accéder à une estimation de la distribution des t_g sous l'hypothèse nulle, le gène n'est pas différentiel (Pan *et al.* 2002 ; Pan 2002 ; 2003). Cette technique non-paramétrique a l'avantage de permettre un contrôle du nombre de faux positifs (FDR). Elle nécessite, en revanche, un grand nombre de répétitions. Broberg (2003) propose une implémentation intéressante (fonction « samroc ») de la méthode SAM avec un contrôle du nombre de faux positifs et de faux négatifs en utilisant les courbes ROC (Receiver Operating Characteristics) dans la bibliothèque de fonctions « *SAG* ». Mais cette bibliothèque de fonctions n'a pas été maintenue de R dans le projet Bioconductor. Kim et Park (2004) proposent d'estimer a en utilisant la variance associée au gène g dans un jeu de lames similaires issu des bases de données d'expression. Enfin, Wu (2005) propose une formalisation de la méthode SAM en utilisant une méthode de régression linéaire pénalisée. La méthode a été généralisée au test de F pour la comparaison de plusieurs conditions expérimentales et présente des caractéristiques de puissance et de contrôle de faux positifs intéressantes. Une implémentation R est téléchargeable sur le site : http://www.biostat.umn.edu/~baolin/research (janvier 2006).

Modélisation de la variance

Il s'agit d'introduire une modélisation de la relation entre la variabilité et l'intensité de l'expression des gènes. De nombreuses approches ont été développées dans ce sens. Baggerly *et al.* (2001) utilisent des modèles β-Binomiaux pour modéliser la relation entre les répétitions issues d'une même lame et des modèles γ-Poisson pour les répétitions entre des lames différentes. Ces modèles permettent une calibration des rapports d'expression et la sélection de gènes différentiels. Rocke et Durbin (2001 ; Durbin *et al.* 2002) proposent un modèle exponentiel (erreurs multiplicative et additive) reliant l'intensité et l'abondance (Eq. 3). Les auteurs proposent une estimation séparée des paramètres sur des sous-ensembles de points sur- et sous-exprimés, puis proposent l'estimation d'intervalles de confiance pour les deux groupes de gènes de façon à sélectionner les gènes différentiels. Huber *et al.* (2002) et Durbin *et al.* (2002) proposent une transformation *arsinh* en faisant l'hypothèse de l'existence d'une relation quadratique entre la variance et l'intensité relative du signal détecté sur la puce. Cui *et al.* (2002) proposent une transformation *linlog* basée sur le modèle de Rocke et Durbin (2001). Leur transformation est basée sur le fait que, dans le modèle d'erreur proposé, les erreurs additives sont prédominantes pour les faibles valeurs d'intensité, alors que les erreurs multiplicatives apparaissent prédominantes pour les fortes valeurs d'intensités. La transformation *linlog* combine une transformation linéaire et une transformation logarithmique avec une transition lissée prenant en compte les avantages des deux méthodes. Ces transformations ne sont pas détaillées dans ce travail, on les trouve dans la bibliothèque de fonctions R « *vsn* ». Cui *et al.* (2002) affirment que même si

les modèles permettent de simuler des données d'expression avec des distributions très proches des distributions observées, ils fonctionnent relativement mal sur les données réelles pour détecter les gènes différentiels.

Les modèles de mélange

L'utilisation des modèles de mélange pour les analyses de données d'expression a été introduite par Ghosh et Chinnayian (2002) et Martin (2002) dans le contexte de la classification, et Lee *et al*. (2000), Irizarry *et al*. (Irizarry *et al.* 2003), Pan *et al*. (2003) et Delmar *et al.* (2005) pour l'analyse différentielle. L'analyse complète de ce type de modèles n'a pas été réalisée. Seul le principe de la méthode pour l'analyse différentielle des lames à double couleur sera présenté ici. On se réfèrera à la thèse de Delmar (2005) pour des informations plus complètes.

Considérons le modèle suivant :

$$y_{gr} = \mu_g + E_{gr} \qquad \text{Eq. 16}$$

y_{gr} représente le logarithme du rapport d'expression normalisé du gène *g* pour la répétition *r* du plan d'expérience. Le paramètre μ_g représente le rapport d'expression différentielle moyen du gène *g* au sein des différentes répétitions du jeu de données. E_{gr} représente le terme d'erreur, c'est une variable aléatoire de moyenne nulle et de variance σ_g^2. Les méthodes de mélange supposent que l'ensemble des gènes peut être partitionné en *m* groupes de variance homogène :

$$E_{gr} \rightarrow N\left(0,\ \sigma_g^2\right) \text{ avec } \sigma_g^2 \in \left[\sigma_{c1}^2, \sigma_{c2}^2, ..., \sigma_{cm}^2\right] \qquad \text{Eq. 17}$$

Les termes *cm* désignent les différentes classes de gènes qui constituent le modèle de mélange. On raisonne ensuite sur la somme des carrés des écarts des résidus, $X_g = \nu s_g^2$, avec, ν=R-1 le degré de liberté (R est le nombre de répétitions supposé constant pour tous les gènes), et s_g^2 l'estimation de la variance σ_g^2. X_g est distribué selon une loi *Gamma* de paramètres σ_g^2 et ν. Le principe des modèles de mélange consiste à considérer que la distribution des X_g suit un mélange de *m* lois *Gamma* correspondant aux *m* classes de gènes de variance homogène. Il est alors possible de quantifier la vraisemblance de l'échantillon en définissant pour chaque gène une probabilité *a priori* d'appartenir à l'une des *m* classes. La vraisemblance du modèle est optimisée pour permettre l'estimation des paramètres du modèle (notamment les σ_g^2) et il est possible d'estimer une probabilité *a posteriori* pour chaque gène d'appartenir à l'une des *m* classes. La détection des gènes différentiels

sera ensuite réalisée en analysant la répartition des rapports d'expression dans chacune des *m* classes et en sélectionnant les gènes montrant des rapports d'expression significatifs par rapport à la variance estimée pour leur classe d'appartenance. Une implémentation de ces méthodes est accessible dans la bibliothèque de fonctions R « *varmixt* » (Delmar *et al.* 2005).

5.2. Comparaison de plusieurs conditions expérimentales : anova

5.2.1. Modèles à effets fixes

La formulation du problème de l'analyse des données d'expression par l'ajustement de modèles d'analyse de la variance a été présentée par Kerr et ses collaborateurs (2000) du groupe de G. Churchill (Jackson Laboratory, USA). Ces modèles ont été implémentés dans la bibliothèque de fonctions R *maanova* développée par Wu *et al.* (2003). Le modèle proposé s'écrit de la façon suivante :

$$y_{lftg} = \mu + L_l + D_f + V_t + G_g + LG_{\mathrm{lg}} + VG_{tg} + \varepsilon_{lftg} \qquad \text{Eq. 18.}$$

y_{lftg} représente le logarithme de l'intensité du gène *g* sur la lame *l*, marqué par le fluorochrome *f*, dans la condition expérimentale *t*. μ représente le signal moyen de toutes les expériences, *L* est l'effet lié aux différentes lames, *D* est l'effet du colorant, *V* représente l'effet des traitements testés et *G* l'effet des gènes. Seules quelques interactions sont considérées dans le modèle car la prise en compte de toutes les interactions saturerait le modèle (cf. § 2.2). La plupart des effets introduits sont des variations systématiques qu'il est nécessaire d'éliminer, c'est la partie normalisation du modèle. Par exemple, l'effet lame introduit une calibration inter-lame, l'interaction *LG* correspond essentiellement à l'effet dépôt. Dans ce type de modèle, tous les effets sont supposés additifs et linéaires, ou log-linéaires si l'intensité est exprimée en logarithme. Une façon d'intégrer la non linéarité de la relation fluorochrome-expression, par exemple, consiste à travailler sur des données préalablement normalisées par la méthode *loess* (Eq 3.4.8). L'interaction *VG* (traitements x gènes) permet de tester l'existence de gènes différentiels en utilisant les contrastes souhaités pour chaque gène. Enfin, ε_{lftg} représente le seul terme aléatoire du modèle, il est distribué selon une loi normale centrée sur 0 et de variance σ^2. Nous verrons plus loin que, comme dans le cas des tests de Student, il est possible de constituer différentes hypothèses pour l'estimation de la variance résiduelle du modèle. Outre les bibliothèques de fonctions R *maanova* et *lima*, de nombreux outils ont été développés pour ajuster des modèles d'analyse de la variance sur des données de transcriptome. On peut citer, par exemple, les logiciels *GeneAnova* (Didier *et al.* 2002) et GeneSight (Draghici *et al.* 2001 ; 2003).

Outre sa capacité de normalisation, l'ajustement du modèle permet d'appréhender l'influence des différents facteurs du plan d'expérience sur la variabilité de l'expression des gènes. Ainsi, Jin *et al.* (2001) ont testé l'influence du sexe, du génotype et de l'âge sur l'expression des gènes de différentes souches de drosophiles. Ils montrent ainsi que le facteur le plus influent est le sexe, avant le génotype. L'âge n'intervient que minoritairement. Les auteurs analysent également certaines interactions, et les gènes significatifs caractéristiques des différents facteurs. On peut citer également le travail de Pritchard *et al*. (2001) portant sur l'analyse de la variabilité de l'expression des gènes chez différents tissus murins. Ce travail a généré le jeu de données *kidney* encore très utilisé dans la littérature.

Cet ajustement global est néanmoins problématique à cause de la taille du jeu de données atteignant rapidement la capacité de mémoire limite des ordinateurs classiques. Dans l'expérience AAS par exemple, avec seulement 617 gènes analysés, 16 lames, deux fluorochromes et deux facteurs croisés à deux niveaux (A et S), la matrice à inverser après simplification du modèle possède encore 19 744 lignes. L'ajustement a néanmoins pu être réalisé (tableau 3). Nous avons introduit le terme *E* (série d'échantillons) dans le modèle, afin d'intégrer le fait que les deux lots de huit lames n'ont pas été réalisés simultanément et pas sur les mêmes échantillons. L'introduction du terme *E* (sous-ensemble de lames) a pour effet de diminuer le degré de liberté du carré associé au facteur lame. Il est à noter que l'échantillon soit le facteur intégrant le plus de la variabilité dans le modèle. Les carrés moyens associés aux effets d'intérêts (*G:S, G:A et G:S:A*) sont de 1 000 à 10 000 fois moins importants. De plus, les carrés moyens associés aux traitements A et S ne sont pas indépendants, à cause du déséquilibre du plan (cf. figure 10), leur estimation dépend de l'ordre d'introduction, nous avons vérifié que cette variation est toutefois très faible.

Source	**ddl**	**CM**
E	1	$3{,}55\ 10^{10}$
L	14	$1{,}40\ 10^{9}$
D	1	$1{,}92\ 10^{7}$
G	584	$8{,}18\ 10^{8}$
A	1	$7{,}37\ 10^{6}$

Source	**ddl**	**CM**
S	1	$8{,}39\ 10^{5}$
A:S	1	$3{,}05\ 10^{6}$
G:S	585	$6{,}48\ 10^{5}$
G:A	585	$7{,}77\ 10^{5}$
G:S:A	585	$1{,}84\ 10^{6}$

Tableau 3. Décomposition des effets de l'expérience AAS. E : série échantillons, L : lame, D : fluorochrome, *g* : gène, A : acide aminé, S : sucre.

Pour gérer ce problème de grands tableaux de données, Wu *et al.* (2003) ajustent un premier modèle appelé « normalisation » de la forme suivante :

$$y_{lftg} = \mu + L_l + D_f + LD_{lf} + r_{lftg} \qquad \text{Eq. 19}$$

La matrice de ce modèle n'est pas inversée durant cette étape. Seules les moyennes des effets de normalisation sont estimées et retranchées pour le calcul des résidus. Ceux-ci sont alors insérés dans un modèle gène à gène de la forme suivante :

$$r_{lftg} = G_g + VG_{tg} + DG_{fg} + LG_{\lg} + \varepsilon_{lftg} \qquad \text{Eq. 20}$$

L'ajustement du modèle global en deux étapes ne pose pas de problème spécifique dans la mesure où les gènes sont systématiquement équilibrés sur le plan d'expérience. Par contre la méthode ne permet pas d'accéder aux carrés moyens de la décomposition du modèle complet.

L'estimation de la variance résiduelle du modèle gène à gène va permettre de détecter les gènes différentiels selon le même type d'hypothèses que celles décrites dans le paragraphe 5.1. Quatre types de tests sont implémentés dans la bibliothèque de fonctions *maanova*. Les trois premiers tests F_1, F_2 et F_3 correspondent respectivement aux hypothèses d'hétéroscédasticité, d'homoscédasticité et d'estimation composite :

$$F_1 = \frac{\left(rss_o - rss_1\right)/\left(df_0 - df_1\right)}{rss_1 / df_1}$$

$$F_2 = \frac{\left(rss_o - rss_1\right)/\left(df_0 - df_1\right)}{\sigma^2_{pool}}$$

$$F_3 = \frac{\left(rss_o - rss_1\right)/\left(df_0 - df_1\right)}{\left(rss_1 / df_1 + \sigma^2_{pool}\right)}$$

Les termes rrs_0 et rss_1 sont les sommes des carrés résiduels correspondant aux hypothèses H0 et H1 et possédant les degrés de liberté df_0 et df_1, σ^2_{pool} est l'estimation de la variance globale correspondant à l'hypothèse d'homoscédasticité. Le quatrième test F_s est basé sur une méthode de réduction de la variance utilisant l'estimateur de réduction de James-Stein. La procédure est décrite dans la publication de Cui *et al.* (2005). Brièvement, cette méthode estime une variance réduite s^2_g à partir de l'ensemble des variances individuelles et de leur moyenne géométrique. Si les variances sont très homogènes, le test converge vers le test F_2 (homoscédasticité) ; si les variances sont très hétérogènes, des poids plus importants sont attribués aux variances individuelles. Les auteurs montrent que le test F_s est le plus puissant et le plus robuste des quatre tests, il a été implémenté pour l'utilisation

de modèles mixtes. Ce test a été utilisé dans l'analyse des effets du plan d'expérience AAS.

Enfin, Kerr *et al.* (2002) pour faire face aux problèmes de dépendance de la variance des gènes et de leur intensité d'expression, proposent d'estimer les p-values correspondant aux différents F, non plus sur des hypothèses de distributions, mais en réalisant des permutations (bootstrap) soit sur les résidus de l'analyse, soit directement sur les conditions expérimentales si celles-ci sont assez nombreuses. Ces méthodes sont implémentées dans la bibliothèque de fonctions R *maanova*.

5.2.2. Modèles mixtes

Les plans d'expériences de puces à ADN incluent systématiquement des répétitions des gènes sur la lame, des répétitions de lames pour le même individu (échantillon) et/ou des « dye swap », et enfin des répétitions d'individus au sein des différentes conditions expérimentales. Il existe donc une hiérarchisation des sources d'erreurs. De plus, certains facteurs sont fixes alors que d'autres sont aléatoires. La prise en compte de ces différentes sources de variabilité est importante. Le risque majeur à ne pas tenir compte des effets hiérarchisés est de surestimer fortement le nombre de gènes différentiels, dans la mesure où les tests sont réalisés en prenant en compte une variabilité résiduelle d'un niveau trop inférieur. D'autre part, les modèles mixtes vont permettre de mieux utiliser les données d'expression des lames à double couleur. En effet, dans cette technologie, la véritable mesure est le rapport d'expression des gènes (hybridation compétitive des ADNc fluorescents), plutôt que la mesure absolue de chacune des fluorescences.

Wolfinger *et al.* (2001) ont été les premiers à utiliser les modèles mixtes dans une analyse de données d'expression en introduisant une composante « lame » aléatoire dans le modèle de Kerr. D'autres travaux ont ensuite été développés. A titre d'exemple, Wernisch *et al.* (2003) ont appliqué un modèle mixte pour l'analyse de l'expression des gènes chez *Mycobacterium tuberculosis*. Les auteurs souhaitaient comparer un mutant porteur d'une délétion sur le gène *trcS* (histidine kinase) relativement à la souche sauvage. A partir de trois échantillons d'une même culture des deux souches, trois hybridations ont été réalisées, puis la culture a été recommencée et trois nouvelles hybridations ont été réalisées. Sur les six lames de l'analyse, la souche mutante est systématiquement associée à la couleur rouge alors que la souche témoin est verte. Le modèle de Wernisch s'écrit alors de cette façon :

$$M_{gsl} = \mu + G_g + S_s + GS_{gs} + (S/L)_{sl} + (GS/L)_{gsl} + \varepsilon_{gsl} \qquad \text{Eq. 21}$$

M_{gsl} représente le logarithme du rapport d'expression entre les souches sauvages et mutante pour le gène g, de la culture s et hybridé sur la lame l. Les termes μ et G sont fixes dans le modèle. Les autres sont aléatoires et sont supposés normalement

distribués, de moyenne nulle et de variance σ_x^2. *S* représente l'effet culture (variance σ_s^2), *L* est l'effet lame (variance σ_a^2), *S/L* et *GS/L* indiquent la hiérarchisation des effets *S* et *GS* sous le facteur lame *L*. Ces deux termes sont également aléatoires et participent aux composantes de la variance du modèle mixte. Il est alors possible de déterminer une expression de la variance de l'effet gène (effet d'intérêt du modèle en fonction des composantes du modèle). Cette variance s'exprime de la façon suivante :

$$Var(\hat{G}_g) = \frac{1}{n_S}\left(\sigma_S^2 + \sigma_{GS}^2\right) + \frac{1}{n_S n_L}\left(\sigma_{SL}^2 + \sigma_{GSL}^2 + \sigma^2\right) \qquad \text{Eq. 22.}$$

L'ajustement du modèle n'est pas trivial, les estimateurs classiques de la variance au maximum de vraisemblance sont biaisés et des valeurs négatives sont obtenues. Les auteurs utilisent alors la méthode du maximum de vraisemblance résiduel (REML) pour réestimer des variances positives. Néanmoins, cette étape nécessite une simplification (basée sur l'équilibrage du plan d'expérience) pour être utilisée avec une matrice de si grande taille (3474 gènes analysés). Ces aspects relativement complexes ne sont pas développés dans ce mémoire. Après ajustement et avec les hypothèses de normalité émises, il est possible de calculer un intervalle de confiance autour de la valeur des rapports d'expression et de sélectionner ainsi les gènes différentiels. Comme les rapports d'expression sont centrés sur 0, on peut déterminer un seuil de significativité δ de la façon suivante :

$$\delta = \Phi^{-1}\left(1 - \alpha / n_G\right)\sqrt{Var(\hat{G}_g)} \qquad \text{Eq. 23}$$

Φ^{-1} est la fonction inverse de la distribution normale cumulée, α est le risque de première espèce et n_G le nombre de gènes de l'analyse permettant un contrôle du FWER selon la méthode de Bonféronni. En prenant un risque de 0,01 et avec 3474 gènes analysés, les auteurs trouvent un seuil δ égale à 0,89. C'est-à-dire que tous les gènes exprimés à plus de $2^{0,89} = 1,86$, ou moins de $2^{-0,89} = 0,54$, sont différentiels.

Enfin, les auteurs remarquent que les hypothèses émises sur la distribution des paramètres du modèle sont relativement fortes, notamment celle de l'indépendance entre la variance du rapport d'expression et l'intensité de l'expression des gènes (hypothèse d'homoscédasticité versus hétéroscédasticité). Les auteurs suggèrent alors d'utiliser des transformations stabilisatrices de la variance au cours de la normalisation pour résoudre le problème (Huber *et al.* 2003). Delmar *et al.* (2005) proposent une approche alternative intéressante qui combine des modèles de mélange (hétéroscédasticité) avec l'ajustement d'un modèle mixte.

L'utilisation de modèles mixtes nécessitera très souvent une collaboration étroite avec une équipe de statisticiens, car les méthodes d'ajustement et les interprétations des sorties de logiciels deviennent relativement complexes à analyser. La définition du plan d'expérience est, elle aussi, rendue beaucoup plus délicate et devra nécessairement être réalisée en partenariat avec les bioanalystes. Quelques articles ont été publiés récemment sur ces aspects et notamment les travaux de Rosa *et al.* (2005) et Dobbin *et al*. (2005) qui mettent en avant une polémique sur le choix des plans en référence ou en boucle associés aux modèles mixtes. Il est très difficile de résumer l'ensemble de ces travaux et de tirer une conclusion sur les avantages et les inconvénients de ces deux types de plans, car l'optimalité d'un plan d'expérience peut s'exprimer de façon très différente en termes de puissance, de robustesse vis-à-vis des hypothèses ou des valeurs manquantes et aberrantes, du coût de l'expérience ou du matériel biologique. Les plans sont adaptés à un type de questionnement très précis, ils présentent alors des inconvénients chaque fois qu'on les utilise pour tester une hypothèse pour laquelle ils ne sont pas directement conçus. Ces aspects ne seront pas plus développés ici et l'on se réfèrera à la revue de Templeman (2005) pour des informations plus détaillées. Le rapport de Master d'O. Bagland (Paris XI) est également très informatif et propose un exemple de plan d'expérience impliquant 40 lames pour comparer l'expression de 10 patients trisomiques relativement à des patients sains avec la prise en compte du facteur sexe et un déséquilibre des effectifs des patients dans les classes (Bagland 2005).

Les modèles mixtes n'ont pas été utilisés sur nos deux expériences YF et AAS. Pourtant des répétitions techniques sont présentes sur la puce, sous la forme de quadruplets issus de deux aiguilles différentes, et sous la forme de différentes sondes pour les 617 gènes de *Buchnera.* Des effets aléatoires sont également présents : l'effet « lame » et l'effet « série d'échantillons ». Les effets d'intérêts (acides aminés et pression osmotique) sont fixes. La hiérarchisation des erreurs techniques a été gérée en effectuant des moyennes de façon à ne conserver qu'une seule valeur par gène et par couple de lame-fluorochrome. Néanmoins, ce calcul de moyenne est problématique dans la mesure où il est déséquilibré, certains gènes présentant 12 répétitions techniques (quatre fois trois sondes), d'autres seulement quatre (une seule sonde). Et ceci sans tenir compte de l'étape de filtration qui peut creuser encore les disparités de 1 à 12. L'effet principal de ce calcul de moyenne est la sous-estimation de la variabilité des gènes présentant le plus de répétitions, ce qui favorisera leur sélection dans le jeu de gènes différentiels. Enfin, l'effet échantillon n'a pas été intégré et l'effet lame a été considéré comme fixe. Une perspective à court terme sera donc d'utiliser des modèles adéquats pour réanalyser les données des deux expériences.

5.3. Méthodes bayésiennes pour la sélection de gènes différentiels

Comme leur nom l'indique, les méthodes Bayésiennes sont des méthodes probabilistes basées sur l'utilisation du théorème de Bayes. Ce théorème permet de calculer la probabilité *a posteriori P(M | D)* d'une hypothèse portée par un modèle *M*, si l'on dispose des observations *D* :

$$P(M \mid D) = \frac{P(D \mid M)P(M)}{P(D)} \qquad \text{Eq. 24}$$

Dans cette formulation *P(D | M)* est la vraisemblance des données sous les hypothèses du modèle, *P(M)* est la probabilité du modèle *M* sur la base de connaissances *a priori* des données. *P(D)* représente la vraisemblance des données sous l'hypothèse nulle.

Le premier article présentant une méthode probabiliste pour la recherche de gènes différentiels est celui de Chen et Zhu (1997). Les auteurs se sont essentiellement basés sur la constance du coefficient de variation de l'expression des gènes sur une seule puce ($c = \sigma_{Rg} / \mu_{Rg} = \sigma_{Gg} / \mu_{Gg}$). En faisant l'hypothèse de normalité des distributions des signaux rouge et vert sur la lame, les auteurs proposent une approximation de la loi de distribution du rapport d'expression ne dépendant que du paramètre *c*. L'estimation au maximum de vraisemblance de ce paramètre peut alors être réalisée avec l'ensemble des gènes de la lame, si on considère que très peu de gènes différentiels sont présents, ou sinon sur un sous-ensemble de gènes invariants. La détermination des intervalles de confiance pour des valeurs de rapports moyens a finalement été réalisée pour sélectionner les gènes différentiels. Baldi et Long (2001) ont implémenté une véritable optimisation Bayésienne du modèle de Chen dans le logiciel Cyber-T (fonction R *hdarray*). Ils intègrent la relation de dépendance entre l'intensité et la variance du signal de fluorescence dans la formulation de leur probabilité *a priori* en calculant les paramètres de distribution au moyen d'une fenêtre glissante incluant les 100 gènes d'expression les plus proches de celle du gène analysé. L'estimation de la variance moyenne du rapport de fluorescence estimée par la méthode aboutit à une formule très proche de celle définie par Tusher *et al.* (2001) dans la méthode SAM. Enfin, Newton *et al.* (2001) ont affiné la méthode en utilisant des lois Gamma pour décrire les distributions des signaux rouge et vert, mais là encore, une seule lame (ou une seule source de variabilité technique) est prise en compte dans le modèle.

De nombreux travaux ont ensuite été publiés. Nous pouvons citer : Efron *et al.* (2001), Theilhaber *et al.* (2001), Lönnstedt et Speed (2002), Townsend et Hartl (2002), Lin *et al.* (2003), Lee *et al.* (2003) et Hsiao *et al.* (2004). Ces auteurs ont construit des modèles Bayesiens intégrant des hypothèses supplémentaires (comme

une modélisation de la variance dans les deux derniers articles), ou des méthodes d'estimations et des implémentations différentes. Enfin, plus récemment Blangiardo *et al.* (2005) ont utilisé l'information apportée par des hybridations témoins (le même échantillon est hybridé en rouge et en vert sur la lame) pour enrichir leur connaissance *a priori* sur la variabilité expérimentale dans un modèle Bayesien hiérarchique.

Nous n'avons pas mis en œuvre de telles approches sur nos données expérimentales, c'est pourquoi ce chapitre est très peu détaillé et ne propose qu'une liste de publications.

6. Intégration des résultats dans la problématique biologique

L'analyse différentielle sélectionne des gènes surexprimés ou réprimés dans les différentes conditions expérimentales. L'étape suivante consiste à interpréter ces résultats dans le contexte biologique. Cette étape n'est pas triviale, c'est elle qui demande souvent le plus de temps de travail. Il s'agit d'intégrer des connaissances biologiques souvent très hétérogènes, comme l'annotation fonctionnelle, la localisation chromosomique, les interactions protéines-ADN ou les différentes structures de régulation (opérons et régulons). Cette intégration nécessite bien souvent un deuxième niveau d'analyse statistique. La dimension importante des jeux de données nécessite, de plus, l'utilisation d'outils spécifiques et des moyens de représentation graphique nouveaux. Il s'agit donc là d'un problème complexe et qui ne peut pas être abordé de façon exhaustive. Je présente brièvement ici les quelques outils et méthodes que nous avons utilisés dans nos expériences sur *Buchnera*.

Analyse fonctionnelle

Il s'agit de déterminer la fonction des gènes sélectionnés par l'analyse différentielle. L'ontologie GO (http://www.geneontology.org/) est classiquement utilisée pour cette tâche. GO est un vocabulaire contrôlé décrivant les fonctions moléculaires, les processus physiologiques (métabolisme et transduction du signal) et la compartimentation cellulaire des protéines. Ce vocabulaire est organisé selon une arborescence (graphe acyclique direct) qui permet d'associer un gène à plusieurs catégories fonctionnelles. L'arborescence permet, de plus, une description à différents niveaux de précision : au bout d'une branche, la précision est maximale et plus on remonte la branche plus la description est générale. Ainsi, le gène *flgA*, codant une protéine du flagelle chez *Buchnera*, est annoté dans les processus biologiques en bout de branche comme « *flagellum biogenesis* », puis en descendant vers la racine, on trouvera successivement : « *flagellum organization and biogenesis* », « *cell projection organization and biogenesis* », « *cellular*

morphogenesis », « *cell organization and biogenesis* », « *cell growth and/or maintenance* », « *cellular physiological process* », et « *cellular process* ».

Des outils spécifiques ont été développés pour associer des listes de gènes à des annotations GO, les classer par catégories fonctionnelles en fonction du niveau de précision de la description, et enfin réaliser des tests statistiques de comparaisons d'effectifs de classes entre deux listes de gènes. Ils sont pour la plupart disponibles sur le Web depuis la page d'accueil de Gene Ontology. Une bibliothèque de fonctions R *GOtools* est également disponible. Dans notre travail, nous avons utilisé l'outil Web *fatigo* (http://www.fatigo.org/) pour réaliser de telles requêtes (Al-Shahrour *et al.* 2004).

Ce type d'analyse permet de mettre en évidence des classes fonctionnelles de gènes spécifiquement induits ou réprimés. Des tests statistiques ont été développés sur *fatigo* pour évaluer si la proportion de gènes différentiels est significativement différente de la proportion de gènes invariants. Ces tests sont, de mon point de vue, assez peu informatifs. En effet, l'hypothèse alternative testée (la classe fonctionnelle est induite) ne sous-entend pas forcément qu'une majorité de gènes soit induite pour produire une réponse fonctionnelle. La régulation de la classe peut passer par l'activation ou la répression d'un seul gène clé.

Analyse génomique

Chez les bactéries, l'organisation du chromosome est très importante pour la régulation de l'expression des gènes. Cette organisation se manifeste à différents niveaux, le plus bas étant celui de la séquence elle-même avec, par exemple, les biais de composition en bases et l'usage du code génétique. Au niveau des gènes, il est possible d'étudier les structures en opérons ou en régulons. Enfin à l'échelle du chromosome, on pourra s'intéresser à la distance des gènes, à l'ordonnée à l'origine ou à la répartition des gènes sur les brins directs et retardés (Lobry 1996). Cet ensemble d'analyses concerne le domaine de l'évolution moléculaire et nous avons essayé d'intégrer ce type de données dans notre problématique (chapitre 4).

Analyse des voies métaboliques

L'interprétation des données d'expression dans le contexte biologique passe par l'analyse des voies métaboliques. On peut inclure dans ce paragraphe l'analyse des mécanismes de transport et de la transduction du signal. L'intégration de ces données est des plus complexes car les informations nécessaires à la construction de ces voies sont souvent hétérogènes et issues d'analyses comparatives automatiques plus ou moins expertisées. La fiabilité des outils est ici est un problème majeur.

Peu d'outils descriptifs fonctionnels sont actuellement disponibles pour représenter une projection des données d'expression sur le réseau métabolique spécifique d'un organisme donné. Avec nos données, nous avons utilisé la base Kegg (http://www.genome.jp/kegg/). Cette base est riche en informations et d'utilisation très conviviale. Elle propose des visualisations de cartes métaboliques théoriques sur lesquelles il est possible de situer les gènes présents chez un organisme spécifié. Des outils de coloration permettent ensuite de repérer les gènes induits ou réprimés dans les différentes voies. La base possède des outils de représentation des génomes et des liens vers les principales bases de données pour l'annotation fonctionnelle. La principale limite de Kegg est liée à son cloisonnement et à son système d'annotation automatique basé sur les relations d'orthologie entre les gènes qui permet l'affiliation des enzymes dans les différentes voies. Notre expérience de l'utilisation de cette base montre que, pour *Buchnera*, beaucoup d'erreurs d'affiliations sont commises, les fusions de gènes ne sont pas intégrées et le système de coloration présente de nombreux dysfonctionnements. Aucun outil de représentation (ni statistique) n'est implémenté pour comparer l'activation ou l'inhibition d'une voie entre différentes conditions expérimentales. La description des systèmes de transport et de transduction du signal est très partielle.

La plateforme Genostar (http://www.genostar.org/) possède une structure modulaire et une implémentation objet qui lui procure une adaptabilité et une évolution beaucoup plus importantes que Kegg. Elle dispose de modules pour l'analyse génomique comparative, pour l'annotation fonctionnelle et la modélisation des voies métaboliques. Elle dispose également d'outils statistiques sous forme de tâches prédéfinies, mais adaptables à des objets de types différents. Elle est complètement basée sur une annotation expertisée. Enfin, l'intérêt majeur de cette plateforme est la possibilité d'effectuer des requêtes complexes sur le modèle de la base. Par contre, la modélisation des données d'expression n'a pas encore été intégrée dans le système, elle n'a donc pas été réellement utilisée sur nos données. Un projet de collaboration avec le consortium Genostar (F. Rechenmann) a débuté en 2004 pour initier l'implémentation d'un modèle de description des données d'expression. Le projet est en cours. La base de donnée biocyc (http://www.biocyc.org/) est également très intéressante, mais ne présente pas d'outils pour l'intégration des données d'expression.

On peut noter le travail très intéressant de G. Thomas (Université de York) qui a développé *buchneraBASE* (http://www.buchnera.org/). Cette base de données, dédiée aux génomes séquencés de *Buchnera,* possède des outils de génomique comparative, ainsi qu'une description très précise des réactions biochimiques catalysées par les enzymes de *Buchnera*. Elle possède le degré d'expertise le plus fort que l'on puisse trouver actuellement pour *Buchnera* puisqu'une réannotation complète a été réalisée en tenant compte des données récentes publiées dans la littérature. Cette base dérive de la base dédiée à *E. coli* ECHOBASE (Misra *et al.*

2005) qui a été un élément essentiel dans nos analyses. Les informations relatives aux systèmes de transport et de transduction de signaux sont disponibles mais non structurées pour des requêtes globales. Ici encore, les données d'expression ne peuvent pas être intégrées.

Ces différentes méthodes n'ont pas été utilisées sur nos données. Nous avons utilisé le test non paramétrique des rangs signés de Wilcoxon pour tester la régulation globale des voies métaboliques. Ce test très simple peut s'appliquer à n'importe quel groupe de gènes pour lequel on souhaite savoir si le rapport d'expression différentiel est globalement différent de 1 (ou de 0 si on travaille sur le logarithme du rapport d'expression). Sous l'hypothèse nulle, les rangs des M_g sont répartis aléatoirement autour de 0, le test unilatéral permet de tester une activation ou une répression globale de la classe. On pourra observer que l'activation d'une voie métabolique ne sous-entend pas forcément l'activation de la majorité des gènes de cette voie. Il peut exister des gènes limitants qui, à eux seuls, bloquent ou activent le flux de métabolites, ou à l'inverse, un ensemble de gènes très faiblement régulés qui, cumulant leurs effets, ont une influence majeure sur le flux de métabolites. Ce test reste un outil exploratoire.

Analyse du réseau de régulation

Le transcriptome est une vision globale de l'expression des gènes d'un organisme. Un des objectifs majeurs de ces prochaines années pour notre groupe est d'essayer de reconstruire le réseau d'interactions génétiques sous-jacent aux données d'expressions chez *Buchnera*. L'analyse des corégulations de gènes dans des conditions expérimentales variées permet d'inférer la structure statique du réseau (*i.e.*, les gènes sous le contrôle d'un même régulateur). Une analyse dynamique est ensuite nécessaire pour comprendre les cascades de régulation impliquées dans le phénomène biologique étudié.

Une littérature très abondante est maintenant disponible sur le sujet. On peut globalement séparer trois types de méthodes. Les premières utilisent des approches de simulations visant à reproduire le fonctionnement du réseau de régulation à partir de connaissances *a priori*, les données expérimentales d'expression venant ensuite valider les prédictions et/ou permettant de reformuler le modèle. Dans ce type de méthodes, on peut citer le travail du groupe de H. de Jong (De Jong *et al.* 2003), G. Bernot (Bernot *et al.* 2004 ; Guespin-Michel *et al.* 2004) ou P. Brazhnick (2005), par exemple. Ces travaux utilisent des systèmes dynamiques soit dans des approches quantitatives très gourmandes en paramètres, soit dans des approches qualitatives permettant d'utiliser des valeurs de paramètres relatives.

La deuxième classe de méthodes cherche à inférer le réseau à partir des données d'expression. Il s'agit pour la plupart de méthodes d'inférence Bayésienne. La

presse scientifique sur le sujet a littéralement explosé à partir de l'année 2004. Ces méthodes ne sont pas décrites dans ce mémoire. Pour débuter une analyse bibliographique, on pourra consulter les revues suivantes : Perrin *et al.* (2003), Brown *et al.* (2004), Mazurie *et al.* (2005) et Pe'er (2005).

Enfin, des méthodes de classification ont également été utilisées dans ce cadre et on peut citer, par exemple, le travail de J.-P. Vert (Vert et Kanehisa 2003 ; Yamanishi *et al.* 2003 ; Yamanishi *et al.* 2004). Ces chercheurs ont développé des méthodes de classification supervisée dans lesquelles le facteur optimisé est une combinaison entre la variance du nuage d'expression et un critère de régularité des niveaux d'expression le long du graphe représentant le réseau métabolique. Cette méthode permet ainsi de rechercher des corrélations entre les niveaux d'expression des gènes et la structure du réseau métabolique. Hanish *et al*. (2002) proposent une analyse très similaire. Cette méthode est, de mon point de vue, assez discutable car elle n'intègre pas le niveau de la régulation transcriptionnelle et suppose une corrélation entre le transcriptome et le réseau métabolique qui n'est pas fondée sur des arguments biologiques. Néanmoins, son principe est très intéressant et pourrait être appliqué aux réseaux de régulations génétiques avec peut-être plus de pertinence. Le groupe belge de van Helden propose une approche tournée vers l'extraction de voies métaboliques pertinentes qui visent plus à la caractérisation du réseau qu'à sa modélisation, sa simulation où son inférence (Croes *et al.* 2005).

Dans notre travail, nous n'avons pour le moment pas envisagé l'inférence du réseau de régulation de la bactérie *Buchnera*. De nombreux problèmes se posent pour envisager de telles études. Tout d'abord le nombre de conditions expérimentales est encore très faible dans notre collection de données, et les connaissances *a priori* sur les régulateurs transcriptionnels de la bactérie sont encore très fragmentaires. Les aspects cinétiques n'ont pas été intégrés également. Nos analyses sont réalisées « à l'équilibre », après deux jours de nutrition sur les différents milieux. L'isolement des bactéries dans les bactériocytes du puceron rend le choix du pas de temps de l'expérience très problématique. En effet, l'effet physiologique mesuré dépend avant tout de la nutrition du puceron sur le milieu artificiel.

Analyse systémique

Les approches de Biologie des Systèmes cherchent à caractériser une entité biologique comme un tout, et non pas comme la somme de ses différents composants (Aderem 2005 ; Kirschner 2005 ; Liu 2005). Deux écoles se sont constituées. La première défend le principe des bases de données et de l'extraction de connaissances ou « data mining ». Il s'agit d'accumuler de façon structurée un maximum d'informations de toute nature sur le système pour effectuer ensuite des requêtes transversales permettant la description du système. De très nombreux outils

et bases de données ont été développés dans ce sens. Le logiciel Blastsets (Barriot *et al*. 2004) que nous avons utilisé dans nos analyses, est un exemple d'outils (comparaison de classifications) permettant de réaliser de telles analyses. On peut citer également les outils « data mining » de la plateforme *GeneSpring GX* développée par la société Agilent (Palo Alto, USA), les systèmes *Indigo* (Nitschke *et al.* 1998), *GABRIEL* (Pan *et al.* 2002) ou *VAMPIRE* (Hsiao *et al.* 2005).

L'autre école est celle de la simulation (Adami 1998). A partir d'entités virtuelles extrêmement simplifiées, il s'agit de les faire évoluer et/ou interagir dans un contexte lui aussi très simplifié. Ces entités sont, de fait, bien souvent très peu réalistes. Mais l'intérêt de la simulation réside dans le fait de pouvoir tester des propriétés globales d'un système (comme l'émergence ou l'auto-organisation) avec un minimum d'hypothèses et de paramètres dans le modèle. Les modèles sous-jacents à ces techniques de simulation peuvent être des modèles individus centrés, et les techniques de simulation peuvent faire appel aux algorithmes génétiques (Soula *et al.* 2005). Mais évidemment, il n'est pas question de présenter ici une vision exhaustive de l'ensemble de ces approches. Le travail du groupe de Beslon est très intéressant pour notre problématique. Il consiste à modéliser la dynamique de l'évolution des génomes en relation avec la fonction des protéines (Knibbe *et al*. 2005). Dans le modèle, les gènes sont des suites de 0 et de 1, les protéines sont des fonctions triangulaires et les fonctions biologiques de ces protéines sont des sommes d'intégrales. Bien que ces objets paraissent très éloignés des entités biologiques qu'ils représentent, les résultats obtenus jusqu'alors sont très informatifs, notamment pour ce qui concerne l'évolution de la taille des gènes et des génomes en relation avec la robustesse du génome vis-à-vis des processus de mutations. Un des enjeux majeurs de ces prochaines années sera d'essayer de coupler nos résultats biologiques avec les simulations réalisées (cf. chapitre 5).

7. Conclusion : pourquoi des modèles et pourquoi toujours plus complexes ?

Des méthodes informatiques d'extraction de règles ou « data mining » ont été utilisées pour analyser des données de puces à ADN (Boulicault et Gandrillon 2004 ; Georgii *et al.* 2005). Certaines de ces méthodes sont des heuristiques très performantes permettant d'extraire d'un jeu de données préalablement discrétisé toutes les associations possibles entre les niveaux d'expression et les conditions expérimentales qui répondent à un ensemble de contraintes définies par l'utilisateur. J'ai entendu parfois des informaticiens défendre ces méthodes sous prétexte que, contrairement aux méthodes statistiques, elles ne supposaient aucune hypothèse sur les données et ne nécessitaient pas de calculs probabilistes complexes pour exploiter les résultats. Je souhaite conclure ce chapitre, non pas en dénigrant l'intérêt des méthodes d'extraction de connaissances, mais en montrant la nécessité des modèles

statistiques et leurs apports pour répondre aux questions posées par l'expérimentation.

Ainsi la règle suivante peut-elle être extraite : *« si les gènes X et Y sont surexprimés dans les conditions A et B et réprimés dans les conditions C et D, alors le gène Z est réprimé dans la condition E, 80 fois sur 100 ».* Cette règle représente un événement (ou un ensemble d'évènements) observé sur le jeu de données associé à une expérience particulière. Elle ne présente donc en elle-même aucun intérêt pour l'expérimentateur. Par contre, trois informations au moins motivent l'expérimentateur en biologie autour de cette règle : (1) Est-ce qu'elle aurait pu être observée par hasard (hypothèse H0), et sinon quelle est la raison qui a favorisé son apparition (hypothèse H1) ? (2) Peut-on généraliser cette règle pour la population dont est issu l'échantillon analysé ? (3) Quelles sont les hypothèses qui permettent d'assurer ces prédictions, et si ces hypothèses ne sont pas vérifiées, peut-on proposer une autre solution d'analyse ?

La première information est relative au risque α, à la puissance des tests statistiques, ainsi qu'à la définition des hypothèses H0 et H1. Nous avons vu dans ce chapitre que le contrôle du risque ou du nombre de faux positifs est un problème crucial dans ce contexte de tests multiples impliquant bien souvent plusieurs milliers de gènes pour un nombre très faible de conditions. La définition des hypothèses H0 et H1 est un problème encore bien plus fondamental, puisqu'il s'agit du principe même de l'expérimentation (Legay 1973).

La deuxième information est relative à la portée des conclusions que l'on veut émettre, c'est ce que l'on appelle l'inférence statistique. Généralement, on ne souhaite pas conclure simplement sur les trois ou quatre sujets de l'expérience, on souhaite pouvoir généraliser ces conclusions à la population entière. On ne peut pas concevoir une expérience sans élaborer une stratégie d'échantillonnage, une planification expérimentale et les modèles adéquats d'analyse.

La troisième information est relative à la robustesse des tests statistiques. Il est toujours très important de connaître les hypothèses sur les variables permettant la validité des analyses. Il s'agit ici encore d'un principe fondamental de l'expérimentation (Tomassone *et al*. 1993). La formalisation de ces hypothèses fait partie intégrante de la description du phénomène analysé.

Enfin, on peut noter que dans le cadre de la détection de gènes différentiels pour les données transcriptomiques, les modèles utilisés sont des modèles prédictifs. Néanmoins, dans certains cas, ils peuvent également avoir valeur de modèles explicatifs, comme nous l'avons vu pour des modèles d'erreurs de la variance par exemple. L'ajustement et la validation de tels modèles peuvent alors permettre d'adopter des stratégies d'amélioration de la technique et/ou du protocole

expérimental (effets de saturation, variabilité liée à la soustraction du bruit de fond, non linéarité des signaux, hybridations aspécifiques…).

En conclusion, la complexité des modèles ne doit pas rebuter les expérimentateurs. Elle est rendue nécessaire par la complexité des objets analysés et des expériences réalisées. L'analyse du transcriptome est une technique de pointe qui demande une maîtrise technologique très importante à cause de la robotisation des différentes étapes, de la miniaturisation et de la multiplicité des facteurs de variabilité s'accumulant au cours du processus. Si la compléxité technique nécessite plusieurs semaines d'expérimentation, c'est toujours plusieurs mois qui seront nécessaires à l'analyse et à l'intégration des résultats biologiques. Les modèles de plus en plus complexes que l'on rencontre dans la littérature correspondent à la prise en main des problèmes posés par les données d'expression par des statisticiens de plus en plus spécialisés. Il est certain qu'il n'est plus possible, pour une seule personne, de posséder les compétences pour construire des expériences autour d'une problématique biologique, pour effectuer les mesures d'expression et pour les analyser dans leur globalité. J'espère que ce travail pourra aider quelques chercheurs d'une des différentes communautés à mieux comprendre les enjeux pluridisciplinaires de cette technologie.

Chapitre 4

Analyse du transcriptome de *Buchnera aphidicola* en conditions de stress nutritionnel du puceron hôte, *Acyrthosiphon pisum*

1. Introduction

Le modèle biologique a été présenté au chapitre 1. Cette introduction définit brièvement la problématique : il s'agit de caractériser, au niveau transcriptionnel, le rôle trophique de la fonction symbiotique dans le couple *Buchnera* - *A. pisum*.

Deux questions principales sont posées :

(1) dans quelle mesure la régulation transcriptionnelle est-elle fonctionnelle chez *Buchnera* ?

(2) *Buchnera* est-elle capable de répondre significativement à une variation de la demande nutritionnelle de la part de son hôte ?

Pour répondre à la première question, nous avons analysé les résultats d'expression obtenus grâce à la puce *Buchnera*, indépendamment du contexte expérimental, pour essayer de caractériser une réponse globale significative de la bactérie. C'est ainsi que nous avons étudié la corrélation entre l'expression des gènes et l'organisation du génome de la bactérie (§ 3.2). Nous avons également analysé l'usage du code et l'expression des ARN de transfert de *Buchnera* (§ 3.3). Pour répondre à la seconde question, les résultats d'expression ont été replacés dans le contexte fonctionnel avec l'analyse des deux expériences YF et AAS (§ 3.4). La première expérience (YF) est très spécifique, elle se caractérise par la déplétion

complète du milieu nutritif du puceron en tyrosine et phénylalanine. La deuxième (AAS) est plus globale, elle consiste à analyser l'effet d'une diminution de la fourniture en acides aminés essentiels (de 50 à 25 %) dans le substrat du puceron, combiné avec un effet de la pression osmotique (0,5 M et 1 M de saccharose) selon un plan factoriel croisant les quatre modalités. L'augmentation de la pression osmotique a pour effet principal de diminuer le taux de croissance du puceron. En effet, dans ces conditions expérimentales relativement physiologiques le puceron est capable de faire face à l'hyperosmolarité du puceron. Une partie matériel et méthodes a été rédigée de façon condensée (§ 2, ci-dessous). Des descriptions plus complètes sont disponibles dans les articles associés à ces travaux (en cours de publication). Enfin, une discussion et une conclusion générale sont fournies à la fin du chapitre.

2. Matériel et méthodes

2.1. Pucerons, milieux artificiels et performances

Le clone parthénogénétique (LL01) du puceron *A. pisum* (Harris), maintenu en élevage au laboratoire depuis de nombreuses années (21°C, 16h/8h jour/nuit), a été utilisé dans ces expériences. Des femelles adultes ailées vivipares sont élevées, à faible densité, sur des jeunes pousses de fèves (*V. fabae*) pendant deux jours. Les larves pondues sur cette période sont transférées sur des milieux artificiels pour les expériences.

Dans l'expérience YF, deux types de milieux ont été préparés. La composition du milieu complet AP_3 est basée sur le profil d'acides aminés de pucerons élevés sur plantes (Febvay *et al.* 1999). Le milieu carencé YF ne diffère du milieu AP_3 que par l'absence de tyrosine et de phénylalanine. Dans la deuxième expérience (AAS), le milieu B (Prosser *et al.* 1992) a été utilisé comme référence. Ce milieu possède une composition proche de la composition de la sève phloémienne de *V. fabae*. Par contre, la concentration en acides aminés dans ce milieu est de 0,15 M, avec un rapport entre acides essentiels et non essentiels équilibré (50 %). Ce rapport ne correspond pas à la teneur moyenne du phloème, il est plus favorable au puceron, puisque les acides aminés essentiels sont normalement très rares dans la sève. La concentration en saccharose est de 0,5 M. Trois autres milieux ont été dérivés de ce milieu de référence, en diluant les acides aminés essentiels par deux (25 %) et/ou en augmentant la concentration en saccharose à 1 M.

Pour l'expérience YF, l'analyse des performances biologiques des pucerons a été conduite sur des larves néonates âgées de 0 à 12h et transférées sur milieu artificiel pendant sept jours. La mortalité des pucerons, les temps de développements

larvaires et la fécondité ont été mesurés chaque jour. Les femelles adultes ont été pesées individuellement le jour de leur mue imaginale. Un lot indépendant de cinq cages (3 individus par cage) a été conservé systématiquement pour l'analyse des acides aminés libres.

Pour l'expérience AAS, des larves de deux jours sont pesées et placées sur les différents milieux artificiels pour une durée de cinq jours. A la fin de l'expérience, les larves sont à nouveau pesées et la différence entre les deux masses permet d'estimer l'accroissement absolu de chaque puceron. Il est possible de calculer le taux d'accroissement protéique des pucerons en utilisant des concentrations protéiques de 60 et 53 µg de protéines par mg de puceron frais pour les larves néonates de 2 et 7 jours respectivement. Ces valeurs ont été estimées sur des expériences antérieures (Prosser *et al.* 1992). Si on divise ce taux par le poids des pucerons, on obtient l'accroissement protéique relatif individuel. Douglas *et al.* (2001) ont déterminé les concentrations en acides aminés protéiques chez des pucerons de différents stades, à partir de ces compositions. Il est alors possible de dériver les besoins spécifiques de chaque puceron pour tous les acides aminés. Finalement, afin de déterminer les quantités d'acides aminés apportées par le milieu nutritif au puceron, il faut connaître l'ingestion moyenne des pucerons. Pour cela des expériences ont été conduites grâce à une méthode de quantification utilisant de l'inuline radioactive selon la méthode de Karley *et al.* (2002) sur des lots de 14 pucerons pour les quatre milieux. La différence entre l'apport du milieu et l'accroissement individuel peut être déterminée pour chaque acide aminé essentiel. Il est ainsi possible de quantifier ce que fournit *Buchnera* en nutriments, c'est ce qu'on appellera « analyse du budget » dans la suite de ce mémoire. Cette analyse a été réalisée sur l'expérience YF en prenant des valeurs théoriques d'ingestion.

Les analyses statistiques (ANOVA et courbes de survies) ont été réalisées à l'aide du logiciel R et JMP (SAS Institute Inc., USA). L'index d'accroissement de la population r_m a été calculé comme indiqué dans Wyatt et White (1977). Pour les expériences de puces, les bactéries symbiotiques ont été purifiées par filtration comme décrit dans Charles et Ishikawa (1999).

2.2. Préparation de l'ARN et synthèse des ADNc

Le détail des protocoles pour la purification de l'ARN et le marquage sont décrits dans l'article de Calevro *et al.* (2004). Brièvement, l'ARN total est purifié à partir du culot de purification de *Buchnera* en utilisant le kit commercial Trizol (Invitrogen). L'ADN génomique est digéré grâce à l'enzyme DNAse RQ1 (Promega), puis éliminé sur colonne Rneasy (Quiagen). Un marquage fluorescent indirect des ADNc est alors réalisé par transcription reverse sur 15 µg d'ARN total en utilisant le kit CyScribe (Amersham). La solution d'amorces utilisée dans cette

réaction est constituée d'un mélange d'amorces aléatoires et d'amorces semi-spécifiques dédiées à *Buchnera*. L'ADN marqué est enfin purifié sur colonne d'exclusion G50 Autoseq (Amersham) juste avant l'hybridation.

2.3. Fabrication des lames

Les sondes utilisées sont des oligonucléotides de 35-mer aminomodifiés à l'extrémité 5' d'un bras carboné en C_6. Ces sondes ont été synthétisées par la société Eurogentec. L'optimisation de la séquence des sondes a été réalisée grâce au logiciel ROSO (Reymond *et al.* 2004a). Trois témoins positifs ont été utilisés : les gènes *pcp* et *pelL* d'*Erwinia chrysanthemi* et le gène *pelK* de *Bacillus subtilis*. En plus de spots « tampons de dépôt », des témoins négatifs ont été déterminés en utilisant des gènes de *Buchnera* auxquels cinq mutations ont été ajoutées ainsi que des séquences aléatoires ne présentant pas d'affinité avec les gènes de *Buchnera*. Chaque sonde est déposée sous forme de doublet sur la puce et chaque doublet est déposé par deux aiguilles différentes. Pour chaque gène, on trouve donc de quatre à 12 mesures d'expression selon que le gène est représenté par une ou trois sondes différentes. La puce complète présente 6144 spots qui ne sont pas répartis aléatoirement sur la lame, mais dans l'ordre chromosomique (ordre des numéros d'accession BU). Par contre, comme deux à trois sondes sont disponibles pour chaque gène, l'effet gène mesuré après le calcul des moyennes est issu de régions géographiques éloignées sur la puce.

La solution de dépôt (200 µM d'oligonucléotides) est constituée de tampon SSC 3x (Sigma-Aldrich) contenant 1.5 M de bétaine (*N,N,N*-trimethyl-glycine). Les lames utilisées sont des lames de verre aldéhydes QMT (Interchim).

2.4. Hybridations, plans d'expériences et acquisition des données

L'expérience YF comporte huit lames selon un plan en « dye swap » (tableau 3). L'expérience AAS correspond à un double plan en boucle de deux fois huit lames (figure 10).

L'efficacité de la réaction de marquage est estimée par mesure au spectrophotomètre à 550 nm (Cy3) ou à 650 nm (Cy5). Une quantité égale de cible marquée (basée sur la concentration en fluorochrome) est appliquée sur chaque lame. Les hybridations sont réalisées soit manuellement sous lame de verre (50°C, douze heures), soit automatiquement dans un appareil à hybridation (Ventana Medical Systems) à 45°C pendant huit heures. Après hybridation, les lames sont lavées dans des solutions de stringences croissantes. Les lames sont finalement séchées par centrifugation puis analysées à l'aide d'un scanner GeneTAC LSIV

(Genomic Solutions). Les intensités de fluorescence pour chaque spot sont ensuite estimées à l'aide du logiciel GenePix 4.0 (Axon Instruments).

2.5. Analyse qualité et normalisation

Pour chaque lame, une analyse qualité est réalisée et un indice est associé à chaque point à l'aide du logiciel GenePix. Pour l'expérience YF (lames de moins bonne qualité), le bruit de fond a été retranché, contrairement à l'expérience AAS où le signal brut a été utilisé. Les niveaux de fluorescence sont alors normalisés en utilisant un jeu de gènes invariants déterminés *a posteriori* sur chaque lame et par une approximation loess globale prenant en compte la non linéarité de la relation Rouge - Vert, mais pas la différence entre bloc d'aiguilles de dépôt. Pour plus de détails sur ces différentes étapes et méthodes, on se réfèrera au chapitre 3 de ce mémoire.

2.6. Analyses statistiques

Le principe des analyses statistiques utilisées a été présenté d'un point de vue théorique dans le chapitre 3. Les applications de ces méthodes à nos deux expériences sont décrites brièvement ci-dessous.

Détection des gènes différentiels

L'expérience YF compare deux conditions expérimentales. Des tests de Student modifiés selon la méthode SAM (Tusher *et al.* 2001) ont été réalisés sur les données normalisées, ainsi que des anova (Wu *et al.* 2003). Dans ces analyses, les répétitions techniques ont été moyennées, et le facteur aléatoire « lame » a été considéré comme fixe. Dans la deuxième série d'expériences (AAS), un modèle d'analyse de la variance a été ajusté, intégrant les effets « lame », « fluorochrome » et « traitement ». Ce dernier est constitué par deux facteurs croisés (acides aminés essentiels et saccharose). Les p-values sont estimées par permutation et en utilisant une variance contrainte selon la méthode de James Stein (cf. chapitre 3 § 5.2.1). Pour les deux expériences, un gène est considéré comme significatif s'il remplit les trois critères suivants : (1) une p-value inférieure à 5 %, (2) un rapport d'expression extérieur à l'intervalle (0,83 – 1,2), et (3) un bon indice qualité.

Une deuxième analyse a parfois été réalisée au niveau d'un bloc de gènes (voie métabolique, région chromosomique, opéron ou régulon). Pour l'ensemble des gènes du bloc, la valeur du logarithme du rapport d'expression différentielle est relevée et un test des rangs signés de Wilcoxon est réalisé sur la série. Ce test permet de savoir

si une voie est régulée globalement, même si aucun de ces gènes n'est significatif individuellement.

Contrôle du nombre de faux positifs (FDR)

L'estimation du nombre de faux positifs a été réalisée grâce à une démarche de simulation à partir des lames témoins de notre jeu de données. La procédure suivie est décrite dans le § 2.3 du chapitre 3. A l'issue de l'analyse, nous avons montré que les taux de faux positifs sont respectivement de 25 et 5 % pour les expériences YF et AAS.

2.7. Structure du génome et évolution moléculaire

Les opérons potentiels de *Buchnera* ont été déterminés manuellement et par comparaison avec le génome d'*E. coli.* Pour cela, chaque orthologue de *Buchnera* chez *E. coli* a été recherché dans la base de données ECHOBASE (http://www.biolws1.york.ac.uk/echobase/). Puis, grâce au lien « Transcription Unit » pointant vers la partie régulation de la base Ecocyc (http://biocyc.org/), les orthologues de *Buchnera,* composants des « unités de transcription » fonctionnelles chez *E. coli,* ont été rassemblés. Ces unités de transcription correspondent dans notre analyse aux opérons, car toutes les unités ne contenant qu'un seul gène ont été éliminées de notre jeu de données. Ces opérons sont théoriques chez *Buchnera* puisque la présence des séquences régulatrices n'a pas pu être recherchée systématiquement et que beaucoup de régulateurs transcriptionnels sont absents chez la bactérie. Néanmoins, Moreno-Hagelsieb *et al*. (2001) montrent la pertinence d'une telle approche. Nous avons ainsi déterminé 81 opérons chez *Buchnera* impliquant 291 gènes différents. La taille des opérons de *Buchnera* est comprise entre 2 et 13 gènes.

Les syntons sont des fragments de gènes consécutifs conservés chez d'autres bactéries phylogénétiquement proches. Ils ont été calculés sur le site du TIGR avec l'outil « operon predictions » (http://www.tigr.org/tigr-scripts/operons/). La plupart du temps les syntons, par définition, englobent les opérons, ce sont des structures supérieures ou égales en taille aux opérons. Après élimination des singletons, nous avons déterminé 97 syntons de tailles comprises entre 2 et 36 gènes et intégrants 494 gènes au total.

Les ka (taux de substitution non synonyme) ont été estimés par Tamas *et al.* (2002) en utilisant le triplet d'orthologue *Buchnera*, *E. coli* et *Salmonella*.

2.8. RT-PCR en temps réel

La RT-PCR en temps réel (couramment appelée PCR quantitative) a été utilisée pour confirmer les résultats d'expression de l'expérience YF, caractérisée par un taux de faux positifs important. Douze gènes ont été choisis, soit parce qu'ils montraient des rapports d'expression différentielle inattendus, très forts ou encore non significatifs. Trois gènes de contrôle ont été ajoutés dans l'expérience : *rpsF* et r*plO* et *atpA*. Ces gènes montrent un niveau de transcription constant dans les différentes conditions testées. Les amorces utilisées pour l'amplification ont été conçues pour amplifier de courtes portions de chacun des gènes choisis (90 à 197 pb). L'optimisation de la séquence des amorces a été réalisée à l'aide du logiciel oligo6.6. Les réactions de RT-PCR en temps réel ont été conduites avec un Lightcycler Roche et les réactifs du Kit Light Cycler Fast Start DNA Master SYBR Green I, en suivant le protocole suggéré par le fabricant. La PCR débute par une première étape de dénaturation à 95°C pendant 10 minutes, suivie de 45 cycles : 10s à 95°C, 4s à 47 ou 51°C et 8s à 72°C. Une courbe de calibration a été construite pour chaque couple d'amorces, en utilisant des dilutions effectuées à partir de produits de PCR purifiés, obtenus par amplification de l'ADN génomique de *Buchnera*. Le nombre de copies d'ARN pour chaque gène a été calculé sur la base de la courbe de calibration ainsi construite. La même extraction d'ARN a été utilisée pour analyser les gènes d'intérêt et les gènes de contrôle dans un même essai. De cette façon, des variations dans l'efficacité d'extraction affecteront au même degré les gènes à tester et les contrôles. Chaque essai a été réalisé sur trois extractions d'ARN indépendantes. La réaction de retrotranscription préalable à l'amplification a été réalisée en partant de 2 µg d'ARN avec le kit Superscript™ First-Strand Synthesis for RT-PCR d'Invitrogen. Chaque produit de RT a été contrôlé par PCR et stocké à -20°C. Les amorces et les résultats de cette expérience sont présentés dans le tableau A7 en annexes.

3. Résultats

3.1. Performances biologiques des pucerons sur les différents milieux

3.1.1. Expérience YF

L'analyse du budget est très simple dans cette expérience. Le taux d'accroissement relatif des pucerons est 0,309 g par g de puceron frais et par jour. Seuls les acides aminés tyrosine et phénylalanine sont déficitaires dans le milieu YF. On peut ainsi calculer, pour un accroissement protéique moyen de 56,2 µg sur la période d'étude, que 19,7 nM de phénylalanine et 4,7 nM de tyrosine sont nécessairement fournies par *Buchnera* au puceron, sous la forme de 24,4 nM de

phénylalanine. Rapellons que *Buchnera* n'est pas capable de synthétiser la tyrosine, l'interconversion entre la phénylalanine et la tyrosine est probablement réalisée dans le cytoplasme du bactériocyte.

Les performances biologiques des pucerons élevés sur le milieu complet AP_3 et sur le milieu carencé YF sont représentées sur le tableau 4.

Milieu	Survie larvaire (%)	Poids (mg)	Durée de développement (jour)	Fécondité par adulte	r_m
AP_3	100	1.57 ± 0,04	7.38 ± 0,08	18.1 ± 2.1	0,284
YF	98	1.22 ± 0,02	7.31 ± 0,10	16.0 ± 2.2	0,271
p-value	ns	<0,0001 [a]	ns (0,25 [b])	ns (0,49 [a])	-

Tableau 4. Performances biologiques du puceron *A. pisum* sur les milieux complets (AP_3) et carencés (YF). Valeurs moyennes (n=40) ± SE. [a] : test ANOVA ; [b] : courbe de survie de Kaplan-Meier (test de Logrank).

Ces résultats montrent clairement que *Buchnera* est capable de compenser la déficience en tyrosine et phénylalanine. La carence se manifeste tout de même par une diminution du poids des larves d'environ 20 %, mais ni la durée de développement, ni la fécondité ne sont altérées. Les courbes de fécondité ne sont pas affectées comme le montre la figure 17.

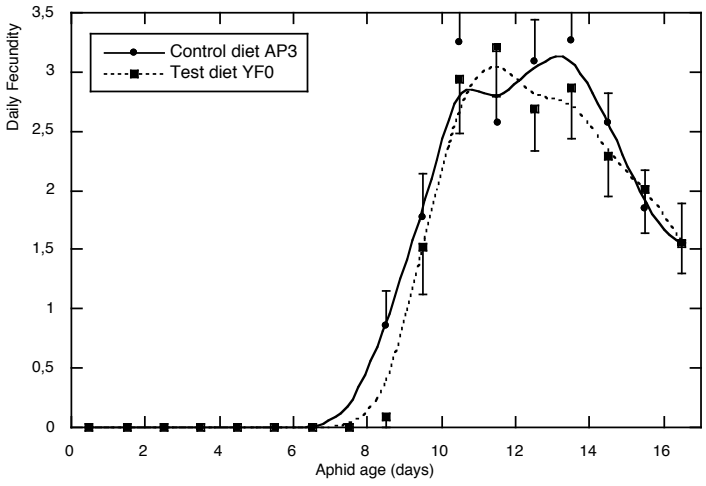

Figure 17. Comparaison des courbes de fécondité de *A. pisum* élevé sur milieu complet AP_3 ou sur milieu carencé YF. Les erreurs standard sont reportées sur le graphe (n=20)

3.1.2. Expérience AAS

L'analyse du budget de l'expérience AAS est la suivante. Les pucerons montrent une croissance plus importante sur les milieux à 0,5 M de saccharose relativement aux milieux contenant 1 M de saccharose. Cette inhibition de croissance s'explique par une baisse significative de l'ingestion des milieux riches en sucre (tableau 5, ci-dessous). Le saccharose est pourtant un phagostimulant important pour les pucerons. On suppose alors que l'effet inhibant du saccharose dans cette expérience est lié à l'hyper-osmolarité induite par la forte concentration de sucre. A l'inverse, le rapport entre acides aminés essentiels et non-essentiels n'a pas d'influence significative sur la croissance des pucerons, ni sur leur appétance.

Si on intègre la composition des milieux, les ingestions et les taux de croissance des pucerons, il est possible de déterminer quelles quantités d'acides aminés ont été fournies par *Buchnera* au cours de l'expérience. Le tableau 6 (ci-dessous) résume le budget en acides aminés essentiels de l'expérience AAS.

Milieu	**TCR** ($g\ g^{-1}\ jour^{-1}$)	**TCP** (µg)	**Ingestion** (µl)	**qPCR**
0,5 M sac. 50 % aa	0,317 ± 0,0162	39,7 ± 6,11	3,20 ± 0,255	2,11 ± 0,782 (7)
0,5 M sac. 25 % aa	0,308 ± 0,0209	30,8 ± 3,35	4,41 ± 0,391	1,94 ± 0,493 (10)
1,0 M sac. 50 % aa	0,233 ± 0,0165	21,3 ± 1,96	2,94 ± 0,222	1,19 ± 0,340 (9)
1,0 M sac. 25% aa	0,226 ± 0,0131	15,9 ± 1,50	1,70 ± 0,251	2,54 ± 0,673 (10)
ANOVA				
Saccharose	$F_{1,36} = 24,22$ *	$F_{1,36} = 20,34$ *	$F_{1,52} = 26,80$ *	$F_{1,32} = 0,08$ (NS)
Ac. aminés	$F_{1,36} = 0,21$ (NS)	$F_{1,36} = 3,71$ (NS)	$F_{1,52} = 0,01$ (NS)	$F_{1,32} = 1,03$ (NS)
Interaction	$F_{1,36} = 0,01$, (NS)	$F_{1,36} = 0,21$ (NS)	$F_{1,52} = 18,1$ *	$F_{1,32} = 1,70$ (NS)

Tableau 5. Performances biologiques des pucerons sur les quatre milieux de l'expérience AAS. Ce travail a été réalisé par A. Douglas (communication personnelle). TCR : taux de croissance relatif, TCP : taux de croissance protéique par puceron sur la durée de l'expérience, qPCR : estimation du nombre de copies ($x10^{-10}$) du gène *dnaK* copies par mg de puceron frais. Les valeurs sont exprimées sous la forme d'une moyenne ± erreur standard (nombre de répétitions). * : p-value < 0,05, NS : test non significatif.

acides aminés	*µg de Buchnera (% du total)*			
	0,5 M sac 50 % aa	**0,5 M sac 25 % aa**	**1 M sac 50 % aa**	**1 M sac 25 % aa**
His	0	0	0	0
Ile	0	0,040 (1,6 %)	0	0,079 (7,5 %)
Lys	0	1,588 (39 %)	0	0,713 (42 %)
Leu	0	0,956 (25 %)	0	0,461 (30 %)
Met	0	0,132 (12 %)	0	0,078 (17 %)
Phe	0,225 (13 %)	1,190 (53 %)	0	0,514 (56 %)
Thr	0	0	0	0,010 (1 %)
Trp	0	0	0	0
Val	0	0,428 (16 %)	0	0,231 (21 %)

Tableau 6. Budget en acides aminés pour l'expérience AAS. Les quantités sont exprimées en µg et en pourcentage de la quantité totale fournie par *Buchnera* (travail réalisé par A. Douglas). Les valeurs nulles indiquent que l'acide aminé n'est pas limitant (il peut même être excédentaire).

Le milieu de référence (0,5 M saccharose, 50 % acides aminés essentiels) est déficient en phénylalanine. La demande en acides aminés est nulle sur le milieu correspondant avec 1 M de saccharose en raison de la diminution du taux de croissance des pucerons. Sur les milieux carencés en acides aminés, tous les acides aminés sont déficients sauf l'histidine et le tryptophane, ainsi que la thréonine sur le milieu à 0,5 M de saccharose.

3.2. Expression des gènes chez *Buchnera*

Les expériences de puces à ADN permettent d'obtenir pour chaque gène, d'une part une mesure absolue (A) du niveau moyen d'expression, et d'autre part, une mesure relative (M) représentant le différentiel d'expression entre deux conditions expérimentales. Dans ce travail, nous avons tout d'abord cherché à caractériser globalement la régulation transcriptionnelle de *Buchnera*.

La première question que nous nous sommes posée est très simple : « Est-ce que tous les gènes de *Buchnera* sont exprimés à un même niveau, et sinon peut-on caractériser les gènes les plus et les moins exprimés ? La figure 18 montre la corrélation entre le niveau moyen d'expression et le taux de substitution des gènes chez *Buchnera*.

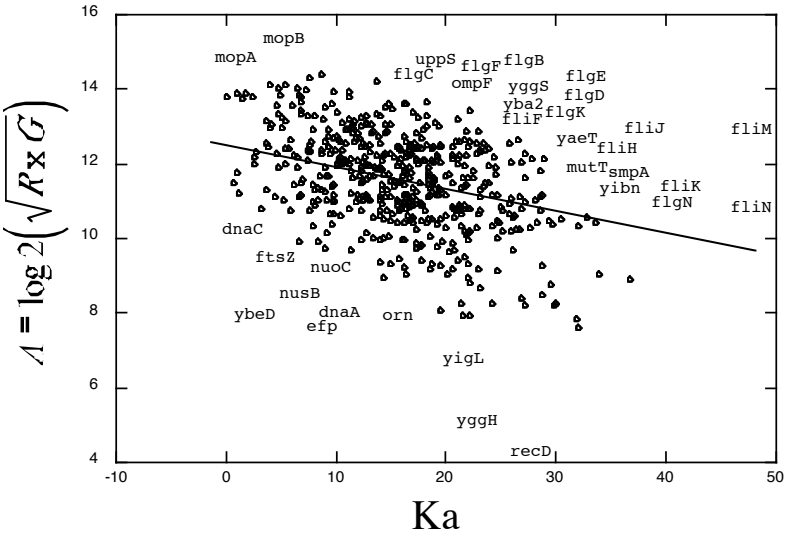

Figure 18. Corrélation entre niveaux moyens d'expression (*A*) des gènes et taux de substitution non-synonyme (Ka) chez *Buchnera* (r^2 = 0,1, p = 10^{-5}). L'estimation des Ka est issue de Tamas *et al.* (2002).

La figure 18 montre une corrélation entre conservation et expression. Les gènes les plus exprimés chez *Buchnera* sont les gènes les plus conservés. Dans la partie inférieure du graphe sont reportés les gènes montrant des niveaux d'expression très faibles (*recD*, *yggH* et *yigL*). A l'inverse, dans la partie supérieure du graphe, figurent les gènes montrant une forte vitesse d'évolution et une expression très forte. Ces gènes, fortement exprimés mais évoluant vite, sont de bons candidats à la sélection positive. C'est-à-dire qu'ils pourraient avoir évolué dans le contexte symbiotique pour changer de fonction. Parmi ceux-ci, on trouve les gènes, codant les chaperones *mopA* et *mopB,* pour lesquels une sélection positive a déjà été évoquée (Fares *et al.* 2002). Ces deux gènes sont, malgré tout, très conservés. De façon très claire, la majorité des gènes codant le flagelle sont repérés dans la partie haute et droite de la figure 18 (opérons *flg* et *fli*). Nous avons fait l'hypothèse que le flagelle chez *Buchnera*, non fonctionnel pour la motilité, pourrait servir de transporteur de polypeptides. Ce point sera repris dans la discussion de ce chapitre.

La deuxième question que nous nous sommes posée concerne la relation entre le niveau d'expression et la structure en unités de transcription chez *Buchnera*. Pour cela, nous avons déterminé les opérons et les syntons potentiels dans le génome de la bactérie (cf. § 2). Une corrélation entre les niveaux d'expression moyens (A) et relatifs (M) des gènes a été recherchée en fonction de ces structures potentielles de régulation (tableau 7).

Enfin la dernière question que nous nous sommes posée concerne l'existence d'une structure en régulon chez *Buchnera*. Nous n'avons pas réalisé une étude systématique de tous les régulons conservés chez *Buchnera*, mais les résultats préliminaires obtenus dans ce travail montrent que cette étude sera nécessaire dans des perspectives à court terme (un projet d'étudiants en 4ème année Bioinformatique INSA est actuellement en cours sur ce sujet). Notre travail a consisté à mesurer l'expression des gènes dont les régulateurs étaient induits dans nos expériences. Il est remarquable que les trois régulateurs que nous avons détectés dans nos expériences appartiennent au même groupe des protéines « histone-like ». Ce sont les régulateurs IHF (*Integration Host Factor*), activés par le stress d'acides aminés dans l'expérience AAS, H-NS (*Histone-like Nucleoid Structuring*) et FIS (*Factor for Inversion Stimulation*), tous deux inhibés par le stress en sucre dans l'expérience AAS. Ces trois régulons ont été plus particulièrement étudiés.

	Source	DF	SCE	R^2	$pval_A$ [1]	$pval_R$ [1]
A_{YF}	*inter UT*	80	391,93	0,63	10^{-16}	10^{-11}
	Intra UT	210	225,36			
	inter synton	96	426,96	0,43	10^{-16}	10^{-13}
	intra synton	393	544,32			
M_{YF}	*inter UT*	80	10,89	0,29	0,29	0,01
	intra UT	210	25,98			
	inter synton	96	17,63	0,26	0,07	0,001
	intra synton	393	49,13			
A_{AAS}	*inter UT*	80	$4{,}76\ 10^{12}$	0,54	10^{-11}	10^{-7}
	Intra UT	210	$4{,}08\ 10^{12}$			
	inter synton	96	$4{,}73\ 10^{12}$	0,41	10^{-12}	10^{-8}
	intra synton	393	$6{,}91\ 10^{12}$			
$Maa05_{AAS}$	*inter UT*	80	28,48	0,35	0,03	0,09
	intra UT	210	53,21			
	inter synton	96	55,46	0,33	10^{-6}	0,01
	intra synton	393	115,19			

Tableau 7. Analyse de la variabilité intra- et inter-unités de transcriptions (opérons et syntons) des niveaux d'expression absolus (A) et différentiels (M) pour les deux expériences YF et AAS. [(1)] $pval_A$ et $pval_R$ sont respectivement les p-values estimées par l'anova et par le test des rangs non paramétrique de Kruskal-Wallis. Les M représentent les log_2 des rapports d'expression pour l'expérience YF et pour le stress en acides aminés essentiels à 0,5 M de saccharose de l'expérience AAS.

Le régulon IHF

IHF régule l'expression de plus de 200 gènes chez *E. coli*. Pour 64 d'entre eux, présentés dans la base regulondb (Salgado *et al.* 2004), des sites de fixation du régulateur ont été déterminés expérimentalement. Sur ces 64 gènes, neuf sont

conservés chez *Buchnera*. Le tableau 8 résume les principaux résultats de cette recherche. Pour six des neuf gènes de *Buchnera*, le facteur IHF est répresseur. Seul le gène *fis* est activé par IHF. Les opérons *carAB* et le gène *cysG* sont régulés positivement ou négativement par IHF en fonction de la présence ou de l'absence d'autres régulateurs chez *E. coli*. Notons que dans les conditions où *himA* est significativement surexprimé (stress d'acides aminés à 0,5 et 1 M de saccharose), le gène *fis* est activé et que tous les gènes théoriquement réprimés le sont à l'exception de *himA* lui-même ; le régulateur n'est cependant auto-réprimé qu'à des taux très élevés d'IHF dans la cellule. Six des 13 gènes de l'opéron *nuo* montrent également une régulation contraire à celle d'*E. coli*. Pour ce dernier, il faut noter que *nuo* est régulé de façon antagoniste par le facteur FIS, significativement induit dans nos expériences.

opéron	**IHF (*E. coli*)**	**Sites IHF (*E. coli*)**	**Régul.**	**IHF (*Buc*)**	**IHF (*Buc*)**	**RE (*Buc*)**
himA	-47	TATCCGAATGTAA	R	-190	TATTTTTTATAAA	A(1)*
	-21	AATCAGGTAGTTG	R	-63	TAATGAAATTAAT	
	+54	CATAAGCCTGATC	R			
himD	-320	TACCTGCGTGCTT	R	-5	AATTTATAAATAT	R(1)
	-304	AAGCATCCCGTGA	R			
	-207	AATCAGCCTGTCT	R			
	-157	TCGCAACTGTTAA	R			
	-77	TAATTCTCTGACT	R			
	-2	AATCAAGCACTAA	R			
SucAD°	ND	ND	R	-230	TATTCATTTTTTT	R(2)
				-130	TATTGATATTTAT	
				-110	TAATTTTATTAAA	
				-40	TATTTTTTATTTA	
ompF	-176	AATCATATAAATA	R	-130	TATTTTTAATGAA	R(1)
	-108	AAAGTTCCTTAAA	A	-46	TATTTATTTTAAT	
	-62	AACCAAATCTTTA	R	-32	AACTGATATTAAT	
	ND	ND	R	-13	TATTTTTACTAAA	
sodA	-66	CAAAACCATACCC	R	-340	CATTTTTAATGTT	R(1)
	-46	AAGTACGGCATTG	R	-116	TATGTATTATGAT	
	-22	CAATATCATTTAA	R	-74	AATTTATAAAAAT	
	+2	TGAACCAACTGCT	R	-14	ATTTTATTATCAT	
NuoAN	-83	GAAACTTTTGTTA	R	-103	TATTTATTAAAAA	A(6)
	-74	GTTAAAGTTGACA	R	-73	TATTTAATAACAT	R(7)
	-66	TGACAAAAGGTTA	R	+15	TAATGTTAATAAA	
CysG°	-115	CAGACAATTCCGT	A/R	-122	TATTGTATATCAA	A
	-88	TATACAAATCAGC		-51	AATTGATAATTTT	
				-45	TAATTTTATTAAA	
carAB	-305	TTGATTATAAAGG	A/R	-132	AAATGTATTTTAT	A(1)-
				-85	AAATGTATATCAT	R(1)
fis°	-114	GATGATTTTTATG	A	-132	TATTGAATATATT	A(1)
				-122	TATTGAATATATT	
				-38	TAATTTTTTTTAA	

Tableau 8. Liste des opérons régulés par le facteur IHF chez *E. coli* et conservés chez *Buchnera*. Les sites de fixation ont été prédits à l'aide du logiciel partser (Hertz et Stormo 1999). * : (n) nombre de gènes activés ou réprimés dans l'opéron. ° : opérons partiellement dégradés chez *Buchnera.*

Nous avons ensuite cherché à savoir si ces gènes corégulés possèdaient des signaux de régulation IHF dans leur partie 5'-terminale. Une procédure de recherche de régions conservées a été lancée à l'aide des logiciels MEME et MAST (Bailey et Elkan 1995 ; Bailey et Gibskov 1998). Des sites potentiels ont été retrouvés pour tous les gènes sauf *himD* qui ne présente pas de probabilité significative. Néanmoins, l'analyse est très peu informative dans la mesure où la dégénérescence du site IHF, le biais en bases A et T des régions intergéniques et l'absence de données expérimentales sur la fixation du facteur chez *Buchnera* rendent ces prédictions très hypothétiques (données non présentées). Nous avons alors effectué la démarche inverse, c'est-à-dire qu'à partir de la matrice consensus du facteur IHF d'*E. coli*, nous avons détecté les sites potentiels de fixation en amont des neuf gènes de *Buchnera*. La figure 19 montre que, pour tous les gènes, des sites de fixation significatifs sont trouvés (ceux-ci sont également reportés dans le tableau 8). Bien sûr, ces deux analyses successives n'ont pas convergé vers les mêmes sites potentiels, et ne peuvent pas constituer une preuve formelle de l'existence de sites de régulation du facteur FIS chez *Buchnera*. Il existe en effet des cas où des organismes phylogénétiquement proches possèdent des sites de fixation très différents pour un même facteur de transcription (Rodionov *et al.* 2004).

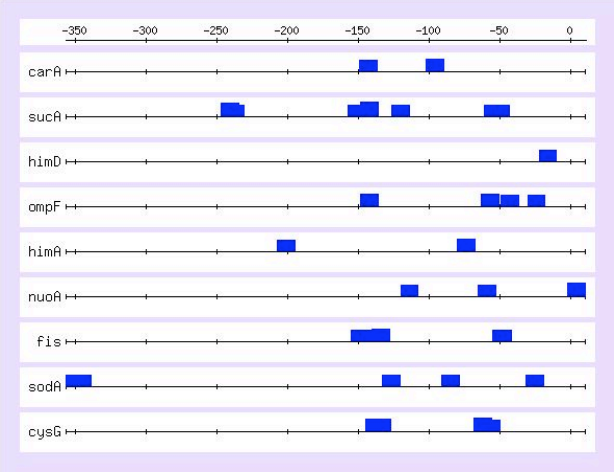

Figure 19. Localisation des sites potentiels significatifs du facteur de transcription IHF, détectés à l'aide du logiciel patser, en amont des gènes de *Buchnera*. La matrice utilisée est la matrice d'*E. coli*.

L'interprétation fonctionnelle de cette régulation est par contre assez difficile à réaliser. On comprend pourquoi, si IHF est activé par le stress en acides aminés, *ompF* et *carB* sont réprimés alors que *fis* est activé. Mais il est difficile de voir l'impact de cette régulation sur la réponse physiologique. OmpF est une protéine de

transport, *cysG* est impliqué dans la biosynthèse de la cystéine et *carB* serait impliqué dans la liaison entre le métabolisme des bases et des acides aminés. Les autres gènes ne semblent pas impliqués dans une réponse directe (ou directement compréhensible).

Le régulon FIS

La même étude que précédemment a été entreprise sur le facteur FIS. Parmi les 57 gènes de regulondb régulés par FIS chez *E. coli*, 16 sont conservés chez *Buchnera* (sans compter les ARN ribosomaux dont les opérons sont fortement dégradés chez *Buchnera*). Par contre, contrairement au facteur IHF, l'analyse des profils d'expression des gènes théoriquement sous le contrôle de FIS ne montre une cohérence vis-à-vis de *E. coli* que pour deux d'entre eux, *mtlA* et *gyrB* (données non présentées). Des sites conservés significatifs ont été trouvés pour tous les gènes. La recherche inverse des séquences potentielles de fixation du régulateur FIS basée sur la matrice d'*E. coli* a été réalisée et pour les 16 gènes de *Buchnera* des sites ont été trouvés (données non présentées).

Le gène *fis* est surexprimé par le stress d'acides aminés à 1 M de saccharose et dans une moindre mesure à 0,5 M. Il est par contre significativement inhibé par le stress en saccharose à 50 % d'acides aminés et invariant à 25 %. Ce gène est typiquement un gène d'interaction.

Le régulon H-NS

H-NS est un régulateur transcriptionnel important conservé chez *Buchnera* (Hommais *et al.* 2001). Cependant, l'analyse du régulon H-NS chez *E. coli* est relativement peu informative pour *Buchnera*, puisque sur les 35 gènes décrits dans la bases regulondb, un seul (*bolA*) est conservé chez la bactérie symbiotique. Le gène *hns* est fortement réprimé en moyenne dans notre analyse (sauf en stress saccharose à 25 % d'acides aminés où il est invariant). Le gène *bolA* semble réprimé par H-NS ce qui est compatible avec la régulation observée chez *E. coli*. Néanmoins, ce résultat n'est pas suffisant pour apporter la preuve d'une régulation significative par Hns chez *Buchnera*. On peut noter que *bolA* est également un régulateur transcriptionnel chez *E. coli* (Santos *et al.* 2002).

3.3. Régulation de l'expression des ARNt chez *Buchnera*

Durant son évolution, *Buchnera* a perdu de nombreux gènes devenus inutiles dans son environnement symbiotique intracellulaire. Ainsi, les mécanismes de réplication et de réparation de l'ADN sont en partie dégradés ; de même la machinerie de transcription est très altérée par la perte des principaux régulateurs et

promoteurs. Par contre, le système de traduction est fortement conservé chez *Buchnera*. Cette conservation semble une caractéristique générale des organismes vivants. Koonin (2003) a déterminé, par une approche de génomique comparative, le jeu de gènes minimal essentiel à la survie d'un organisme bactérien ; notons que, parmi les 60 protéines sélectionnées par l'analyse, figuraient essentiellement les protéines de la traduction. Les 55 protéines ribosomales, les 12 facteurs de traduction et les 21 ARNt aminoacyl-transférases trouvées chez *E. coli* sont tous conservés chez *Buchnera* (http://www.genome.jp/kegg). On peut noter également que les gènes codant les ARN ribosomaux (5S, 16S et 23S) sont présents en simple copie chez *Buchnera*, tandis que sept copies coexistent chez *E. coli*. La perte des paralogues ribosomaux est souvent interprétée comme une propriété des organismes à croissance lente (Lambert et Moran 1998 ; Sharp *et al.* 2005).

Au cours de l'expériences YF, nous avons noté une surexpression très forte des ARNt en condition de stress trophique (figure 20). Ce paragraphe analyse en détail le répertoire d'ARNt isoaccepteurs, la relation entre leur expression absolue et l'usage du code, le biais d'usage du code et enfin la régulation de l'expression des ARNt chez *Buchnera*.

3.3.1. Répertoire d'ARNt isoaccepteurs chez *Buchnera*

Le code génétique comporte 61 codons pour spécifier les 20 acides aminés. Cependant, il n'existe pas 61 ARNt spécifiques de chacun de ces codons. Un même ARNt reconnaît plusieurs codons, grâce à un appariement partiel du codon avec son anticodon. On parle alors d'ARNt isoaccepteur. *Buchnera* possède un jeu de 32 ARNt isoaccepteurs. Les codons complémentaires parfaits de ces 32 isoaccepteurs sont les codons gras et soulignés dans le tableau 9 (ci-dessous). Le jeu d'ARNt de *Buchnera* est minimaliste. En effet, seuls huit organismes, parmi ceux entièrement séquencés à ce jour, possèdent moins de 32 ARNt. *Mycoplasma mobile* est l'organisme possédant le moins d'ARNt, avec un jeu de 28 isoaccepteurs (Sharp *et al.* 2005). *M. mobile* est notamment le seul à ne posséder que deux codons méthionine, alors que tous les autres organismes en possèdent au moins trois.

La distribution des ARNt chez *Buchnera* semble optimisée pour utiliser au plus juste la règle standard de la dégénérescence de la paire codon-anticodon (tableau 9, ci-dessous). C'est-à-dire que pour la majorité des acides aminés, les codons XXpyrimidine et XXPurine sont reconnus par deux ARNt isoaccepteurs porteurs des anticodons GXX et UXX. Des réductions plus drastiques semblent avoir eu lieu dans le groupe des mollicutes (un seul ARNt-Ala ou deux ARNt-Arg chez *M. mobile*). Le chargement des acides aminés proline et isoleucine par un seul ARNt semble une caractéristique bien conservée chez les organismes à petit génome. Cette réduction implique des modifications de la troisième base de l'anticodon (Inosine, ou thio-uridine) pour une partie de la population des deux isoaccepteurs

correspondants. Notons que les acides aminés arginine et proline ne suivent pas la règle générale. Les ARNt correspondant aux codons CGC et CCC sont évités contrairement à ce qui est fait pour les autres acides aminés. Nous avons observé que cette situation est également retrouvée chez d'autres bactéries intracellulaires d'insectes : *Wolbachia pipientis*, *Wigglesworthia glossinidia brevipalpis* et *Blochmania floridanus*.

3.3.2. Usage du code chez *Buchnera*

La fréquence d'utilisation des différents codons synonymes est variable entre les organismes, mais également entre les différents gènes d'un même organisme. Parmi les facteurs susceptibles d'interagir avec l'usage du code chez les bactéries figurent le biais de substitution intra-brin d'ADN (Lobry 1996), le transfert horizontal de gènes entre des bactéries montrant des taux de GC différents (Ochman *et al.* 2000), l'évitement des séquences palindromiques (Fuglsang 2004) et enfin la sélection pour l'efficacité de la traduction. Cette dernière hypothèse a été formulée initialement chez *E. coli*. Ikemura (1981 ; 1985) a en effet montré que l'abondance des ARNt est corrélée à l'usage du code. En d'autres termes, cette règle (N°1) s'énonce de la façon suivante : les ARNt isoaacepteurs les plus exprimés sont ceux qui correspondent aux codons majoritaires de l'organisme. Cette règle a ensuite été vérifiée chez d'autres organismes (voir Sharp *et al*. 2005 pour une étude plus complète). Au contraire, chez les bactéries montrant des compositions en bases GC extrêmes, une absence de sélection pour l'usage du code est souvent observée (Wright et Bibb 1992 ; Andersson et Sharp 1996 ; Kerr *et al.* 1997). Chez *Buchnera*, plusieurs auteurs ont prétendu que l'usage du code a été effacé par le biais de substitution (Wernegreen et Moran 1999 ; Rispe *et al.* 2004). Plus récemment, Dittmar *et al*. (2004) ont mesuré l'abondance des ARNt chez *Bacillus subtilis* grâce à une puce à ADN, et ont montré l'absence de corrélation entre l'abondance des isoaccepteurs et l'usage du code chez cette bactérie.

Comme un ARNt isoaccepteur reconnaît plusieurs codons différents, la règle N°1 ne suffit pas pour expliquer complètement l'usage du code. Trois règles supplémentaires doivent être énoncées. La règle N°2 stipule que les anticodons possédant une base modifiée du type thio-uridine ou carboxyméthyl-uridine en troisième position préfèrent les codons finissant par adénine, relativement aux codons finissant par une guanine. La règle N°3 affirme qu'une base modifiée du type inosine en troisième position d'anticodon produira une préférence pour les codons finissant par U ou C relativement à A. Enfin, la règle N°4 s'applique aux acides aminés à deux codons et pour les codons du type (A/U)-(A/U)-pyrimidine, le choix d'un C plutôt que d'un U sera optimal. Cette dernière règle, appelée règle du codon « bon choix » a été décrite initialement par Grosjean *et al.* (1978) et Grantham *et al.* (1981).

	T				C				A				G			
	cod[b]	AA	RSCU[a]	N	cod	AA	RSCU	N	cod	AA	RSCU	N	cod	AA	RSCU	N
T	TTT	Phe	**1.83**	8662	TCT	Ser	**2.56**	5821	TAT	Tyr	**1.73**	5882	TGT	Cys	**1.64**	1874
	TTC		0.17	814	**TCC**		0.25	570	**TAC**		0.27	910	**TGC**[c]		0.36	409
	TTA	Leu	**3.87**	12029	**TCA**		**1.49**	3388	TAA	STOP	2.27	433	TGA	STOP	0.34	65
	TTG		0.62	1932	TCG		0.21	468	TAG		0.39	75	**TGG**	Trp	1.00	1701
C	CTT	Leu	0.73	2251	CCT	Pro	**1.89**	2678	CAT	His	**1.75**	3470	**CGT**	Arg	**2.00**	2384
	CTC		0.10	320	CCC		0.29	404	**CAC**		0.25	504	CGC		0.29	342
	CTA		0.56	1727	**CCA**		**1.52**	2144	**CAA**	Gln	**1.74**	5226	CGA	Arg	0.95	1128
	CTG		0.12	370	CCG		0.30	423	CAG		0.26	790	**CGG**		0.09	102
A	ATT	Ile	**1.65**	11901	ACT	Thr	**1.82**	3900	AAT	Asn	**1.72**	11599	AGT	Ser	**1.23**	2801
	ATC		0.25	1783	**ACC**		0.26	551	**AAC**		0.28	1864	**AGC**		0.26	596
	ATA	Ile	**1.11**	8013	**ACA**		**1.71**	3674	**AAA**	Lys	**1.84**	17059	**AGA**	Arg	**2.50**	2962
	ATG	Met	1.00	4023	ACG		0.21	453	AAG		0.16	1514	AGG		0.18	211
G	GTT	Val	**1.80**	4102	GCT	Ala	**1.76**	3688	GAT	Asp	**1.76**	7153	GGT	Gly	**1.67**	4244
	GTC		0.28	640	**GCC**		0.25	524	**GAC**		0.24	966	**GGC**		0.30	782
	GTA		**1.61**	3665	**GCA**		**1.72**	3618	**GAA**	Glu	**1.82**	9417	**GGA**		**1.78**	4529
	GTG		0.32	726	GCG		0.27	572	GAG		0.18	942	GGG		0.25	643

Tableau 9. Usage du code et RSCU chez *Buchnera*. a : RCSU (Relative Synonymous Codon Usage), les valeurs en gras sont celles des codons préférés chez *Buchnera* (RCSU > 1). b : les codons gras soulignés correspondent à la complémentarité parfaite avec les ARNt isoaccepteurs du génome de *Buchnera*. c : BU330, décrit comme pseudo-ARNt (GCA anticodon) par l'annotation automatique du TIGR-CMR.

	T				C				A				G			
	cod	AA	Exp[b]	Hi/Lo	cod	AA	Exp	HI/Lo	cod	AA	Exp	HI/Lo	cod	AA	Exp	HI/Lo
T	TTT	Phe	-	0.97	TCT	Ser	-	0.98	TAT	Tyr*	-	0.94	TGT	Cys	-	1.00
	TTC		10.73	**1.28**[a]	**TCC**		11.02	**1.25**	**TAC**		10.71	**1.48**	**TGC**[c]		11.87	1.00
	TTA	Leu	9.86	0.98	**TCA**		10.21	0.99	TAA	STOP	-	1.01	TGA	STOP	-	**1.62**
	TTG	*	-	**0.77**	TCG		-	**0.54**	TAG		-	**0.69**	**TGG**	Trp	10.18	1.00
C	CTT	Leu *	-	**1.23**	CCT	Pro	-	0.90	CAT	His*	-	0.93	**CGT**	Arg*	11.33	**1.38**
	CTC		10.55	**1.55**	CCC		-	1.16	**CAC**		13.04	**1.60**	CGC		-	**1.35**
	CTA		11.72	1.08	**CCA**		9.71	1.10	**CAA**	Gln	11.39	1.01	CGA	Arg*	-	**0.87**
	CTG		-	1.10	CCG		-	1.12	CAG		-	0.93	**CGG**		8.34	**0.50**
A	ATT	Ile*	-	1.06	ACT	Thr	-	1.03	AAT	Asn *	-	0.96	AGT	Ser	-	1.04
	ATC		12.66	1.04	**ACC**		10.80	1.15	**AAC**		12.32	**1.28**	**AGC**		7.56	**1.33**
	ATA	Ile*	-	**0.89**	**ACA**		10.23	1.01	**AAA**	Lys	10.69	1.00	**AGA**	Arg*	8.80	**0.87**
	ATG	Met	10.81	1.00	ACG		-	**0.69**	AAG		-	1.00	AGG		-	**0.33**
G	GTT	Val*	-	0.98	GCT	Ala	-	0.91	GAT	Asp	-	0.97	GGT	Gly	-	0.99
	GTC		9.85	1.00	**GCC**		10.70	**1.26**	**GAC**		13.04	**1.23**	**GGC**		10.29	1.19
	GTA		11.16	0.93	**GCA**		13.47	1.04	**GAA**	Glu	13.58	1.01	**GGA**		10.74	0.98
	GTG		-	**1.71**	GCG		-	1.04	GAG		-	0.95	GGG		-	0.96

Tableau 10. Expression des ARNt et rapports des RSCU entre les gènes fortement et faiblement exprimés. a : les rapports en gras correspondent aux valeurs les plus éloignées de 1 (* acide aminé montrant un test de chi-2 significatif à 5 %), b : l'expression est donnée en unité de fluorescence pour chaque isoaccepteur (log_2).

L'usage du code chez *Buchnera* est présenté grâce aux valeurs de RSCU (Relative Synonymous Codon Usage) dans le tableau 9 (ci-dessus). Le RSCU correspond à une mesure intra acide aminé de l'usage du code. Pour un acide aminé à quatre codons, la somme des RSCU associés à chaque codon est de quatre, et les valeurs individuelles sont proportionnelles aux fréquences des différents codons. De façon très évidente, il apparaît que les codons préférentiels chez *Buchnera* sont systématiquement les codons terminés par A ou T. L'usage du code est donc majoritairement dicté par le biais AT et bien souvent la règle du codon « bon choix » est violée chez *Buchnera*.

3.3.3. Expression des ARNt et abondance des isoaccepteurs

Nous avons mesuré l'expression des ARNt chez *Buchnera* grâce à notre puce à ADN. Nous avons montré que la mesure de l'expression des ARNt n'était possible que lorsque les lames sont hybridées manuellement (entre lames et lamelles). En effet, lorsque l'hybridation est automatique (station d'hybridation Ventana Discovery^{TM}), les ARNt apparaissent systématiquement colorés en rouge. Cet artéfact est vraisemblablement lié à une interaction entre les ADNc complémentaires des ARNt et l'huile utilisée au cours de la réaction d'hybridation. Cette huile (mineral oil, LCS) est fluorescente à la même longueur d'onde que le colorant Cy5. Les expériences d'hybridation manuelle et automatique montrant ce biais de coloration sont présentées sur la figure 20.

Nous avons cherché à savoir si cette coloration rouge pouvait être expliquée par le taux de GC des ARNt, leur structure secondaire ou encore leur faible taille, et enfin si ce biais était retrouvé sur d'autres types d'ARN messagers ou ribosomaux.

Mais aucun résultat significatif n'a pu être obtenu ; les autres gènes de *Buchnera*, riches en bases G et C, possédant des structures secondaires ou une petite taille, ne montrent pas de biais systématique. C'est finalement la spécificité de liaison des ARNt aux acides aminés qui est suspectée pour expliquer cet artéfact, néanmoins la caractérisation du problème n'a pour le moment pas été approfondie.

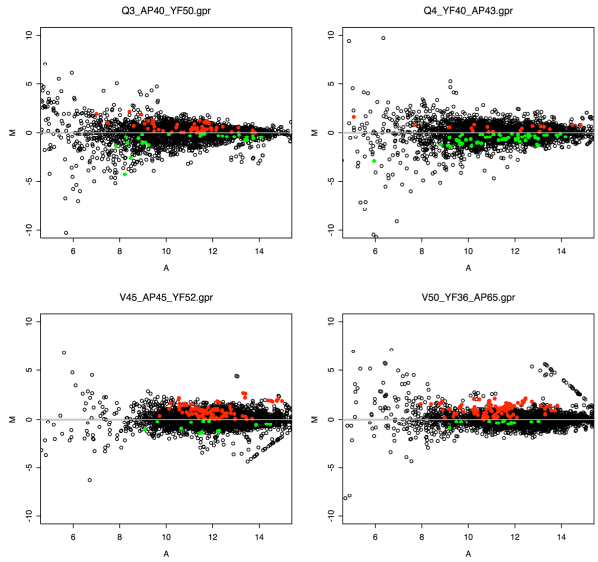

Figure 20. Graphes des $\log_2$ des rapports d'expression normalisés (M) en fonction de l'intensité du marquage (A) dans une expérience en « dye swap ». En rouge sont représentés les ARNt caractérisés par des valeurs positives, en vert sont pointées les valeurs négatives. Le « dye swap » est bien visible pour l'hybridation manuelle (en haut) alors qu'un biais systématique rouge est observé avec l'hybridation automatique (en bas).

Les hybridations manuelles réalisées dans l'expérience YF nous ont ainsi permis d'estimer les niveaux d'expression des différents isoaccepteurs en condition « normale » de croissance des pucerons (milieu AP_3). Les données sont présentées dans le tableau 10 (ci-dessus). Nous avons ensuite représenté l'abondance des ARNt en fonction de la composition en codons des 50 gènes les plus exprimés de *Buchnera* (Figure 21). Une corrélation de 0,32 a été trouvée entre la composition en codons et l'abondance des isoaccepteurs correspondants, elle n'est cependant pas significative ($p = 0,09$). Néanmoins, si on élimine du jeu de données les deux points extrêmes correspondant aux seuls anticodons entièrement composés de A et de T (TTA et AAA), la corrélation devient significative ($r = 0,52$, $p = 0,006$). Les mêmes résultats sont obtenus avec la composition moyenne de l'ensemble des gènes de *Buchnera* (données non présentées).

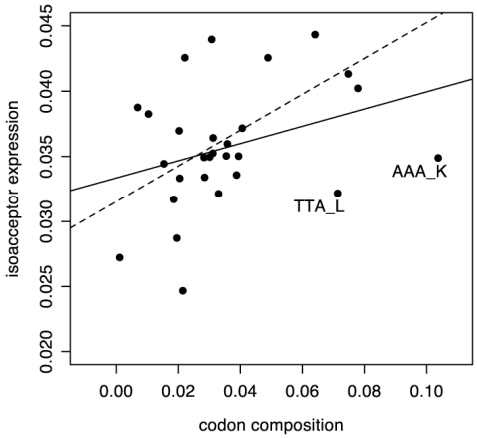

Figure 21. Niveau d'expression des isoaccepteurs (pourcentage de l'expression totale) en fonction de la composition en codons correspondant dans les 50 gènes les plus exprimés de *Buchnera*. La ligne pleine correspond à la régression linéaire intégrant tous les points et la ligne pointillée correspond à la régression sans les deux codons AT riches TTA et AAA.

Ce résultat montre que l'expression des ARNt chez *Buchnera* est ajustée à la fréquence des codons dans les gènes de *Buchnera*. L'expression des isoaccepteurs correspondant aux codons AAA et TTA, très largement sur-utilisés à cause du biais AT chez *Buchnera*, ne semble pas avoir été augmentée en conséquence par la bactérie.

3.3.4. Biais d'usage du code et règles d'appariement

Le tableau 10 (ci-dessus) permet la comparaison des RSCU entre les 50 gènes les plus fortement et les plus faiblement exprimés chez *Buchnera*. Cette différence est significative (test de Chi-2 sur les effectifs) pour la leucine, l'isoleucine, la valine, la tyrosine, l'histidine, l'asparagine et l'arginine. Pour tous ces acides aminés, sauf l'isoleucine, on note une utilisation plus fréquente par les gènes fortement exprimés des codons terminés en C (et souvent un rejet de la terminaison G). Il est remarquable que cette préférence soit systématique pour les 16 codons finissant par un C. Cette répartition est biaisée de façon très significative ($p = 10^{-5}$, test de la médiane). Cette préférence est partiellement expliquée par la règle N°4 du codon « bon choix » pour les acides aminés Phe, Tyr, Asn, Asp et His. Ces derniers codons, terminés par un C, ont été trouvés optimaux chez de nombreux organismes

(Kanaya *et al.* 1999). Le codon AUA-Ile (nécessitant une modification de l'ARNt) est rejeté dans les gènes les plus exprimés. La même observation est faite pour le codon CGA-Arg. Ce choix pourrait être expliqué par une modification de type inosine de l'isoaccepteur (règle N°3). Enfin, le rejet des codons terminés par un G (sauf pour la valine), chez les gènes fortement exprimés, pourrait impliquer la règle N°2, si les isoaccepteurs possèdent une modification de type thio-uridine. Cependant, la préférence de la terminaison A n'est pas véritablement observée. Notons enfin que les choix énoncés ci-dessus ne concernent que des codons rares.

La détermination du nombre de codons utiles (Nc), pour chaque gène de *Buchnera*, permet de mieux visualiser le biais d'usage du code révélé par le RSCU différentiel. Nc correspond à la moyenne du nombre de codons utilisés, normalisée par la taille et la composition en acides aminés du gène considéré (Peden 1999). Cet estimateur varie donc entre 20, si le gène utilise un seul codon par acide aminé, et 61, si le gène utilise aléatoirement tous les codons possibles. Une valeur théorique de Nc est estimable sous l'hypothèse que l'usage du code n'est soumis qu'à l'influence du biais de composition. Ainsi, la représentation de Nc en fonction du taux de GC en position 3 synonyme de codon permet d'évaluer le biais d'usage du code sous un nouvel angle (Figure 22).

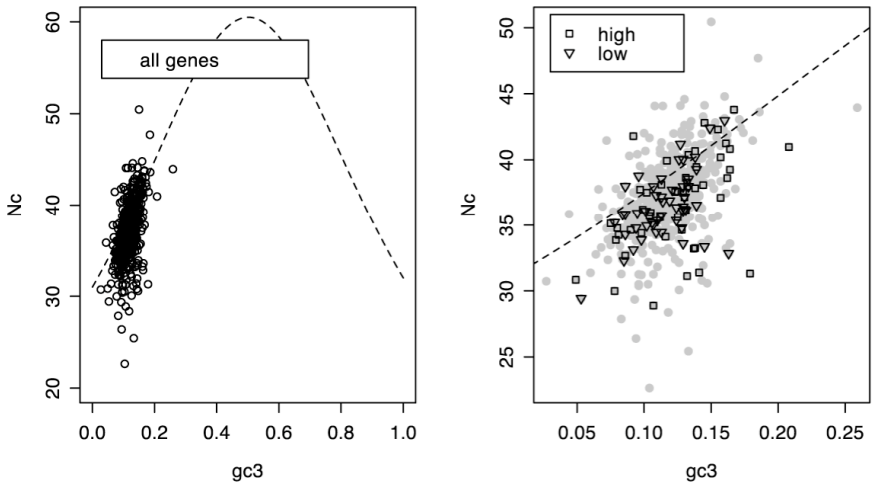

Figure 22. Estimation du nombre de codons utiles pour les gènes de *Buchnera*. La ligne en pointillés est l'estimation théorique de Nc pour chaque valeur du GC_3. A gauche, représentation du nuage de point des gènes de *Buchnera* sur toute la gamme des GC_3. A droite, grossissement du même nuage et visualisation des gènes fortement et faiblement exprimés. On remarquera une abondance un peu plus marquée des gènes fortement exprimés

(carrés) sur la gauche et dans la partie basse du nuage par rapport aux gènes faiblement exprimés (triangles).

La figure 22 (à gauche) montre tout d'abord que le biais AT est très fort chez *Buchnera* puisque le taux de GC3, compris entre 10 et 20 %, est extrêmement peu variable. Le nuage est bien réparti autour de la courbe théorique, ce qui indique que le biais d'usage du code, s'il existe, est très faible. Lorsque l'on s'intéresse plus précisément à la comparaison des gènes fortement et faiblement exprimés (figure 22, à droite), on remarque que les plus exprimés sont plutôt localisés en-dessous de la courbe (ils possèdent un usage du code particulier) et un peu plus sur la gauche (ils utilisent plutôt les codons avec G ou C en troisième position). Ce résultat semble donc confirmer le biais détecté dans le paragraphe précédent.

3.3.5. Usage du code et asymétrie de composition

L'asymétrie de composition intra-brin d'ADN est forte chez *Buchnera* (http://pbil.univ-lyon1.fr/software/Oriloc). Le brin avancé est plus riche en bases C et T, alors que le brin retardé est plus riche en bases A et G (Lobry 1996). Le biais d'usage du code observé chez *Buchnera* pourrait alors être expliqué par un biais de distribution des gènes fortement et faiblement exprimés sur les deux brins d'ADN. Néanmoins, la répartition des gènes n'est pas significativement biaisée dans notre analyse. Le brin avancé porte respectivement 59 % et 60 % des gènes fortement et faiblement exprimés.

Pour tenter, néanmoins, d'analyser l'influence de l'asymétrie sur l'usage du code chez *Buchnera*, nous avons séparé notre jeu de données de gènes selon les deux brins d'ADN, et nous avons réitéré les analyses de RSCU différentiels. Le choix préférentiel des codons rares terminés par un C a alors été observé de façon beaucoup plus marquée sur le brin retardé (riche en C et T), et de façon moins nette mais toujours significative sur le brin avancé. Le rejet des codons terminés par un G (sauf pour la valine) a également été observé sur les deux brins (données non présentées).

Ces résultats sont tout à fait contradictoires avec les travaux de Sharp *et al*. (2005) qui ont analysé la force de sélection pour l'usage du code chez 80 bactéries. Ces auteurs ont montré une préférence inattendue (car en désaccord avec les règles de codons optimaux) pour les codons terminés par un T chez les gènes fortement exprimés de *Buchnera*. Ils expliquaient ce résultat par un fort biais de distribution des gènes entre les deux brins d'ADN. Ces résultats sont le fruit d'une double erreur. D'une part, le jeu de gènes fortement exprimés était un jeu de gènes théoriques essentiellement composé de protéines ribosomales. Nous avons montré dans nos analyses que les protéines ribosomales ne sont ni les seules ni toutes fortement exprimées. Nous avons, d'autre part, montré qu'il n'existe pas de dissymétrie de

distribution de la localisation des gènes fortement exprimés sur les deux brins d'ADN. Ce résultat est en accord avec les travaux de Rocha et Danchin (2003 a et b) qui montrent que la localisation des gènes sur le brin avancé est influencée par le caractère essentiel, et non l'expressivité des gènes. Les protéines ribosomales, protéines essentielles, sont majoritairement localisées sur le brin avancé chez *Buchnera*.

3.3.6. Régulation de l'expression des ARNt

Chez *E. coli*, l'expression des ARNt est régulée de façon complexe. Emilsson *et al*. (1993) montrent que l'expression des isoaccepteurs majoritaires augmente tandis que celle des isoaccepteurs minoritaires diminue lorsque le taux de croissance de la bactérie augmente. Gourse *et al*. (1986) montrent que l'expression des opérons ribosomaux et des ARNt est réprimée par les ribosomes libres. Cependant, le mécanisme de régulation n'est pas clairement élucidé. Il pourrait être associé à la transcription, à la maturation et/ou à la dégradation des précurseurs, ou encore à la dégradation des ARNt mûrs. Enfin, Dittmar *et al.* (2004) ont analysé la régulation des ARNt chez *Bacillus subtillis* grâce à une puce à ADN spécifique. Ces auteurs ont montré que les niveaux d'ARNt étaient plus élevés en condition de stress nutritionnel, et associés à un fort taux de croissance de la bactérie.

Chez *Buchnera*, les ARNt sont organisés en 18 régions contiguës (deux d'entre elles sont associées aux gènes ribosomaux rrs et rrf-rrl). A titre de comparaison, il existe près de 87 gènes d'ARNt chez *E. coli* répartis en 40 unités de transcription dont sept contiennent les ARN ribosomaux. Chez *E. coli*, les régulateurs transcriptionnels généraux FIS et H-NS ont un rôle primordial dans la régulation de l'expression des ARN ribosomaux et des ARNt. Ces régulateurs sont conservés chez *Buchnera*, mais leur rôle n'a pas été démontré.

Grâce aux deux lames hybridées manuellement dans l'expérience YF, nous avons montré que la majorité des ARNt est surexprimée après une déplétion nutritionnelle en acides aminés aromatiques dans l'alimentation du puceron. En effet, 26 des 32 ARNt de *Buchnera* montrent des rapports d'expression positifs. Les résultats sont présentés dans le tableau A.3 en annexe. Le test de Wilcoxon donne une probabilité de 10^{-5} d'une distribution aléatoire des rapports autour de la valeur 1. Si les rapports d'expression extérieurs à l'intervalle (0,8 – 1,2) sont éliminés de l'analyse, la probabilité reste très significative (10^{-3}).

L'analyse complète de l'expérience YF (§ 3.4, ci-dessous) montre que cinq autres gènes associés au métabolisme des ARNt sont détectés comme significativement activés par le stress nutritionnel : *glnS* (Glutaminyl-tRNA synthase), *gltX* (Glutamyl-tRNA synthase), *glyS* (Glycyl-tRNA synthase β), *tdhF* et

trmD (deux gènes impliqués dans les modifications des tRNA). Au contraire, le gène *pheT* (Phénylalanine-tRNA synthase β) est réprimé.

Les ARNt sont répartis chez *Buchnera* dans 18 régions chromosomiques (UTP, unité de transcription potentielle). Ces régions sont visualisées par les lignes pointillées du tableau A.3. Bien que le test anova soit non significatif, la variance des rapports d'expression (M) inter-UTP est toujours supérieure à la variance intra UTP. Pour gagner en puissance, les valeurs de M ont été discrétisées selon la dernière colonne du tableau A.3. Un test de chi-2 a alors été réalisé, en estimant la probabilité par permutation pour s'affranchir des problèmes de classes à effectif faible. Une dépendance significative entre les deux variables qualitatives (contigs et expression) a ainsi été trouvée ($p = 0,05$). Ce résultat signifie que la régulation des ARNt est liée à l'organisation chromosomique, c'est-à-dire que les gènes situés côte à côte sur le génome sont globalement corégulés.

3.4. Analyse fonctionnelle

L'analyse du transcriptome s'incrit dans le cadre des analyses systémiques (ou de biologie des systèmes) puisqu'elle propose une vision globale d'un niveau d'organisation cellulaire (l'ensemble de la population des ARN messagers). Pour tenter de comprendre dans sa globalité la réponse fonctionnelle de la bactérie, il est nécessaire de décrire intégralement les résultats obtenus. Je n'ai donc pas cherché à résumer cette partie qui constitue pour notre équipe une description de nos expériences que le format des publications traditionnelles ne permet pas. Dans ce qui suit, les gènes différentiels ont été regroupés dans sept classes fonctionnelles arbitrairement définies par rapport au contexte de nos recherches : (1) Métabolisme des acides aminés et des vitamines, (2) Métabolisme des sucres et du peptidoglycane, (3) Transport, système flagellaire, (4) Régulation de la transcription, de la traduction et métabolisme de l'ADN, (5) métabolisme des lipides, (6) Repliement des protéines et réponse au stress, et (7) Divers et protéines hypothétiques.

3.4.1. Analyse fonctionnelle de l'expérience YF

La liste des gènes différentiels de l'expérience YF est donnée dans le tableau A.4 en annexe. L'examen de ce tableau montre que 22 gènes sont significativement surexprimés, tandis que 31 sont réprimés chez *Buchnera* lorsque le puceron est nourri sur un milieu sans tyrosine ni phénylalanine. Dix gènes montrent des rapports d'expression différentiels inférieurs à 0,5 et huit supérieurs à deux.

Comme le taux de faux positifs est important dans cette expérience, nous avons vérifié les valeurs d'expression différentielle pour un certain nombre de gènes

(tableau A7 en annexes). Parmi les 12 gènes testés, 10 montrent des rapports concordant entre les deux techniques. Les gènes *hisH* et *aroH* montrent des rapports d'expression contradictoires. Nous verrons par la suite que les corrections apportées par la PCR quantitative sur ces deux gènes apportent de la cohérence dans l'analyse fonctionnelle.

Métabolisme des acides aminés et des vitamines

Au total, dix gènes significatifs individuellement sont impliqués dans le métabolisme des acides aminés. La biosynthèse du tryptophane, de l'histidine et de l'arginine semble affectée par le stress. On peut remarquer que les gènes codant les enzymes de la branche finale de la biosynthèse de la phénylalanine ne sont pas induits.

Dans les voies de biosynthèse des acides aminés aromatiques. Le gène *aroH* est induit (mesure par PCR quantitative), par contre dans la branche terminale de biosynthèse du tryptophane, le gène *trpE* est significativement réprimé ; les gènes *trpA*, *trpB* et *trpG* sont également réprimés, tandis que *trpC* est invariant et *trpD* est légèrement surexprimé. La voie semble globalement réprimée même si le test global sur la voie n'est pas significatif (test des rangs signés de Wilcoxon, $p_w = 0,07$).

Dans la voie de l'histidine, cinq des huit gènes de l'opéron *his* sont réprimés par la déplétion (*hisC* et *hisG* sont significatifs). Le gène *hisG* a été décrit chez d'autres bactéries comme codant une enzyme rétroinhibée par le substrat dans la biosynthèse de l'histidine et pourrait donc être un gène clé de cette voie (Hoppe *et al.* 1979). La mesure du rapport d'expression différentielle par PCR quantitaive révèle une répression du gène *hisH* (contrairement à la mesure par puce estimant une surexpression forte). Après élimination du gène *hisH*, la voie est globalement réprimée ($p_w = 0,01$).

Enfin, le gène *carB* est réprimé par la déplétion en YF. Le complexe protéique CarAB est impliqué dans la biosynthèse de l'arginine ; c'est également un régulateur important du flux d'azote entre la biosynthèse des bases pyrimidiques et celles des acides aminés via la glutamine. Il est à noter que le gène codant la sous-unité CarA est également réprimé dans l'expérience.

Dans les voies de biosynthèse des vitamines et des groupes prosthétiques, on trouve les gènes *bioA* (biosynthèse de la biotine), *gshA* (biosynthèse du glutathion), *ispA* (métabolisme des isoprénoïdes), *hemD* et *hemK* (métabolisme de la porphyrine) qui sont réprimés par le stress en acides aminés YF. Seul le gène *ribE*, codant la chaîne α de la riboflavine synthétase, est activé par la déplétion YF.

Métabolisme des sucres et du peptidoglycane

Aucun gène des voies de la glycolyse ou des pentoses phosphates n'est régulé significativement par la déplétion en acides aminés. Par contre, dans les voies associées à la biosynthèse du peptidoglycane, trois gènes sont significativement activés : *amiB* (acétylmuramoyl-L-alanine amidase), *murC* (UDP-N-acétylmuramate-L-alanine ligase) et *uppP* (undecaprenyl-diphosphatase). Globalement, ni la voie métabolique, ni l'opéron *mur* ne sont significativement réprimés.

Transport, système flagellaire

Deux gènes annotés comme transporteurs sont retrouvés dans notre analyse différentielle. Le gène *cyoB* (transporteur d'électrons attaché à la membrane interne) est significativement réprimé, alors que *ptsH* (transporteur PTS de glucose) est activé. L'autre transporteur PTS de *Buchnera* codé par le gène *ptsG* est également surexprimé (non significatif).

Le gène *flgA*, codant une protéine du flagelle, est significativement surexprimé. Globalement l'opéron *flg*, codant la partie du flagelle associée à la membrane externe, semble surexprimé (mais le test n'est pas significatif). A l'inverse, l'opéron *fli* codant la partie basale semble plutôt réprimé, mais le test n'est pas significatif non plus ($p_w = 0,4$).

Régulation de la transcription, de la traduction et métabolisme de l'ADN

Le métabolisme des ARNt est activé par le stress nutritionnel. Vingt-six des 32 ARNt sont surexprimés ($p_w = 10^{-5}$), ainsi que de nombreux gènes de la biosynthèse des ARNt. Ce point a été évoqué dans le paragraphe précédent.

Le gène *ligA*, codant la protéine Dnlj essentielle pour la réplication, est surexprimé en condition de stress, comme les gènes *greA* (régulateur de l'élongation) et *nusG* (anti-terminaison). Inversement, le gène *apaH* (métabolisme des nucléosides) est réprimé par le stress en acides aminés. Le gène *rep1*, localisé sur le plasmide leucine, est surexprimé dans notre expérience. Ce gène avait été initialement annoté comme codant la protéine RepA1 associée à la réplication. Il est maintenant annoté comme erreur de chaîne de lecture dans la base de données. Nos expériences suggèrent que l'annotation initiale était correcte et que le gène est vraisemblablement fonctionnel dans la mesure où son niveau d'expression est fort ($155^{ème}$ rang d'expression moyen sur 617). Le gène *rep2*, codant une autre protéine de réplication, probablement organisé en opéron avec *rep1*, est également surexprimé.

Repliement des protéines et réponse au stress

Parmi les gènes surexprimés par la déplétion en acides aminés YF, on trouve les gènes *dnaK* (codant la protéine de choc thermique Hsp70 impliquée dans le repliement et le transport des protéines), *ibpA* (codant la protéine de choc thermique de faible masse moléculaire) et *surA* (codant un chaperon moléculaire). Deux gènes sont significativement réprimés dans l'expérience : *ppiD* (repliement des protéines) et *gcp* (catabolisme des protéines).

Métabolisme lipidique

Les gènes *lpdA* (dihydrolipoyl dehydrogenase) et *fabI* (énoyl-ACP reductase) sont réprimés par la déplétion YF. Ils interviennent dans le métabolisme lipidique, une fonction très altérée chez *Buchnera*.

Divers

Le gène codant la protéine inconnue Yba4 est activé par la déplétion en acides aminés YF.

3.4.2. Analyse fonctionnelle de l'expérience AAS

L'expérience AAS est plus complexe à analyser dans la mesure où sont mesurés à la fois des effets principaux (stress osmotique et stress d'acides aminés) et des effets d'interaction. A partir des quatre conditions expérimentales de l'expérience AAS (concentration en acides aminés essentiels normale et faible, concentration en sucre normale et élevée), il est possible d'estimer quatre rapports d'expression : les deux rapports différentiels induits par le stress en acides aminés à 0,5 et 1 M de saccharose, et les deux rapports différentiels induits par le stress en saccharose pour les conditions 50 et 25% d'acides aminés essentiels. L'ensemble des rapports d'expression varie entre 0,1 et 63 (avec environ 4 % de valeurs comprises entre 0,5 et 2). Cependant, les rapports d'expression des gènes différentiels sélectionnés montrent un écart beaucoup plus réduit (de 0,4 à 1,8).

3.4.2.1. Effet du stress en acides aminés

La liste des gènes différentiels après un stress en acides aminés essentiels et pour les deux conditions de pression osmotique, est donnée dans le tableau A.5 en annexe. Il est possible de représenter ces résultats sous une forme plus synthétique, et intégrant déjà une première information fonctionnelle. Ainsi, la figure 23 représente des ensembles de gènes différentiels avec les intersections contenant des gènes significatifs pour les effets principaux ou pour l'interaction. Les gènes localisés dans des ensembles uniques sont des gènes significatifs pour les contrastes correspondants.

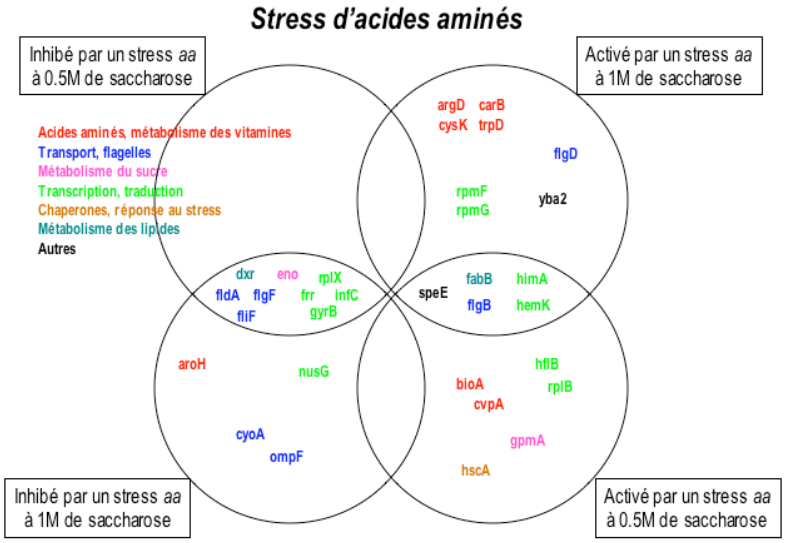

Figure 23. Diagramme de Venn correspondant aux gènes différentiels en condition de stress d'acides aminés.

Au total, 32 gènes ont été détectés comme différentiels dans cette expérience. Parmi ceux-ci 13 sont réprimés et 19 sont activés par le stress en acides aminés. Aucun gène n'est réprimé pour la seule condition 0,5 M de saccharose. Les principales catégories fonctionnelles concernées par la réponse de *Buchnera* sont détaillées ci-dessous.

Métabolisme des acides aminés et des vitamines

Dans la voie de biosynthèse des acides aminés aromatiques (tryptophane, phénylalanine, tyrosine), le gène *aroH*, codant l'aldolase permettant l'entrée du phosphoénolpyruvate (PEP) dans la voie, est réprimé à 1 M de saccharose. A l'inverse, dans la branche finale de la biosynthèse du tryptophane (à partir du chorismate), tous les gènes sont surexprimés, à la fois ceux localisés sur le plasmide (*trpE* et *G*) et les quatre gènes de l'opéron *trpABCD* situé sur le chromosome à 1 M de saccharose ($p_w = 0,01$). Dans les deux autres branches aboutissant à la tyrosine et à la phénylalanine, les gènes sont invariants dans les deux conditions de saccharose.

Dans la voie de biosynthèse de la leucine, aucun gène n'est significatif individuellement, mais les sept gènes du plasmide Leu montrent des rapports d'expression inférieurs à un. Cette tendance significative à 1 M ($p_w = 0,01$) est encore observée à 0,5 M de saccharose, mais n'est plus significative ($p_w = 0,18$).

Dans la voie de biosynthèse de l'histidine, sept des huit gènes de l'opéron *his* sont réprimés à 0,5M (p_w = 0,02), la tendance est non significative à 1 M, mais cinq gènes sur les huit sont encore réprimés (p_w = 0,07). Aucun gène n'est significatif individuellement dans les deux conditions de saccharose.

Dans la voie de biosynthèse de l'arginine, les huit gènes sont surexprimés à 0,5 M de saccharose (p_w = 0,02). A 1 M de saccharose, la tendance est conservée avec six gènes surexprimés (*argD*, *argF* et quatre gènes sur les cinq de l'opéron *argECBGH*), mais le test est non significatif (p_w = 0,07). On peut noter que le gène *argS*, codant l'arginine ARNt synthase, est réprimé (non significativement) à 1 M de saccharose. Ces effets cumulés semblent permettre une augmentation du taux d'arginine libre chez *Buchnera* à 1 M de saccharose. Le gène *carB* est surexprimé à 1 M de saccharose, mais *carA* ne montre pas un profil d'expression différentiel similaire.

Dans la voie de biosynthèse de la cystéine, le gène *cysK* est significativement activé à 1 M de sucrose, les six autres gènes de la voie sont invariants.

Le gène *cvpa* est surexprimé significativement à 0,5 M de saccharose. Ce gène orthologue de la protéine membranaire CvpA chez *E. coli* a été décrit comme un pseudogène chez *Buchnera* à cause d'un décalage de chaîne de lecture en position 481. Un transcrit est cependant détecté dans nos analyses avec un niveau d'expression moyen (rang 284 sur 585). Le peptide codé par le gène *cvpa* est exceptionnellement riche en phénylalanine (20 % de sa composition), alors que les autres peptides de *Buchnera* sont composés de 2 à 10% de cet acide aminé. L'augmentation de la composition en codon phénylalanine (TTT et TTC) dans ce gène peut s'expliquer par une accumulation de bases T suite à la perte de fonction du gène. Néanmoins, le niveau fort d'expression de ce gène et sa régulation laissent penser que ce peptide pourrait avoir un rôle dans le transport vers l'hôte de la phénylalanine. La localisation cytoplasmique de la protéine CvpA, ainsi que le transporteur membranaire restent à trouver pour valider cette hypothèse.

Le gène *bioA* est significativement surexprimé à 0,5 M de saccharose, les deux autres gènes de la voie *bioB* et *bioD* montrent des rapports d'expression supérieurs à un.

La figure 24 résume l'ensemble des régulations observées au niveau des gènes de la biosynthèse des acides aminés pour les deux conditions osmotiques.

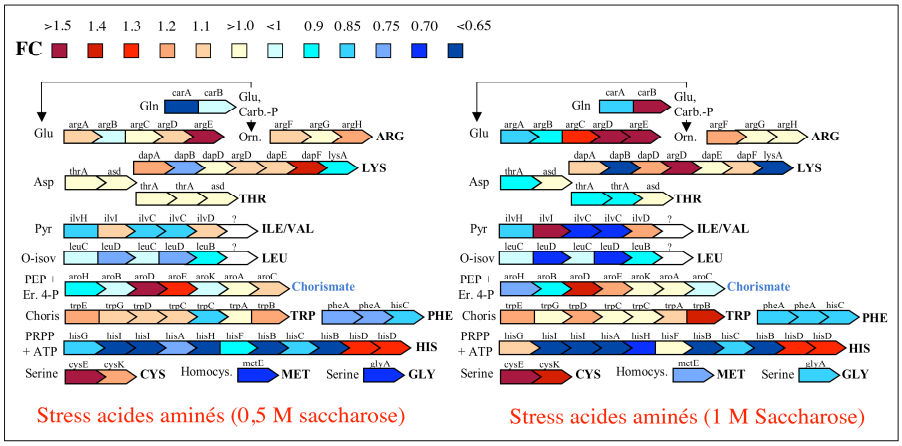

Figure 24. Représentation des rapports d'expression des gènes de la biosynhtèse des acides aminés chez *Buchnera* lorsque le puceron est soumis à un stress en acide aminés essentiels à 0,5 M et 1 M de saccharose.

Métabolisme des sucres et du peptidoglycane

Les gènes *eno* (énolase) et *gpmA* (2,3-biphosphoglycerate mutase) sont réprimés par le stress en acides aminés. Mais aucune tendance globale n'est observée sur ces deux voies.

Les gènes *mur* et *fts* (de BU215 à BU223) impliqués dans la biosynthèse du peptidoglycane sont réprimés à 0,5 M de saccharose (7/9 gènes, $p_w = 0,02$), mais cette différence n'est plus significative à 1 M de saccharose ($p_w = 0,12$). Aucun gène n'est significatif individuellement.

Transport, système flagellaire

Deux transporteurs (donneurs) d'électrons sont significativement réprimés par le stress en acides aminés essentiels : *cyoA* est le premier gène de l'opéron *cyo*, codant la sous-unité II du complexe terminal de la cytochrome oxydase et *fldA*, codant la flavodoxine-1, est un donneur d'électron pour de très nombreuses enzymes oxydoréductases.

Le gène *ompF* (porine) est réprimé à 1 M de saccharose.

Les deux opérons *fli* et *flh*, codant les protéines de la partie basale du flagelle associée à la membrane interne, sont globalement réprimés (11/15 gènes, p_w=0,001) dans les deux conditions de saccharose. Le gène *flgF* est significativement réprimé et les gènes *flgB* et *flgD* sont significativement surexprimés, mais globalement,

l'opéron *flg*, codant les protéines flagellaires associées à la membrane externe, n'est pas régulé dans l'une ou l'autre des conditions (5 gènes sur 12, $p_w = 0,5$).

Régulation de la transcription, de la traduction et métabolisme de l'ADN

Le régulateur transcriptionnel *himA* est significativement induit par le stress en acides aminés essentiels et dans les deux conditions de saccharose. Le régulateur *fis* est surexprimé en conditions de stress en acides aminés à 1 M de saccharose, mais *fis* est inhibé sous l'action du stress en saccharose à 50 % d'acides aminés ; c'est un gène qui se trouve dans la partie interaction. Il n'apparaît pas dans le diagramme de Venn acides aminés (figure 23) car la p-value associée au terme d'interaction pour ce gène est de 0,06. Ces deux gènes ont été analysés plus en détail dans le paragraphe 3.2.

Les gènes *nusG* (antiterminaison) et *gyrB* (topoisomérase) sont significativement réprimés dans les deux conditions de saccharose.

Le gène *speE*, codant la spermidine synthase et permettant la transformation de putrescine fournie par l'hôte à *Buchnera* en spermidine, est surexprimé dans les deux conditions de saccharose. Ce gène est classé dans cette catégorie fonctionnelle parce qu'il a été montré que la spermidine a un rôle important dans la structure d'enroulement de l'ADN et pourrait ainsi avoir un rôle dans la transcription chez les bactéries (Nakabachi et Ishikawa 2000). Cette explication reste hypothétique pour le moment chez *Buchnera*.

Le gène *infC*, codant le facteur d'initiation de la traduction IF3 responsable de l'activation de la sous-unité 30S du ribosome sur laquelle démarre la traduction, est significativement réprimé dans les deux conditions de saccharose. Le gène *hemK* est réprimé dans les deux conditions de saccharose, ce gène code une protéine impliquée dans la méthylation des facteurs de traduction RF-1 et RF-2. Le gène *ffr* codant le facteur de recyclage des ribosomes, activateur de la traduction en permettant à un ribosome d'être recyclé entre deux cycles de traduction différents, est significativement réprimé dans les deux conditions de saccharose. Ce gène est vraisemblablement corégulé avec *dxr* dans une structure en opéron conservée chez *E. coli*.

Les protéines ribosomales *rpmF* et *rpmG* sont significativement surexprimées dans les 2 conditions de saccharose. Une tendance à la surexpression est visible pour les gènes voisins sur cette région du chromosome. Mais une analyse de l'ensemble des « clusters » de gènes codant les différentes protéines ribosomales ne permet pas de dégager de tendance significative pour cette catégorie.

Repliement des protéines et réponse de stress

Le gène *hscA*, codant la protéine chaperone HscA, est réprimé à 0,5 M de saccharose. Le gène *hfld* code pour une protéine essentielle chez *E. coli*, c'est une protéase ATP dépendante de la membrane interne. Ce gène est activé sous l'action du stress en acides aminés à 0,5 M de saccharose.

Métabolisme lipidique

Dans les voies associées au métabolisme lipidique, le gène *dxr*, codant la protéine DXP reductoisomérase impliquée dans la biosynthèse de l'isopentényl diphosphate, est significativement réprimé dans les deux conditions de saccharose. Le gène *fabB*, codant la sous-unité I de la 3-oxalate synthase impliquée dans la biosynthèse des acides gras, est significativement activé dans les deux conditions de saccharose. Le gène *fabG* est invariant, le gène *fabI* est fortement réprimé bien que non significativement.

Divers

Le gène *yba2* (protéine inconnue) est réprimé à 1 M de saccharose par la déplétion en acides aminés essentiels.

3.4.2.2. Effet du stress en saccharose

La liste des gènes différentiels induits par un stress osmotique et pour les deux conditions d'acides aminés est donnée dans le tableau A.5 présenté en annexe. Comme précédemment, la liste des gènes peut être présentée sous forme d'un diagramme de Venn (figure 25).

Au total 62 gènes, dont 24 activés et 38 réprimés, ont été détectés comme exprimés de façon différentielle. Les principales catégories fonctionnelles concernées par la réponse de *Buchnera* sont détaillées ci-dessous.

Métabolisme des acides aminés et des vitamines

Les gènes *leuD*, *argH* et *cvpA* sont significativement surexprimés après le stress osmotique, alors que les gènes *trpB*, *carB*, *cysK* et *thrA* sont réprimés. A part la voie de biosynthèse de la leucine (cinq gènes sur sept, $p_w = 0,04$), aucune voie n'est globalement significative dans cette catégorie fonctionnelle.

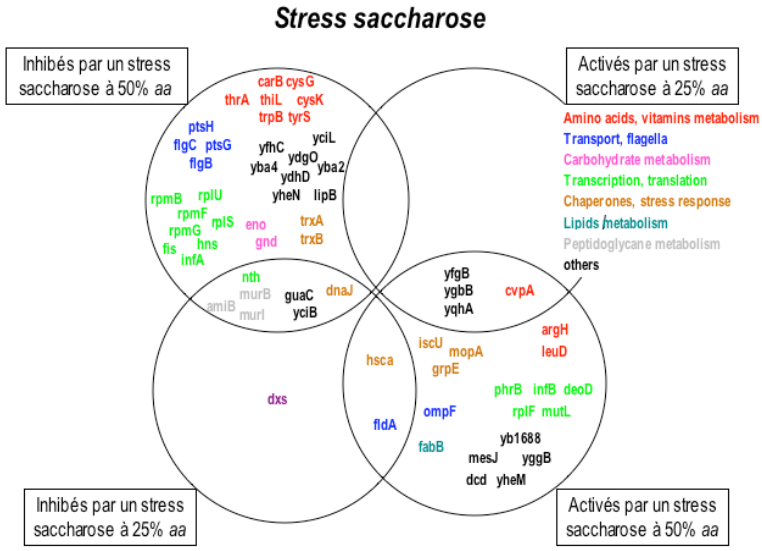

Figure 25. Diagramme de Venn correspondant aux gènes différentiels en condition de stress de saccharose.

Métabolisme des sucres et du peptidoglycane

Deux gènes impliqués dans le métabolisme des sucres ont été détectés par l'analyse différentielle : les gènes *eno* et *gnd*, codant pour l'énolase (glycolyse) et la G-phosphogluconate deshydrogénase (voie des pentoses phosphates). Ces deux gènes sont réprimés significativement. Une analyse plus détaillée des différents gènes de la glycolyse montre que la majorité sont réprimés par le stress osmotique. Ainsi, tous les gènes *pfka*, *fba*, *gapA*, *pgk*, *gpmA*, *eno* et *pyrkA* permettant le passage du fructose-P au phosphoénolpyruvate (PEP) montrent des rapports différentiels inférieurs à un, en réponse au stress osmotique à 50 % d'acides aminés essentiels ($p_w = 0,05$). La voie des pentoses phosphates montre la même tendance ($p_w = 0,01$). Ces résultats restent valables pour les deux conditions de sucrose.

La figure 26 ci-dessous illustre les régulations observées dans le métabolisme du sucre chez *Buchnera* en condition de stress osmotique du puceron.

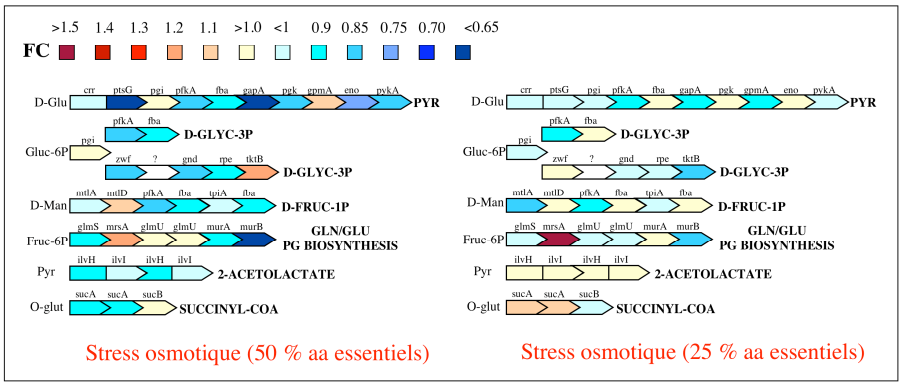

Figure 26. Représentation des rapports d'expression des gènes de la glycolyse et de la voie des pentoses phosphates chez *Buchnera* lorsque le puceron est soumis à un stress osmotique (A) 50 % d'acides aminés essentiels, (B) 25 % d'acides aminés essentiels.

Les gènes *amiB*, *murB* et *murI* impliqués dans la biosynthèse du peptidoglycane sont significativement réprimés par le stress osmotique, les autres gènes *mur* de la région BU215-223 sont globalement réprimés (7/9), mais la moyenne n'est pas significative (p_w = 0.12), vraisemblablement à cause de *murG* qui est assez fortement surexprimé mais avec un niveau d'expression très bas.

Transport, système flagellaire

Le gène *fldA* (flavodoxine I, donneur d'électron) montre une très forte interaction en étant activé à 50 % et inhibé à 25 % d'acides aminés essentiels.

La répression de deux gènes codant le système de transporteur PTS du glucose chez *Buchnera* (*ptsG* et *ptsH*) peut également être observée. Les gènes *crr* et *ptsI*, codant les protéines associées au complexe, sont globalement invariants, tout comme le gène *mtla*, codant la partie membranaire du complexe PTS associée au transport du mannitol.

Le gène codant la porine *OmpF* est significativement surexprimé par le stress osmotique.

Sur les 12 gènes de l'opéron *flg*, dix sont réprimés (p_w=0,01), ainsi que neuf sur 12 des gènes de l'opéron *fli* (p_w = 0,04), les gènes *flhA* et *flhB* sont invariants. L'expression des gènes codant les protéines du flagelle est donc fortement réprimée par le stress osmotique.

Régulation de la transcription, de la traduction et métabolisme de l'ADN

Le régulateur transcriptionnel *fis* est réprimé par le stress osmotique (la régulation de *fis* est décrite dans le § 3.2 ci-dessus).

Le gène *nth*, codant l'endonucléase III responsable de la dégradation des brins d'ADN endommagés, est réprimé alors que les gènes *mutL* (réparation des mésappariements de l'ADN), *phrB* (deoxyribodipyrimidine photo-lyase) et *deoD* (nucleoside purine phosphorylase) sont surexprimés significativement par le stress osmotique.

Le gène *yciL*, codant la protéine RluB, est réprimé. Ce gène code pour la sous-unité B de l'enzyme responsable de la synthèse de la pseudouridine (dégénérescence des ARNt). Le gène *guaC*, codant la GMP réductase, est réprimé. Cette enzyme catalyse la déamination irréversible et NADPH dépendante du GMP en IMP et est impliquée dans le maintien de l'équilibre entre les concentrations en nucléotides libres A et G.

Les gènes *tyrS* (tyrosyl ARNt synthase), *infA* (facteur de traduction IF-1) et *lipB* (lipoylation des apoprotéines) sont réprimés significativement par le stress en acides aminés. A l'inverse, le gène *infB* codant le facteur de traduction IF-2 est surexprimé.

Cinq protéines ribosomales sont significativement réprimées alors qu'une seule est surexprimée. Une analyse plus complète montre cependant qu'il est difficile de trouver une tendance générale sur l'ensemble des protéines ribosomales.

Repliement des protéines et réponse de stress

Sept protéines chaperonnes sont significativement régulées en réponse au stress en saccharose. Les gènes *dnaJ*, *trxA et trxB* sont réprimés significativement. Le gène *dnaK* associé en opéron à *dnaJ* montre un rapport d'expression différentielle inférieur à un. A l'inverse, les gènes *grpE* et *iscU* sont activés dans cette condition. Le gène *mopA* (codant la protéine Hsp60) est significativement surexprimé par le choc osmotique, alors que le gène *mopB* (Hsp10) est réprimé non significativement. Ce résultat est très intéressant dans la mesure où la protéine Hsp10 a été montrée toxique à forte concentration chez *E. coli* (Zeilstra-Ryalls *et al.* 1991). Une régulation différentielle au niveau transcriptionnel existe chez *E. coli* entre les deux gènes pourtant associés dans un même opéron. Cette régulation est suspectée chez *Buchnera* également (Sato et Ishikawa 1997). Enfin, le gène codant la protéine chaperonne HscA est surexprimé en réponse au stress osmotique à 50 % d'acides aminés mais réprimé à 25 %.

Métabolisme lipidique

Le gène *fabB*, impliqué dans la biosynthèse des acides gras, est faiblement surexprimé, les autres gènes de la voie sont invariants. Le gène *dxs*, codant la DXP synthase dans le métabolisme des terpènes, est le seul gène significativement réprimé par le seul stress osmotique à 25% d'acides aminés essentiels.

Divers

Le gène *yciB* code la protéine IspZ probablement impliquée dans la septation cellulaire ; il est réprimé sous l'action du stress osmotique, de même que le gène *mesJ,* codant une protéine de fonction inconnue classée dans la division cellulaire chez *E. coli*.

Treize gènes codant des protéines hypothétiques sont régulées par le stress osmotique. Les gènes *yba2, y187*, *yba4*, *yciL*, *ydhD*, *yfhC*, *yheN*, *yciB* sont réprimés, alors que les gènes *yb1688*, *yggB*, *yheM*, *yfgB*, *yqhA* sont surexprimés.

4. Discussion

La première partie de ce chapitre concerne les capacités de régulation transcriptionelle de la bactérie *Buchnera*. Les études publiées jusqu'à présent au niveau théorique prédisent un potentiel de régulation très réduit de la bactérie. Cependant, notre travail montre que, malgré des rapports d'expression relativement faibles, une réponse significative de la bactérie peut être quantifiée. Cette réponse est corrélée à l'organisation du génome de la bactérie.

Ainsi, les gènes de *Buchnera* ne sont pas tous exprimés au même niveau. Les plus conservés (ce sont également les plus riches en bases G et C chez *Buchnera*) sont les plus exprimés. La présence de gènes très faiblement exprimés signifie que le jeu de gènes conservés chez *Buchnera* pourrait encore être réduit. Tamas *et al.* (2002) affirment que les lignées de *Buchnera* sont entrées dans une phase stationnaire ne disposant plus des séquences répétées et des enzymes nécessaires pour continuer la réduction du génome. Néanmoins, les 450 kb observés chez la bactérie symbiotique du puceron du cèdre, *Cinara cedri*, semblent contredire cette hypothèse (Gil *et al.* 2002). Le génome de cette bactérie est en cours de publication (Moya, communication personnelle) et l'analyse comparative des quatre génomes de *Buchnera* sera certainement très informative sur ce point. A l'inverse, les gènes codant le système flagellaire de *Buchnera* sont très fortement exprimés bien qu'ils évoluent très rapidement. Ces mêmes gènes se sont révélés souvent surexprimés ou réprimés dans les différentes expériences nutritionnelles réalisées. Le flagelle de *Buchnera* est incomplet, le gène codant la protéine du filament FliC est absent. Il est donc non fonctionnel pour la motilité de la bactérie. Au vue de ces résultats, nous avons émis l'hypothèse que le flagelle pourrait servir de transporteur peptidique

chez *Buchnera*. En effet, il a été montré récemment que des mutants *ΔfliC* étaient capables d'exporter des protéines recombinantes à travers le pore flagellaire (Majander *et al.* 2005).

L'organisation en structures de régulation (unités de transcription) chez *Buchnera* est révélée par l'analyse de la variabilité de l'expression absolue des gènes à l'intérieur des opérons (et des syntons) toujours beaucoup plus faible qu'entre ceux-ci. Cette organisation est très importante puisqu'elle explique entre 40 et 63 % de la variabilité totale des niveaux d'expression. Le même résultat est observé avec les niveaux d'expression différentielle, ce qui signifie que si un gène est régulé dans une unité de transcription, les autres gènes de la structure sont corégulés. Cette régulation à l'intérieur des unités est importante, car bien qu'elles ne soient pas régulées simultanément dans nos diverses expériences, la variance intra-structure représente de 23 à 36 % de la variabilité totale des rapports d'expression différentielle. Enfin, les syntons sont toujours plus significatifs que les opérons, ce qui semble indiquer que les unités de transcription correspondent plus aux syntons qu'aux opérons chez *Buchnera*. Cette idée de « super-opérons » est en accord avec les travaux de Moran et Mira (2001a) et Silva *et al.* (2001) qui ont suggéré qu'au court de la dégradation du génome de *Buchnera* certains gènes se seraient rapprochés. Les nouveaux gènes adjacents auraient formé des unités de transcription en bénéficiant de l'effet de certains promoteurs forts. Ces travaux purement théoriques n'ont pas encore été étayés par des preuves expérimentales. Il faut noter que la comparaison des deux modèles opérons et syntons de notre analyse est rendue délicate par des degrés de liberté différents : il y a plus de gènes impliqués dans les syntons et plus de syntons que d'opérons. Les structures en syntons expliquent d'ailleurs toujours moins de pourcentage de variabilité (comparaison des R^2) que les structures en opérons. Ce dernier point méritera donc une analyse plus approfondie.

Nous avons également étudié la structure en régulons chez *Buchnera*. Les régulateurs transcriptionnels induits dans nos expériences de stress trophiques sont tous des régulateurs globaux de la famille des protéines « histone-like ». Le régulon IHF est activé par un stress nutritionnel subi par le puceron hôte. Les régulateurs *fis* et *hns* sont, eux aussi, régulés dans nos expériences, mais la trace de leur régulation n'a pas pu être détectée de façon claire parmi les effecteurs de *Buchnera*. Peut-être parce que ces deux facteurs ne sont plus fonctionnels chez *Buchnera*. Ils peuvent également avoir évolué en inversant leur régulation. Il est enfin possible que nos conditions expérimentales et les interactions entre les différents facteurs ne nous aient pas permis de qualifier correctement ces régulations. Néanmoins, nos analyses suggèrent que l'absence de régulateurs transcriptionnels spécifiques est peut-être compensée par un rôle accru de certains régulateurs globaux comme les protéines « histone-like » chez *Buchnera*. Cette hypothèse mérite bien sûr une analyse plus poussée. Un groupe de quatre étudiants de la promotion 4 BIM INSA est en train de travailler sur ce sujet, dans le cadre de leur projet annuel.

Nous avons essayé d'analyser les contraintes évolutives agissant sur l'usage du code chez *Buchnera*. Le biais mutationnel est la principale force, indépendante de l'expressivité, déterminante du code génétique. Cependant, une corrélation significative a été observée entre l'abondance des ARNt isoaccepteurs et la composition en codons des gènes de *Buchnera*, à l'exception des deux isoaccepteurs $ARNt^{UUU}$ et $ARNt^{TUU}$. Pour ces deux derniers, l'ajustement de leur niveau d'expression n'a apparemment pas pu être réalisé par la bactérie. Il faut noter que les deux codons lysine et leucine correspondants représentent près de 15 % de la totalité des codons de *Buchnera*. D'autre part, un biais d'usage du code significatif a été observé pour les codons rares de *Buchnera*, avec une préférence globale pour les codons terminés par un C alors que les codons terminés par un G sont plutôt rejetés. Nous avons montré que la dissymétrie de répartition des gènes sur les deux brins de l'ADN ne permet pas d'expliquer ce biais d'usage du code. Nous n'avons cependant pas été en mesure de vérifier si ce biais avait un rôle fonctionnel ou s'il correspondait plutôt à un vestige issu d'une bactérie libre ancestrale. Il existe des preuves expérimentales montrant que les codons rares peuvent réduire le rendement de la réaction de traduction chez *E. coli*. Cette propriété est bien connue des chercheurs travaillant dans la production de protéines hétérologues (Sharp et Li 1986 ; McNulty *et al.* 2003). Chez les organismes montrant une forte sélection d'usage du code, on doit donc s'attendre à une éviction des codons rares dans les gènes fortement exprimés. Par ailleurs, il a été démontré chez la drosophile que la fidélité de la traduction était meilleure lorsque la paire codon / anticodon ne présente pas une complémentarité parfaite (Akashi 1994). La conservation des codons rares dans les gènes essentiels et/ou fortement exprimés pourrait alors correspondre chez *Buchnera* à une sélection, pour une meilleure qualité de la traduction, de certains sites ou acides aminés ayant un rôle fonctionnel prépondérant pour la protéine. Cette hypothèse a été émise par C. Rispe dans son mémoire d'Habilitation à Diriger des Recherches (2005). Nous avons alors tenté de la vérifier en utilisant le gène *mopA* codant la protéine GroEL pour laquelle une information sur les sites fonctionnels était disponible. Nous n'avons pas pu mettre en évidence une relation claire entre les sites fonctionnels et l'utilisation de codons terminés par un C sur ce gène. Ce travail nécessitera une analyse plus poussée intégrant une plus grande diversité de protéines.

Nous avons ensuite montré que lorsque le milieu nutritif des pucerons est incomplet en acides aminés aromatiques, les ARNt sont surexprimés par la bactérie. Une surexpression des ARNt est généralement corrélée à une augmentation du métabolisme cellulaire et de la division. L'analyse des paramètres biologiques des pucerons élevés sur le milieu YF montre que les pucerons souffrent assez peu du manque d'acides aminés. Seule la variable poids (-20 %) est affectée, la fécondité et la survie ne sont pas affectées. Ces résultats sont en accord avec une augmentation du métabolisme de *Buchnera* pour la complémentation nutritionnelle du puceron. La mesure de la densité bactérienne n'a pas été réalisée dans l'expérience YF, elle a par

contre été faite dans l'expérience AAS qui présente un déficit en tyrosine et phénylalanine entre les conditions 0,5 et 1 M de saccharose pour une concentration en acides aminés essentiels de 50 % (tableau 6). Le nombre de *Buchnera* est presque deux fois plus grand dans la condition déficiente. La différence, même si elle s'accorde avec cette hypothèse de croissance bactérienne, est toutefois non significative.

Dans la deuxième partie de ce chapitre, les résultats des deux expériences ont été analysés au niveau fonctionnel. L'expérience YF est très ciblée sur un métabolisme crucial du système symbiotique : la biosynthèse de la phénylalanine et de la tyrosine. La deuxième expérience (AAS) est beaucoup plus générale et porte globalement sur la fourniture d'acides aminés essentiels par *Buchnera* au puceron, combinée avec l'analyse de la tolérance à l'hyper-osmolarité du milieu.

Les expériences sont réalisées avec des pucerons élevés sur des milieux artificiels, mais elles restent toutefois assez physiologiques. En effet, l'analyse des paramètres biologiques des pucerons sur les milieux témoins montre qu'ils ont un développement quasi normal. Sur les milieux incomplets, on peut noter que *Buchnera* parvient à complémenter le déficit nutritionnel de son hôte. On doit donc considérer que, lorsque les pucerons sont placés sur un milieu incomplet, leur demande nutritionnelle vers *Buchnera* est accrue. Mais cette analyse peut être compliquée par des facteurs comportementaux du puceron. Parmi ceux-ci, le plus important est sans doute l'appétence du milieu. En effet, même si le saccharose est un facteur phagostimulant pour les pucerons, ceux-ci, soumis à un régime fortement hyper-osmotique vont avoir tendance à moins s'alimenter. L'analyse du budget, quantifiant les apports du milieu et les besoins nutritionnels des pucerons, permet d'estimer très précisément la demande du puceron envers sa bactérie.

Les rapports d'expression mesurés chez *Buchnera* dans les deux expériences sont relativement faibles, majoritairement compris entre 1,5 et 1,2 pour les gènes surexprimés entre 0,66 et 0,83 pour les gènes sous-exprimés. Ces faibles valeurs s'expliquent tout d'abord par notre plan expérimental qui comportent de nombreuses répétitions de lames et de redondances intra-lames, si bien que les valeurs finales obtenues sont très fortement moyennées. D'un point de vue physiologique, il faut souligner que les conditions nutritionnelles testées sont appliquées sur le puceron hôte. Elles sont donc tamponnées par l'homéostasie de l'insecte. Ainsi, lorsque *E. coli* est soumis à un stress nutritionnel modéré, sa réponse transcriptionnelle est tout à fait comparable à la réponse observée chez *Buchnera* (Tao *et al.* 1999 ; Berka *et al.* 2003) pour la grande majorité des gènes. Par ailleurs, la mesure d'expression est réalisée à l'équilibre des flux (5 à 7 jours de nutrition), on peut supposer que certains gènes sont plus fortement régulés au début du processus d'adaptation au stress trophique. On peut noter encore que la mesure réalisée correspond à une moyenne des niveaux d'expression des gènes des différentes populations de

Buchnera dans le puceron. Les *Buchnera* des embryons montrent un métabolisme beaucoup plus actifs que les *Buchnera* issues des bactériocytes maternels (Wilkinson et Ishikawa 1999). Enfin, d'un point de vue évolutif, de la même façon que le génome de *Buchnera* a été épuré de son contenu génétique à cause de la stabilité des conditions environnementales, on doit imaginer que sa régulation transcriptionnelle a également été épurée. Il paraît donc relativement logique d'observer une réponse transcriptionnelle limitée comparativement à celle d'autres bactéries libres. La réponse de *Buchnera* paraît plus forte et plus spécifique dans l'expérience de déplétion en acides aminés essentiels (AAS) que dans celle de déplétion en tyrosine et phénylalanine (YF). Néanmoins, ce résultat est certainement lié à la différence de qualité entre les deux jeux de données. Par ailleurs, la réponse quantitativement la plus forte (impliquant le plus de gènes différentiels) est la réponse au stress osmotique, ici encore ce résultat peut être un artéfact lié au déséquilibre du plan expérimental décrit au chapitre 3 (§ 2.1.2).

Métabolisme des acides aminés

Pour ce qui concerne les gènes codant les enzymes du métabolisme des acides aminés, la réponse la plus marquée est celle de la déplétion en acides aminés essentiels (AAS), puisqu'elle concerne pratiquement toutes les voies des acides aminés essentiels de *Buchnera*. Cette réponse est plus marquée à 1 M de saccharose qu'à 0,5 M. Cette observation est cohérente avec l'analyse du budget montrant une demande du puceron plus forte à 1 M qu'à 0,5 M. Elle se caractérise par une activation de la voie du tryptophane, de l'arginine et de la cystéine, et une répression des voies de biosynthèse de la leucine et de l'histidine.

Pour les acides aminés aromatiques, la branche du chorismate est globalement activée ce qui est cohérent avec une production globale d'acides aminés aromatiques augmentée. La demande en phénylalanine étant forte pour les deux expériences YF et AAS, on attendait une activation du gène *pheA* pour orienter le flux de métabolites vers la production de cet acide aminé. Mais *pheA* reste invariant dans les deux expériences. La surexpression du gène *cvpa* pourrait indiquer une activation de la « navette » phénylalanine – tyrosine, mais cette hypothèse devra être confirmée par des expériences supplémentaires.

La forte demande en lysine de l'expérience AAS est à corréler avec l'activation de la voie de biosynthèse de cet acide aminé à 0,5 M de saccharose, et de façon plus forte à 1 M. Néanmoins, les résultats ne sont pas significatifs globalement.

Les voies de biosynthèse de l'arginine et de la cystéine sont activées significativement dans l'expérience. Il n'y a pas d'analyse du budget pour ces acides aminés dans la mesure où ils ne sont pas essentiels. L'arginine est néanmoins un des acides aminés les plus représentés des protéines du puceron et l'on peut s'attendre à

une fourniture par *Buchnera*. De même l'activation de la biosynthèse de la cystéine pourrait s'expliquer par la demande en méthionine assurée par la bactérie grâce à l'interconversion cystéine-méthionine réalisée par la bactérie, et ceci en l'absence du régulateur transcriptionnel *metR* activant chez *BuchneraSG* par exemple la régulation directe de la biosynthèse de la méthionine (Wilcox *et al*. 2004).

La demande en acides aminés branchés est forte dans l'expérience AAS alors que la voie n'est pas significativement activée. On peut noter que le gène *ilvHI* est l'un des gènes les plus exprimés chez *Buchnera*, ce qui est assez surprenant pour un gène codant une enzyme du métabolisme. On peut supposer que l'expression très forte du gène *ilvHI* induit une surproduction constante de ces acides aminés chez *Buchnera*. Cette observation n'explique cependant pas la répression de la voie de la leucine dans nos expériences, alors que la demande pour cet acide aminé est forte.

L'inhibition de la voie de l'histidine n'était pas prévue par l'analyse du budget (la demande est toutefois nulle pour cet acide aminé). Néanmoins, G. Febvay a montré qu'il n'est pas possible d'obtenir une synthèse d'histidine de *Buchnera* dès lors que cet acide aminé est présent dans le milieu nutritif des pucerons (communication personnelle). Nos résultats confortent donc cette observation. De même, la demande en thréonine est très faible, voire nulle dans nos expériences. Les voies de biosynthèse ne sont pas clairement régulées. Notons que la sève phloémienne des légumineuses est toujours relativement riche en thréonine, cette absence de réponse de la part de la bactérie n'est donc pas surprenante.

Métabolisme des sucres

La réponse de *Buchnera* liée au métabolisme des sucres et du peptidoglycane est la plus forte dans le cas du stress osmotique. Cette réponse est beaucoup plus marquée à 50 % d'acides aminés qu'à 25 %. Lorsque le milieu nutritif du puceron est hyper-osmotique, les voies de la glycolyse, des pentoses phosphates et de la biosynthèse du peptidoglycane sont réprimées significativement chez *Buchnera* ; ceci s'estompe si la demande métabolique est accrue. Ce résultat n'est pas facilement interprétable dans la mesure où l'un des rôles de *Buchnera* dans l'osmorégulation est la consommation du saccharose intracellulaire. Il faut noter toutefois que, dans les quatre conditions expérimentales, le taux de saccharose du bactériocyte ne varie pas significativement (Douglas, communication personnelle). La réponse induite sur le métabolisme carboné est donc issue d'un processus de signalisation induit par l'hôte, vraisemblablement au niveau du tube digestif, puis du bactériocyte. Cette voie de signalisation serait très intéressante à étudier.

Transport, système flagellaire

Les gènes codant les transporteurs et le système flagellaire de *Buchnera* sont régulés dans les trois types de conditions expérimentales. Les réponses les plus fortes sont associées à l'expérience AAS pour les deux types de stress (osmotique et acides aminés). Le rôle probable du flagelle dans l'exportation des polypeptides a déjà été discuté. La régulation des gènes codant le flagelle est relativement complexe chez *E. coli* (Lehnen *et al.* 2002). Chez *Buchnera*, l'interprétation fonctionnelle des régulations observées semble également assez complexe, dans une même expérience les gènes codant certaines parties du flagelle sont activés alors que d'autres sont réprimés.

Il est notable que les conditions hyper-osmotiques entraînent la répression des deux gènes codant les transporteurs PTS de *Buchnera*. Ces mêmes gènes sont activés par la déplétion YF. Le gène codant la porine OmpF est réprimé par la déplétion en acides aminés essentiels mais activé par le stress osmotique.

Régulation de la transcription, de la traduction et métabolisme de l'ADN

L'activation ou la répression de gènes comme *ligA*, *repA*, *nth*, *mutL*, *phrB* ou *deoD* pourrait avoir un impact important sur la réplication et le métabolisme de l'ADN, influençant ainsi la transcription via la régulation de la quantité d'ADN cellulaire. Rappelons que *Buchnera* est une bactérie polyploïde contenant de 20 à 80 chromosomes par cellule (Komaki et Ishikawa 1999). Mais il est cependant très difficile d'argumenter sur le rôle et l'impact physiologique de la régulation de ces différents gènes chez *Buchnera* en l'absence d'expérimentations complémentaires. La quantité d'ADN de *Buchnera* dans les quatre conditions physiologiques de l'expérience AAS a été dosée par PCR quantitative par l'équipe de A. Douglas à York, et n'a pas montré de différences significatives.

La régulation de la transcription par des topoisomérases, comme *gyrB,* ou par l'action de la spermidine (gène *speE*) sur la super hélicité de l'ADN est également très spéculative. Par contre, l'observation de l'induction des régulateurs transcriptionnels *fis* et *himA* semble plus significative (ce point a déjà été discuté auparavant). De même, la traduction semble être une fonction affectée par nos conditions expérimentales, la réponse la plus forte concernant le métabolisme des ARNt. Ce point a été détaillé dans la première partie de cette discussion.

Réponse de stress et autres fonctions

Une réponse de stress est observée dans toutes les expériences, la plus forte réponse concerne le stress osmotique puisque sept gènes codant des chaperonnes ont été identifiés dans la réponse de *Buchnera.* La déplétion YF recense également cinq gènes dans cette classe, alors que la déplétion en acides aminés essentiels n'en régule qu'un seul. Beaucoup de protéines impliquées dans les mécanismes de

repliement sont aussi impliquées dans le transport. Comme le transport est un point clé de la relation *Buchnera*-puceron, il est possible que le rôle de ces protéines soit associé à cette fonction. Néanmoins, ici encore, cette hypothèse pour être testée nécessitera des expériences complémentaires. Enfin, on peut noter que de façon très significative beaucoup de protéines hypothétiques ont été détectées dans la liste des gènes différentiels induits par le stress osmotique, contrairement au deux autres conditions expérimentales. Ce résultat traduit sans doute un changement physiologique plus global suite au stress osmotique relativement au stress en acides aminés.

5. Conclusions

Ce chapitre a été consacré à l'étude du transcriptome de la bactérie *Buchnera*. Nous nous sommes placés aux niveaux évolutif puis fonctionnel pour répondre à deux questions : (1) quelles sont les capacités de régulation transcriptionnelle de la bactérie en relation avec l'organisation de son génome et (2) *Buchnera* est-elle capable de répondre à une modification de la demande nutritionnelle de la part de son hôte ?

La première partie du travail montre qu'il est possible de détecter une trace de régulation liée à l'organisation du génome. Les gènes les plus conservés sont les plus exprimés, les gènes les plus exprimés montrent un usage de codons rares particulier, les gènes sont corégulés en opérons et peut-être même en régulons. Même si les résultats de ce travail sont statistiquement significatifs rien ne prouve qu'ils ont une importance fonctionnelle (phénotypique). Ces structures de regulation pourraient correspondre à des vestiges de la bactérie libre ancestrale, en voie de dégradation chez *Buchnera*. Le cas de l'usage de codons rares chez *Buchnera* illustre bien cette idée.

La deuxième partie correspond à l'analyse fonctionelle et cherche des liens de causalité entre les fonctions des gènes et le contexte physiologique des expériences pour « expliquer » les régulations d'expression. Cette partie est également très spéculative. Le piège des tests multiples, évoqué dans le chapitre 3, est présent à ce niveau de l'analyse. En effet, lorsque l'on dispose d'un nombre aussi conséquent d'informations sur des expériences très globales, on peut toujours trouver un argument explicatif pour justifier de la surexpression ou de la répression de chacun des gènes. Les hypothèses émises ne sont pour nous que des pistes de travail pour nos investigations futures (*cf.* § chapitre 5).

De façon très schématique, le stress en acides aminés essentiels dans le substrat nutritif du puceron se traduit par l'activation de gènes spécifiques de la biosynthèse des acides aminés essentiels chez *Buchnera*, alors qu'un stress osmotique induit des

gènes du métabolisme sucré. Le résultat le plus frappant est la régulation de voies de biosynthèse complètes malgré l'absence des régulateurs transcriptionnels correspondants chez *Buchnera*. L'action des régulateurs globaux est une explication possible de ce résultat. Ici encore, les très faibles variations d'expression et parfois l'absence de cohérence dans les réponses transcriptionnelles, nous invite à nous interroger sur la signification phénotypique de la réponse transcriptionnelle de la bactérie. Néanmoins, la succession de petites surexpressions des gènes le long d'une voie se traduit par un effet multiplicatif sur le flux de métabolites dans cette voie. On peut donc penser que les régulations les plus fortes et les plus globales observées dans nos expériences ont une signification biologique. L'incorporation des résultats d'expression dans un modèle de flux, ainsi que le suivi de métabolites chez *Buchnera* pourraient se révéler très informatif pour valider nos résultats.

Chapitre 5

Conclusions générales et perspectives de recherche

Le travail présenté dans ce mémoire correspond à l'acquisition, au sein du laboratoire BF2I, d'une compétence en analyse du transcriptome. Cette compétence porte à la fois sur les aspects méthodologiques (outils bioinformatiques pour la conception et l'analyse des données de puces à ADN) et expérimentaux (technique de fabrication des puces, hybridation et planification expérimentale). Les résultats déjà obtenus démontrent d'une part la faisabilité de ce type d'analyse sur le modèle *Buchnera* et, d'autre part donnent une première vision globale du système de régulation de la bactérie. Les quelques cibles repérées dès à présent permettent d'envisager de nouvelles expérimentations qu'il aurait été impossible de concevoir sans cette analyse globale.

Le chapitre précédent finit sur une liste de résultats issus de nos expériences de stress trophique, suivie d'une question assez « dramatique » : la réponse de *Buchnera* est-elle vraiment fonctionnelle ? Répondre à cette question sera sans doute l'un des enjeux majeurs pour les prochaines années de notre groupe de recherche (symbiose : génomique fonctionnelle des interactions trophiques) au laboratoire BF2I.

Les perspectives de ce travail peuvent se décomposer en trois volets que nous avons hiérarchisés. Le premier constitue le socle de notre projet et il concentrera

l'essentiel des forces de notre thématique au moins à court terme. Le deuxième sera réalisé en collaboration avec MF. Sagot (LBBE, UCBL). La partie expérimentale ne sera développée ultérieurement que si les résultats de ces études théoriques sont concluants. Enfin, le troisième volet ne sera abordé que si les forces de notre groupe sont augmentées. En effet, un poste de CR2 INRA orienté sur cette problématique a été demandé au département SPE de l'INRA pour l'année 2007.

Volet 1. Caractérisation du transcriptome de Buchnera

Il s'agit de la continuation directe de notre travail. Des expériences seront conduites pour enrichir la base de données d'expression de *Buchnera* : analyse de la déplétion en acides aminés branchés (thèse de J. Viñuelas), étude de l'influence de *Buchnera* sur le développement du puceron (collaboration avec T. Wilkinson, Université de Dublin), caractérisation des interactions de *Buchnera* avec les mécanismes de défense des plantes (induction par le jasmonate). Des plans d'expériences judicieux devront être établis et des méthodes statistiques plus performantes (modèles mixtes) seront certainement utilisées. Enfin, des méthodes d'analyses transversales (classification notamment) pourront être envisagées dès lors que le jeu de conditions expérimentales sera devenu plus conséquent.

D'autre part, certaines des hypothèses émises dans ce travail ouvrent la voie à plusieurs champs d'investigations :

(1) Evolution de la régulation transcriptionnelle chez *Buchnera* : recherche automatique des unités de transcription (séquences promotrices, terminatrices, distances intergéniques, génomique comparative) et caractérisation des régulateurs globaux (modélisation moléculaire, recherche des sites de fixation, génomique comparative). Un étudiant de Master sera recruté pour l'année 2006-2007 sur cette thématique.

(2) Caractérisation fonctionnelle et évolution des systèmes de transport chez *Buchnera* : recherche *in silico* des systèmes de transport, analyse comparative et modélisation moléculaire. Mise en évidence de la fonction de transport du flagelle de *Buchnera*. Yvan Rahbé (DR INRA, BF2I) a initié des recherches sur ce sujet.

(3) Rôle du gène *cvpa* chez *Buchnera* : mise au point d'une méthode de dosage HPLC du polypeptide, suivi de la concentration protéique dans les différents compartiments cellulaires, induction du gène par différents stress trophiques et caractérisation du transporteur membranaire chez *Buchnera*. Ce projet n'est pas planifié pour le moment.

Parallèlement, une base de connaissances sur *Buchnera*, incluant les annotations génomique et métabolique ainsi que les régulations transcriptionnelles, devra être développée pour faciliter l'intégration des données dans le contexte physiologique (biologie intégrative). Pour cela, nous travaillons en collaboration avec le groupe

d'A. Viarri (HELIX INRIA Grenoble) et F. Rechenman (plateforme Génostar). Notre collaboration avec A. Douglas et G. Thomas (Université de York) qui développent BuchneraBASE (Prickett et al. 2006) sera également enrichissante sur ces aspects.

Le troisième point de ce programme tente de répondre à la question de la fonctionnalité de la réponse transcriptionnelle de *Buchnera* par la modélisation de son réseau de régulation génétique (modélisation aux niveaux fonctionnel et évolutif). Ce point est très complexe et peut être abordé de différentes façons. Une première façon de procéder consiste à utiliser une approche comparative basée sur les relations d'orthologie entre les gènes de *Buchnera* et *d'E. coli*. Une modélisation dynamique (système d'équations différentielles) peut alors être envisagée, mais elle se heurtera rapidement à la méconnaissance de la régulation transcriptionnelle de *Buchnera*. Ce point est à l'étude grâce à un groupe d'étudiantes de 4ème année de la formation BiM de l'INSA. Une autre façon de procéder consiste à inférer les relations entre gènes directement à partir des données d'expression, par des méthodes d'inférences bayésiennes notamment. Une littérature très abondante est disponible sur ce sujet et nous avons réfléchi à l'utilisation de ces méthodes sur nos données en collaboration avec S. Genieys (Institut C. Jourdan, UCBL) dans le cadre du groupe BSMC (http://bsmc.insa-lyon.fr). Le point clé de ces études est la nature des expérimentations qui permettront d'inférer le réseau. Des expériences cinétiques devront être élaborées. Enfin, l'analyse de l'évolution du génome de *Buchnera* dans son contexte symbiotique devra être poursuivie par la caractérisation des relations entre expression et organisation du génome en utilisant, d'une part les méthodes classiques de l'évolution moléculaire, et d'autre part les méthodes de simulation informatique développé par le groupe de G. Beslon (Prisma, INSA) dans le cadre de BSMC.

Enfin, le quatrième point de ce programme est issu de la biologie des systèmes puisqu'il vise à caractériser les fluctuations aléatoires de l'expression des gènes chez *Buchnera* relativement à *E. coli*. Kupiec et Sonigo (2000) affirment que comme le gène est apparu par sélection naturelle, son fonctionnement est sculpté par la sélection naturelle. Il n'existe donc pas un activateur (une clé) spécifique d'un effecteur donné (une serrure). Les deux entités ont coévolué et ont été sélectionnées pour interagir entre elles. L'expression des gènes doit donc être de nature aléatoire à l'échelle cellulaire, selon cette théorie également désignée comme « Darwinisme moléculaire ». En effet, il a été montré chez les eucaryotes et chez les procaryotes que, même si l'expression moyenne des gènes semble déterministe à l'échelle d'une population de cellules, au niveau cellulaire elle apparaît aléatoire (Elowitz *et al.* 2002 ; Paldi 2003 ; Thattai et van Oudenaarden 2004). *Buchnera* a évolué dans un environnement intracellulaire stabilisé ce qui a conduit à la perte de nombreux gènes devenus inutiles. On peut supposer que la dégradation du génome de *Buchnera* s'est accompagnée d'une diminution de la stochasticité de l'expression de ces gènes par

phénomène d'hyperspécialisation (la variabilité n'est plus sélectionnée puisque non nécessaire dans un environnement stable). C'est cette hypothèse que nous souhaitons tester par l'analyse du transcriptome de *Buchnera,* cellule à cellule, et en comparaison à celui d'*E. coli*. L'analyse du transcriptome d'une cellule bactérienne est encore un défi technologique important, néanmoins l'équipe de F. Chatelain (CEA, Grenoble) a développé des outils très prometteurs sur le sujet. Un projet qui sera soumis à l'ANR dans le cadre de l'appel d'offre « biologie des systèmes » est en cours d'écriture actuellement. Cette thématique ne sera pas abordée sur les moyens propres de l'équipe, nous demanderons le recrutement d'un PostDoc dans le cadre de l'appel d'offre.

Volet 2. Caractérisation du réseau métabolique de Buchnera

Des études théoriques sont actuellement réalisées pour caractériser le réseau métabolique de *Buchnera* en collaboration avec MF Sagot (LBBE, UCBL) dans le cadre de la thèse de L. Cottret (LBBE-BF2I, 2006). La structure du réseau sera analysée comparativement à *E. coli* et à d'autres bactéries intracellulaires comme les *Wolbachia* qui n'évoluent pas dans le même contexte symbiotique. Des modélisations de flux devraient être envisagées sur la bactérie pour étudier notamment la robustesse du réseau vis-à-vis de l'inactivation de certains gènes. Cette question est tout à fait fondamentale dans notre problématique car elle permet, d'une part, de tester la fonctionnalité de notre approche transcriptionnelle, et d'autre part, de tester *in silico* des stratégies de lutte contre les pucerons par l'inactivation de gènes clés du système symbiotique.

En fonction de la pertinence de cette analyse, des mesures de flux entrant et sortant, des mesures d'activités enzymatiques clés du système, ou du radiotraçage de métabolites pourront être élaborés pour confirmer certaines hypothèses importantes. Le laboratoire BF2I possède une très bonne expertise de ce genre d'expérimentation, notamment avec la participation de G. Febvay dans ce travail.

Globalement, cette analyse viendra enrichir notre connaissance des échanges entre les deux partenaires symbiotiques, de façon complémentaire à l'analyse transcriptomique. Elle constitue une étape supplémentaire vers l'approche de physiologie intégrée (ou biologie des systèmes) que nous souhaitons développer.

Volet 3. Caractérisation moléculaire des interactions trophiques entre le puceron hôte et Buchnera

Ce programme de recherche se focalisera sur la caractérisation du contrôle de la fonction symbiotique par l'hôte par la recherche de gènes, enzymes, ou voies clés gérant la complémentation nutritionnelle entre l'hôte et le symbiote. Il se concentrera sur le fonctionnement de la cellule bactériocytaire à l'aide des

techniques d'étude de l'expression génique eucaryote spécifique de ces cellules (dont une puce puceron disponible à l'INRA de Rennes). Il s'agira d'utiliser ces ressources pour étudier la réponse du tissu cible à des stimuli d'ordre trophique, du même type que ceux déjà réalisés dans nos précédents travaux. Un point fort sera l'intégration de ces résultats d'analyse transcriptomique dans une approche physiologique globale. L'analyse fonctionnelle des voies symbiotiques ciblées sera le prolongement naturel de cette approche, et s'appuiera sur le développement d'outils d'inactivation génique, notamment l'utilisation de la technique « ARN interférence » en cours de développement sur ce modèle. Ce travail ne sera pas abordé sans le recrutement d'un chargé de recherche sur cette thématique.

TROISIEME PARTIE

Activités d'enseignement

Chapitre 1

Enseignements au département de Biosciences

Mon recrutement en 1998 au laboratoire BF2I correspondait au remplacement de C. Voulot. J'ai donc repris en grande partie ses enseignements au département : l'anatomie du petit animal de laboratoire, l'entomologie, une partie du cours de statistique, ainsi qu'un bloc de travaux pratiques sur la culture *in vitro* des végétaux initialement développé par Brigitte de la Chapelle (MdC, INSA). Plus récemment, je me suis impliqué dans le projet de création de la filière « BioInformatique et Modélisation » du département Biosciences. J'ai alors pris en charge une partie plus importante du programme de statistique (le modèle linéaire) en troisième année. Je coordonne également un module de 50H sur l'analyse du transcriptome en fin de quatrième année. Mon implication en recherche dans le domaine de la bioinformatique m'a conduit à développer un module de bioinformatique en cinquième année pour la filière Biochimie et Biotechnologie (BB).

Anatomie pratique

Je me suis beaucoup impliqué dans cette série de travaux pratiques de troisième année (filière BB) dans les premières années qui ont suivi mon recrutement. J'ai essayé d'introduire des notions d'anatomie comparée pour replacer les observations dans un contexte d'évolution des structures et des fonctions. Ces travaux pratiques ont subi d'importantes réductions d'horaires et sont maintenant orientés vers une observation très physiologique grâce à l'implication de deux maîtres de conférences nouvellement recrutés au département, F. Calevro (responsable actuelle) et C. Soulage. Je ne participe désormais à ces travaux pratiques que pour l'observation

du système nerveux végétatif (ortho et parasympathique) chez le lapin (4 heures de TP).

Biodiversité, Ecologie et Evolution – Exemples chez les insectes

Ce module de troisième année (filière BB) dérive des travaux pratiques et cours d'entomologie initialement développés par P. Nardon, G. Plantevin et C. Voulot, enseignants du laboratoire BF2I. J'ai entièrement repris ce groupe d'enseignements qui, lui aussi, a subi des réductions d'horaires drastiques (44 heures en 2005). J'ai introduit quatre heures de cours et quatre heures de travaux pratiques de phylogénie moléculaire. Un chercheur du laboratoire intervient pour donner quelques notions sur la physiologie de l'insecte (respiration, excrétion, développement, reproduction) en insistant sur les différences avec le modèle mammifère. La dissection de la blatte illustre cette conférence. Deux séances sont consacrées à l'analyse de la biodiversité dans le groupe des arthropodes avec des travaux pratiques illustrant les différents niveaux taxonomiques. Huit heures sont consacrées à la protection des cultures sous la forme d'exposés par les étudiants et avec la participation de chercheurs du laboratoire. Enfin, une sortie sur le terrain (Château de Crussol, site Natura 2000) permet aux étudiants de discuter avec un professionnel des métiers de l'environnement, de la préservation et de l'aménagement des espaces naturels. C'est aussi, pour beaucoup d'entre eux, l'occasion de découvrir une nature qui leur était jusqu'à présent parfaitement inconnue. J'assure la majorité des enseignements de ce module, j'ai rédigé le cours polycopié correspondant, à partir des nombreux documents fournis par les anciens enseignants.

Transfert de gènes stables et transitoires chez Nicotiana tabacum

Ce module de travaux pratiques de quatrième année (filière BB) est issu de la combinaison entre les travaux pratiques de culture végétale *in vitro* (développé par B. de la Chapelle, MdC INSA) et les travaux pratiques de transgenèse développés par C. Dorel (MdC, INSA). Ces TP sont intégrés dans un module plus général de biologie moléculaire des eucaryotes. Dans ces travaux pratiques, j'ai repris les séances de transgenèse du tabac par la bactérie *Agrobacterium tumefasciens*, et j'ai monté des travaux pratiques de transformation transitoire de protoplastes de tabac. Ces séances sont également l'occasion de discussions sur les OGM avec les étudiants. Les évaluations de ces TP sont faites par des exposés sur les OGM incluant la thérapie génique et abordant les aspects scientifiques, économiques, sociologiques et éthiques associés à cette biotechnologie.

Statistiques, modèle linéaire

Dans les premières années qui ont suivi mon recrutement, j'ai assuré quelques enseignements de statistiques en quatrième année de la filière BB (20 heures) sur les

méthodes de régression linéaire. Depuis la création de la filière BiM, je suis en charge de tout le module de statistique linéaire (30 heures), de statistiques non paramétriques (12 heures) et d'une partie de la planification expérimentale (4 heures). Les travaux dirigés sont réalisés avec le logiciel R. J'ai rédigé une partie du cours polycopié, avec l'aide de JF Pageaux (MdC, INSA).

Analyse du transcriptome

Ce module de 50 heures a été créé *de novo* lors de la mise en place de la quatrième année de la filière BiM. J'ai participé à la définition du contenu pédagogique avec l'équipe enseignante et je coordonne le module. Celui-ci est organisé autour des techniques d'analyse du transcriptome utilisant la PCR (RT-PCR différentielle, PCR soustractive), le séquençage (EST, SAGE et MPSS) et l'hybridation (puces à ADN). Des travaux pratiques d'analyse des données SAGE (data mining) et puces à ADN (analyse différentielle) ont été développés. La modélisation des réseaux de régulation et les approches systémiques sont abordées. J'assure les cours et TP relatifs aux puces à ADN, des chercheurs prennent en charge les autres techniques, O. Gandrillon (CGMC, UCBLyon) et H. De Jong (INRIA, Grenoble) gèrent, par exemple, les parties SAGE et réseaux de régulation.

Bioinformatique dans la filière Biochimie et Biotechnologies

Suite au développement de la filière Bioinformatique, nous avons essayé de faire bénéficier la filière « traditionnelle » de notre expérience dans ce domaine. Un module de Bioinformatique a donc été créé. Je coordonne ce module (32 heures) et j'assure les enseignements en collaboration avec G. Beslon (LIRIS, INSA) et A. Rodrigues (UMGR, INSA). Ce module contient une introduction à l'architecture informatique, le système Linux, les technologies de l'Internet et une ouverture vers l'intelligence artificielle. Les travaux pratiques sont orientés sur la phylogénie moléculaire, les bases de données en génomique et protéomique et le travail à distance (sur un exemple d'optimisation de sondes oligonucléotidique pour les puces à ADN).

Perspectives

Je me suis toujours attaché à faire évoluer mes enseignements vers des notions plus modernes, correspondant au mieux aux métiers de l'ingénieur et du chercheur d'aujourd'hui. Je milite cependant souvent auprès des élèves, mais également auprès de certains collègues enseignants, pour conserver l'observation au cœur de nos travaux pratiques de biologie. Malheureusement, l'introduction de technologies de pointe, comme certaines techniques de biologie moléculaire « enkitée », la microscopie à fluorescence ou confocales et les nanotechnologies, par exemple, ne peu se faire qu'au détriment de séances d'observations « classiques », où l'étudiant

est face à sa dissection, son microscope de paillasse et doit tout « simplement » décrire sa préparation. Je fais ainsi bien souvent figure, au sein du département, de conservateur archaïque contre la modernisation des enseignements.

Les puces à ADN, exemple phare de technologie « à la mode », offrent un magnifique exemple de ce dilemme. L'application la plus industrielle et la plus développée des puces à ADN est le génotypage pour la détection de pathogènes, le contrôle sanitaire, la traçabilité des processus. Le génotypage est également appelé « taxonomie moléculaire ». L'enseignement de la taxonomie (Biodiversité en 3BB) fait partie des cours les plus controversés du département. Comment imaginer à l'heure actuelle que nos étudiants ingénieurs se lancent dans de tels projets sans avoir la moindre notion de biodiversité et de taxonomie ? L'application en recherche la plus développée pour les puces à ADN est l'analyse transcriptomique, outil essentiel de cette « biologie intégrative » dont tout le monde académique raffole. La biologie intégrative n'est rien d'autre que la mise dans le contexte physiologique des résultats acquis au niveau moléculaire. Il paraît donc très important de conserver et de renforcer les enseignements de physiologie du Département de Biosciences pour permettre à nos étudiants d'accéder à ces analyses. Ici encore, il s'agit de préserver un enseignement de l'observation (anatomie pratique) et l'acquisition de compétences de bases (cours de physiologie) qui me paraît beaucoup plus important que d'amener les étudiants à pratiquer ces techniques modernes pour la plupart robotisées.

Chapitre 2

Enseignements à l'Université : le master aMIV

Le Master aMIV (Approches Mathématiques et Informatique du Vivant, http://miv.univ-lyon1.fr) a été créé à l'occasion de la réforme LMD (septembre 2004), essentiellement sous l'impulsion des enseignants chercheurs du laboratoire LBBE de l'Université Claude Bernard de Lyon (Responsable S. Charles). L'INSA est cohabilité pour ce Master et j'en assume la responsabilité. L'offre de ce Master se décline en quatre parcours :

- Bio-Informatique Moléculaire (Resp. D. Mouchiroud et G. Deléage) ;
- Bio-Mathématiques et Pharmacologie (Resp. F. Gueyffier et P. Girard);
- Bio-Statistiques (Resp. P. Roy) ;
- Biométrie Appliquée (Resp. A.B. Dufour).

Les trois premiers parcours sont ouverts à nos étudiants des filières BiM et BB, soit en cinquième année (étudiants en double cursus), soit après la validation de leur diplôme d'ingénieur. J'ai participé à l'élaboration du contenu pédagogique de cette formation, et mon rôle dans ce master est principalement administratif (jury de diplôme) puisque je ne participe à l'enseignement que pour quatre heures de cours dans le parcours Bio-Informatique Moléculaire (analyse du transcriptome).

REFERENCES BIBLIOGRAPHIQUES

Adami, C. 1998. Introduction to artificial life. Springer-Verlag, New-York.

Aderem, A. 2005. Systems biology: its practice and challenges. *Cell* 121:511-513.

Akashi, H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* 136:927-935.

Al-Shahrour, F., R. Diaz-Uriarte and J. Dopazo. 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20:578-580.

Altschul, S., T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389-3402.

Andersson, S. G. and P. M. Sharp. 1996. Codon usage and base composition in *Rickettsia prowazekii*. *J Mol Evol* 42:525-536.

Aubert, J., A. Bar-Hen, J. J. Daudin and S. Robin. 2004. Determination of the differentially expressed genes in microarray experiments using local FDR. *BMC Bioinformatics* 5:125.

Baggerly, K. A., K. R. Coombes, K. R. Hess, D. N. Stivers, L. V. Abruzzo and W. Zhang. 2001. Identifying differentially expressed genes in cDNA microarray experiments. *J Comput Biol* 8:639-659.

Bagland, O. 2005. Etude du transcriptome chez des patients trisomiques. Pp. 30. Rapport de Master 2 : Probabilité statistique. Université Paris XI, Orsay (France).

Bailey, T. L. and C. Elkan. 1995. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* 3:21-29.

Bailey, T. L. and M. Gribskov. 1998. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* 14:48-54.

Baldi, P. and A. D. Long. 2001. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17:509-519.

Barriot, R., J. Poix, A. Groppi, A. Barre, N. Goffard, D. Sherman, I. Dutour and A. de Daruvar. 2004. New strategy for the representation and the integration of biomolecular knowledge at a cellular scale. *Nucleic Acids Res* 32:3581-3589.

Baumann, P., L. Baumann, C.-Y. Lai, D. Rouhbakhsh, N. A. Moran and M. A. Clark. 1995. Genetics, physiology and evolutionary relationships of the genus *Buchnera*: Intracellular symbionts of aphids. *Annu Rev Microbiol* 49:55-94.

Beau, J., F. Delaunay, S. Lacoche, A. Grechez-Cassiau and F. Levi. 2005. Algorithm of determination of circadian gene expression profiles analysed with DNA microarrays. *Pathol Biol (Paris)* 53:295-299.

Benjamini, Y. and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerfull approach to multiple testing. *JRSSB* 99:289-300.

Berka, R. M., X. Cui and C. Yanofsky. 2003. Genomewide transcriptional changes associated with genetic alterations and nutritional supplementation affecting tryptophan metabolism in *Bacillus subtilis*. *Proc Natl Acad Sci U S A* 100:5682-5687.

Bernot, G., J. P. Comet, A. Richard and J. Guespin. 2004. Application of formal methods to biological regulatory networks: extending Thomas' asynchronous logical approach with temporal logic. *J Theor Biol* 229:339-347.

Bilban, M., L. K. Buehler, S. Head, G. Desoye and V. Quaranta. 2002. Normalizing DNA microarray data. *Curr Issues Mol Biol* 4:57-64.

Black, M. A. and R. W. Doerge. 2002. Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics* 18:1609-1616.

Blangiardo, M., S. Toti, B. Giusti, R. Abbate, A. Magi, F. Poggi, L. Rossi, F. Torricelli and A. Biggeri. 2005. Using a calibration experiment to assess gene specific information: full Bayesian and empirical Bayesian models for two-channels microarray data. *Bioinformatics*. 22:50-57.

Boulicault, J.-F. and O. Gandrillon. 2004. Informatique pour l'analyse du transcriptome. Hermès sciences publication, Paris.

Bozdech, Z., J. Zhu, M. Joachimiak, F. Cohen, B. Pulliam and J. DeRisi. 2003. Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biology* 4:R9.

Braendle, C., T. Miura, R. Bickel, A. Shingleton, S. Kambhampati and D. Stern. 2003. Developmental origin and evolution of bacteriocytes in the Aphid-*Buchnera* symbiosis. *PLOS Biology* 1:70-76.

Brazhnik, P. 2005. Inferring gene networks from steady-state response to single-gene perturbations. *J Theor Biol* 237:427-440.

Broberg, P. 2003. Statistical methods for ranking differentially expressed genes. *Genome Biology* 4:R1.

Brown, L. E., I. Tsamardinos and C. F. Aliferis. 2004. A novel algorithm for scalable and accurate Bayesian network learning. *Medinfo* 11:711-715.

Buchner, P. 1965. Aphids. Pp. 297-332 *in* P. BUCHNER, ed. Endosymbiosis of animals with plant microorganisms. Interscience, New York (USA).

Calevro, F., H. Charles, N. Reymond, V. Dugas, J.-P. Cloarec, J. Bernillon, Y. Rahbe, G. Febvay and J.-M. Fayard. 2004. Assessment of 35mer amino-modified oligonucleotide based microarray with bacterial samples. *J Microbiol Methods* 57:207-218.

Charles, H. 1997. Aspects moléculaires de la bactérie symbiotique principale du charançon des céréales *Sitophilus oryzae* (Coléoptère, Curculionidae) et étude de ses interactions avec l'hôte. Pp. 132. Thèse de Doctorat. INSA de Lyon.

Charles, H., G. Condemine, C. Nardon and P. Nardon. 1997a. Genome size characterization of the principal endocellular symbiotic bacteria of the weevil *Sitophilus oryzae*, using pulsed field gel electrophoresis. *Insect Biochem Molec Biol* 27:345-350.

Charles, H., A. Heddi, J. Guillaud, C. Nardon and P. Nardon. 1997b. A molecular aspect of symbiotic interactions between the weevil *Sitophilus oryzae* and its endosymbiotic bacteria: over-expression of a chaperonin. *Biochem Biophys Res Com* 239:769-774.

Charles, H., A. Heddi and Y. Rahbé. 2001. A putative insect intracellular endosymbiont stem clade, within the Enterobacteriaceae, infered from phylogenetic analysis based on a heterogeneous model of DNA evolution. *C R Acad Sci Paris* 324:489-494.

Charles, H. and H. Ishikawa. 1999. Physical and genetical map of the genome of *Buchnera*, the primary endosymbiont of the pea aphid *Acyrthosiphon pisum*. *J Mol Evol* 48:142-150.

Charles, H., H. Ishikawa and P. Nardon. 1995. Presence of a protein specific of endocytobiosis (symbionin) in the weevil *Sitophilus*. *C R Acad Sci Paris* 318:35-41.

Chen, H. and G. Zhu. 1997. Computer program for calculation the melting temperature of de generate oligonucleotides used in PCR. *Biotechniques* 22:1158-1160.

Chen, Y., V. Kamat, E. Dougherty, L. Bittner, P. Meltzer and J. Trent. 2002. Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics* 18:1207-1215.

Churchill, G. A. 2002. Fundamentals of experimental design for cDNA microarrays. *Nat Genet* 32:490-495.

Colantuoni, C., G. Henry, S. Zeger and J. Pevsner. 2002. Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. *Biotechniques* 32:1316-1320.

Croes, D., F. Couche, S. J. Wodak and J. van Helden. 2005. Inferring Meaningful Pathways in Weighted Metabolic Networks. *J Mol Biol*.

Cui, X. and G. Churchill. 2003. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* 4:210.

Cui, X., J. T. Hwang, J. Qiu, N. J. Blades and G. A. Churchill. 2005. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* 6:59-75.

Cui, X., K. Kerr and G. Churchill. 2002. Data transformation for cDNA microarray data. http://www.jax.org/staff/churchill/labsite/pubs/index.html.

D'Ambrosio, C., L. Gatta and S. Bonini. 2005. The future of microarray technology: networking the genome search. *Allergy* 60:1219-1226.

Dagnélie, P. 1980. Théorie et méthodes statistiques, De Boeck.

De Gregorio, E., P. Spellman, G. Rubin and B. Lemaitre. 2001. Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays. *Proc Natl Acad Sci USA* 98:12590-12595.

de Hoon, M. J., Y. Makita, K. Nakai and S. Miyano. 2005. Prediction of Transcriptional Terminators in *Bacillus subtilis* and Related Species. *PLoS Comput Biol* 1:e25.

De Jong, H., J. Geiselmann, C. Hernandez and M. Page. 2003. Genetic Network Analyzer: qualitative simulation of genetic regulatory networks. *Bioinformatics* 19:336-344.

Delmar, P. 2005. Modèle de mélange sur la variance pour l'analyse différentielle des biopuces. Pp. 165. Thèse de Doctorat. Ecole Centrale Paris, Paris (France).

Delmar, P., S. Robin and J. J. Daudin. 2005. VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics* 21:502-508.

DeRisi, J., L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su and J. M. Trent. 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 14:457-460.

Didier, G., P. Brezellec, E. Remy and A. Henaut. 2002. GeneANOVA--gene expression analysis of variance. *Bioinformatics* 18:490-491.

Dittmar, K., E. Mobley, A. Radek and T. Pan. 2004. Exploring the regulation of tRNA distribution on the genomic slide. *J Mol Biol* 337:31-47.

Dixon, A. F. G., F. A. Klingauf and K. Kawada. 1987. World crop pests, Aphids their biology, natural enemies and control, volume A. Amsterdam: Elsevier.

Dobbin, K., E. S. Kawasaki, D. W. Petersen and R. M. Simon. 2005. Characterizing dye bias in microarray experiments. *Bioinformatics* 21:2430-2437.

Dobbin, K., J. Shih and R. Simon. 2003a. Statistical design of reverse dye microarrays. *Bioinformatics* 19:803-810.

Dobbin, K., J. Shih and R. Simon. 2003b. Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. *J Natl Cancer Inst* 95:1362-1369.

Dorman, C. J. and P. Deighan. 2003. Regulation of gene expression by histone-like proteins in bacteria. *Curr Opin Genet Dev* 13:179-184.

Douglas, A. E. 1998. Nutritional interactions in insect-microbial symbioses: Aphids and their symbiotic bacteria *Buchnera*. *Annu Rev Entomol* 43:17-37.

Douglas, A. E. 2003. *Buchnera* bacteria and other symbionts of aphids. Pp. 23-38. Insect Symbiosis. CRC Press Inc, Boca Raton.

Douglas, A. E., L. B. Minto and T. L. Wilkinson. 2001. Quantifying nutrient production by the microbial symbionts in an aphid. *J exp Biol* 204:349-358.

Draghici, S., A. Kuklin, B. Hoff and S. Shams. 2001. Experimental design, analysis of variance and slide quality assessement in gene expression arrays. *Curr Opin Drug Discov and Devel* 4:332-337.

Draghici, S., O. Kulaeva, B. Hoff, A. Petrov, S. Shams and M. Tainsky. 2003. Noise sampling: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. *Bioinformatics* 19:1348-1359.

Dudley, A. M., J. Aach, M. A. Steffen and G. M. Church. 2002. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc Natl Acad Sci USA* 99:7554-7559.

Dudoit, S., J. Shaffer and C. Boldrick. 2003a. Multiple hypothesis testing in microarray experiments. *Statistical Sciences* 18:71-103.

Dudoit, S., M. van der Laan and K. S. Pollard. 2003b. Multiple testing. Part I. Single step procedures for control of general type error rates. *U.C. Berkeley Division of Biostatistics working paper series (138)*.

Dudoit, S., Y. H. Yang, M. J. Callow and T. Speed. 2002. Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. *Stat Sin* 12:111-139.

Durbin, B. P., J. S. Hardin, D. M. Hawkins and D. M. Rocke. 2002. A variance-stabilizing transformation for gene expression microarray data. *Bioinformatics* 18:S105-S110.

Edwards, D. 2003. Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics* 19:825-833.

Edwards-Ingram, L. C., M. E. Gent, D. C. Hoyle, A. Hayes, L. I. Stateva and S. G. Oliver. 2004. Comparative genomic hybridization provides new insights into the molecular taxonomy of the *Saccharomyces* sensu stricto complex. *Genome Res* 14:1043-1051.

Efron, B., R. Tibshirani, J. Storey and V. Tusher. 2001. Empirical Bayes analysis of a microarray experiment. *J Am Statist Ass* 96:1151-1160.

Eickhoff, B., B. Korn, M. Schick, A. Poustka and J. van der Bosch. 1999. Normalization of array hybridization experiments in differential gene expression analysis. *Nucleic Acids Res* 27:e33.

Ekstrom, C. T., S. Bak, C. Kristensen and M. Rudemo. 2004. Spot shape modelling and data transformations for microarrays. *Bioinformatics* 20:2270-2278.

Elowitz, M. B., A. J. Levine, E. D. Siggia and P. S. Swain. 2002. Stochastic gene expression in a single cell. *Science* 297:1183-1186.

Emilsson, V., A. K. Naslund and C. G. Kurland. 1993. Growth-rate-dependent accumulation of twelve tRNA species in *Escherichia coli*. *J Mol Biol* 230:483-491.

Fares, M. A., E. Barrio, B. Sabater-Munoz and A. Moya. 2002. The evolution of the heat-shock protein GroEL from *Buchnera*, the primary endosymbiont of aphids, is governed by positive selection. *Mol Biol Evol* 19:1162-1170.

Febvay, G., B. Delobel and Y. Rahbé. 1988. Influence of the amino acid balance on the improvement of an artificial diet for a biotype of *Acyrthosiphon pisum* (Homoptera : Aphididae). *Can J Zool* 66:2449-2453.

Febvay, G., I. Liadouze, J. Guillaud and G. Bonnot. 1995. Analysis of energetic aminoacid metabolism in *Acyrthosiphon pisum* : a multidimensional approach to amino acid metabolism in aphids. *Arch Insect Biochem* 29:45-69.

Febvay, G., Y. Rahbe, M. Rynkiewicz, J. Guillaud and G. Bonnot. 1999. Fate of dietary sucrose and neosynthesis of amino acids in the pea aphid, *Acyrthosiphon pisum*, reared on different diets. *J Exp Biol* 202:2639-2652.

Feuerstein, B. G., L. D. Williams, H. S. Basu and L. J. Marton. 1991. Implications and concepts of polyamine-nucleic acid interactions. *J Cell Biochem* 46:37-47.

Fielden, M., R. Halgren, E. Dere and T. Zacharewski. 2002. GP3: GenePix post-processing program for automated analysis of raw microarray data. *Bioinformatics* 18:771-773.

Finkelstein, D., J. Gollub and J. M. Cherry. 2002a. Normalization and systematic measurement error in cDNA microarray data. *http://afgc.stanford.edu/afgc_html/site2Stat.htm*.

Finkelstein, D., J. Gollub, R. Ewing, F. Sterky, S. Somerville and J. M. Cherry. 2002b. Iterative linear regression by sector: renormalization of cDNA microarray data and cluster analysis weighted by cross homology. *http://afgc.stanford.edu/afgc_html/site2Stat.htm*.

Freier, S. M., R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson and D. H. Turner. 1986. Improved free-energy parameters for predictions of RNA duplex stability. *Proc Natl Acad Sci USA* 83:9373-9377.

Freundlich, M., N. Ramani, E. Mathew, A. Sirko and P. Tsui. 1992. The role of integration host factor in gene expression in *Escherichia coli*. *Mol Microbiol* 6:2557-2563.

Fuglsang, A. 2004. The relationship between palindrome avoidance and intragenic codon usage variations: a Monte Carlo study. *Biochem Biophys Res Commun* 316:755-762.

Fukatsu, T. and H. Ishikawa. 1998. Differential immunohistochemical visualization of the primary and secondary intracellular symbiotic bacteria of aphids. *Appl Entomol Zool* 33:321-326.

Fukatsu, T., N. Nikoh, R. Kawai and R. Koga. 2000. The secondary endosymbiotic bacterium of the pea aphid *Acyrthosiphon pisum* (Insecta, Homoptera). *Appl Environ Microbiol* 66:2748-2758.

Ge, Y., S. Dudoit and T. Speed. 2003. Resampling-based multiple testing for microarray data analysis. *Test* 12:1-77.

Georgii, E., L. Richter, U. Ruckert and S. Kramer. 2005. Analyzing microarray data using quantitative association rules. *Bioinformatics* 21 Suppl 2:ii123-ii129.

Ghosh, D. and A. Chinnaiyan. 2002. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics* 18:275-286.

Gil, R., B. Sabater-Munoz, A. Latorre, F. J. Silva and A. Moya. 2002. Extreme genome reduction in *Buchnera* spp.: Toward the minimal genome needed for symbiotic life. *Proc Natl Acad Sci USA* 99:4454-4458.

Glonek, G. F. and P. J. Solomon. 2004. Factorial and time course designs for cDNA microarray experiments. *Biostatistics* 5:89-111.

Gomez-Valero, L., M. Soriano-Navarro, V. Perez-Brocal, A. Heddi, A. Moya, J. M. Garcia-Verdugo and A. Latorre. 2004. Coexistence of *Wolbachia* with *Buchnera aphidicola* and a secondary symbiont in the aphid *Cinara cedri*. *J Bacteriol* 186:6626-6633.

Gourse, R. L., H. A. de Boer and M. Nomura. 1986. DNA determinants of rRNA synthesis in *E. coli*: growth rate dependent regulation, feedback inhibition, upstream activation, antitermination. *Cell* 44:197-205.

Grantham, R., C. Gautier, M. Gouy, M. Jacobzone and R. Mercier. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res* 9:r43-74.

Grosjean, H. J., S. de Henau and D. M. Crothers. 1978. On the physical basis for ambiguity in genetic coding interactions. *Proc Natl Acad Sci USA* 75:610-614.

Guespin-Michel, J. F., G. Bernot, J. P. Comet, A. Merieau, A. Richard, C. Hulen and B. Polack. 2004. Epigenesis and dynamic similarity in two regulatory networks in *Pseudomonas aeruginosa*. *Acta Biotheor* 52:379-390.

Guo, X., H. Qi, C. M. Verfaillie and W. Pan. 2003. Statistical significance analysis of longitudinal gene expression data. *Bioinformatics* 19:1628-1635.

Guschin, D. Y., B. K. Mobarry, D. Proudnikov, D. A. Stahl, B. E. Rittman and A. D. Mirzabekov. 1997. Oligonucleotide microchips as genosensors for determinative studies in microbiology. *Appl Env Microbiol* 63:2397-2402.

Hall, D., H. Zhu, X. Zhu, T. Royce, M. Gerstein and M. Snyder. 2004. Regulation of gene expression by a metabolic enzyme. *Science* 306:482-484.

Hanisch, D., A. Zien, R. Zimmer and T. Lengauer. 2002. Co-clustering of biological networks and gene expression data. *Bioinformatics* 18 Suppl 1:S145-154.

Heddi, A., H. Charles, C. Khatchadourian, G. Bonnot and P. Nardon. 1998. Molecular characterization of the principal symbiotic bacteria of the weevil *Sitophilus oryzae*: a peculiar G + C content of an endocytobiotic DNA. *J Mol Evol* 47:52-61.

Herbeck, J. T., P. H. Degnan and J. J. Wernegreen. 2005. Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the Enterobacteriales (gamma-Proteobacteria). *Mol Biol Evol* 22:520-532.

Hertz, G. Z. and G. D. Stormo. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15:563-577.

Hommais, F., E. Krin, C. Laurent-Winter, O. Soutourina, A. Malpertuy, J. P. Le Caer, A. Danchin, and P. Bertin. 2001. Large-scale monitoring of pleiotropic regulation of gene expression by the prokaryotic nucleoid-associated protein, H-NS. *Mol Microbiol* **40**:20-36.

Hoppe, I., H. M. Johnston, D. Biek and J. R. Roth. 1979. A refined map of the *hisG* gene of *Salmonella typhimurium*. *Genetics* 92:17-26.

Houk, E. J., G. W. Griffiths and S. D. Beck. 1976. Lipid metabolism of the primary symbiotes of the pea aphid *Acyrthosiphon pisum*. *Comp Biochem Physiol* 54B:427-431.

Houk, E. J., G. W. Griffiths, N. E. Hadjokas and S. D. Beck. 1977. Peptidoglycan in the cell wall of the primary intracellular symbiote of the pea aphid. *Science* 198:401-403.

Hsiao, A., T. Ideker, J. M. Olefsky and S. Subramaniam. 2005. VAMPIRE microarray suite: a web-based platform for the interpretation of gene expression data. *Nucleic Acids Res* 33:W627-632.

Hsiao, A., D. S. Worrall, J. M. Olefsky and S. Subramaniam. 2004. Variance-modeled posterior inference of microarray data: detecting gene-expression changes in 3T3-L1 adipocytes. *Bioinformatics* 20:3108-3127.

Huber, W., A. Von Heydebreck, H. Sueltmann, A. Poustka and M. Vingron. 2003. Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology* 2:3.1-3.22.

Huber, W., A. Von Heydebreck, H. Sultmann, A. Poustka and M. Vingron. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18:S96-S104.

Huentelman, M. J., D. W. Craig, A. D. Shieh, J. J. Corneveaux, D. Hu-Lince, J. V. Pearson and D. A. Stephan. 2005. SNiPer: improved SNP genotype calling for Affymetrix 10K GeneChip microarray data. *BMC Genomics* 6:149.

Humphreys, N. J. and A. E. Douglas. 1997. Partitioning of symbiotic bacteria between generations of insect: a quantitative study of a *Buchnera sp.* in the pea aphid (*Acyrthosiphon pisum*) reared at different temperatures. *Appl Environ Microbiol* 63:3294-3296.

Ikemura, T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol* 151:389-409.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* 2:13-34.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf and T. P. Speed. 2003. Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249-264.

Ishikawa, H. 1987. Nucleotide composition and Kinetic complexity of the genomic DNA of an intracellular symbiont in the pea aphid *Acyrthosiphon pisum*. *J Mol Evol* 24:205-211.

Ishikawa, H. 1982a. DNA, RNA and protein synthesis in the isolated symbionts from the pea aphid, *Acyrthosiphon pisum*. *Insect Biochem* 12:605-612.

Ishikawa, H. 1982b. Host-symbiont interactions in the protein synthesis in the pea aphid, *Acyrthosiphon pisum*. *Insect Biochem* 12:613-622.

Jansen, R., H. J. Bussemaker and M. Gerstein. 2003. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Res* 31:2242-2251.

Jensen, L., M. Langaas, W. P. Kuo, B. Smith-Sorensen, O. Myklebost and E. Hovig. 2002. Analysis of repeatability in spotted cDNA microarrays. *Nuc Acid Res* 30:2335-3244.

Jin, W., R. Riley, R. Wolfinger, K. White, G. Passador-Gurgel and G. Gibson. 2001. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat Genet* 29:389-395.

Kafadar, K. and T. Phang. 2003. Transformations, background estimation and process effects in the statistical analysis of microarrays. *Comput Stat Data An* 44:313-338.

Kanaya, S., Y. Yamada, Y. Kudo and T. Ikemura. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* 238:143-155.

Karley, A. J., A. E. Douglas and W. E. Parker. 2002. Amino acid composition and nutritional quality of potato leaf phloem sap for aphids. *J Exp Biol* 205:3009-3018.

Kelly, A., M. D. Goldberg, R. K. Carroll, V. Danino, J. C. Hinton and C. J. Dorman. 2004. A global role for Fis in the transcriptional control of metabolism and type III secretion in *Salmonella enterica serovar Typhimurium*. *Microbiology* 150:2037-2053.

Kepler, T. B., L. Crosby and K. T. Morgan. 2002. Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biol* 3:RESEARCH0037.

Kerr, A. R., J. F. Peden and P. M. Sharp. 1997. Systematic base composition variation around the genome of *Mycoplasma genitalium*, but not *Mycoplasma pneumoniae*. *Mol Microbiol* 25:1177-1179.

Kerr, M. K., C. A. Afshari, L. Bennett, P. Bushel, J. Martinez, N. J. Walker and G. A. Churchill. 2002. Statistical analysis of a gene expression microarray experiment with replication. *Stat Sin* 12:203-217.

Kerr, M. K. and G. A. Churchill. 2001a. Statistical design and the analysis of gene expression microarray data. *Genet Res* 77:123-128.

Kerr, M. K. and G. A. Churchill. 2001b. Experimental design for gene expression microarrays. *Biostatistics* 2:183.

Kerr, M. K., M. Martin and G. A. Churchill. 2000. Analysis of variance for gene expression microarray data. *J Comput Biol* 7:819-837.

Kim, R. and P. Park. 2004. Improving identification of differentially expressed genes in microarray studies using information from public databases. *Genome Biol* 5:R70.

Kirschner, M. W. 2005. The meaning of systems biology. *Cell* 121:503-504.

Knibbe, C., G. Beslon, V. Lefort, F. Chaudier, J.M. Fayard. 2005 Self-adaptation of genome size in artificial organisms.(eds) Copcassere et al., Advances in artificial lifes, Proceedings of the 8th European Conference, ECAL 2005, LNAI3630, Springer, pp. 423-432.

Koga, R., T. Tsuchida and T. Fukatsu. 2003. Changing partners in an obligate symbiosis: a facultative endosymbiotnt can compensate for loss of the essential endosymbiont *Buchnera* in an aphid. *Proc R Soc Lond B* 270:2543-2550.

Komaki, K. and H. Ishikawa. 2000. Genomic copy number of intracellular bacterial symbionts of aphids varies in response to developmental stage and morph of their host. *Insect Biochem Molec Biol* 30:253-258.

Komaki, K. and H. Ishikawa. 1999. Intracellular bacterial symbionts of aphids possess many genomic copies per bacterium. *J Mol Evol* 48:717-722.

Koonin, E. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol* 1:127-136.

Kooperberg, C., T. G. Fazzio, J. J. Delrow and T. Tsukiyama. 2002. Improved background correction for spotted DNA microarrays. *J Comput Biol* 9:55-66.

Kupiec, J.J. et P. Sonigo. 2000. Ni dieu ni gène. (eds) Seuil edition. pp. 229.

Laforest, F., A. Tchounikine, T. Chaari, H. Charles and F. Calevro. 2005. SITRANS: a web information system for microarray experiments. XIX International Congress of the European Federation for Medical Informatics, Genève (Suisse), août-septembre 2005.

Lambert, J. D. and N. A. Moran. 1998. Deleterious mutations destabilize ribosomal RNA in endosymbiotic bacteria. *Proc Natl Acad Sci USA* 95:4458-4462.

Lee, K. E., N. Sha, E. R. Dougherty, M. Vannucci and B. K. Mallick. 2003. Gene selection: a Bayesian variable selection approach. *Bioinformatics* 19:90-97.

Lee, M. L., F. C. Kuo, G. A. Whitmore and J. Sklar. 2000. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci U S A* 97:9834-9839.

Lefèvre, C. 2004. Caractérisation et phylogénie des bactéries symbiotiques intracellulaires des charançons de la famille des Dryophthoridae. Pp. 143. Thèse de Doctorat. INSA de Lyon.

Lefèvre, C., H. Charles, A. Vallier, B. Delobel, B. Farrell and A. Heddi. 2004. Endosymbiont phylogenesis in the Dryophthoridae weevils: evidence for bacterial replacement. *Mol Biol Evol* 21:965-973.

Legay, J. M. 1973. La méthode des modèles, état actuel de la méthode expérimentale. Informatique et Biosphère, Paris (France).

Lehnen, D., C. Blumer, T. Polen, B. Wackwitz, V. F. Wendisch and G. Unden. 2002. LrhA as a new transcriptional key regulator of flagella, motility and chemotaxis genes in *Escherichia coli*. *Mol Microbiol* 45:521-532.

Liadouze, I., G. Febvay, J. Guillaud and G. Bonnot. 1995. Effect of diet on the free amino acid pools of symbiotic and aposymbiotic pea aphids, *Acyrthosiphon pisum*. *J Insect Physiol* 41:33-40.

Liang, Y., B. Tayo, X. Cai and A. Kelemen. 2005. Differential and trajectory methods for time course gene expression data. *Bioinformatics* 21:3009-3016.

Lin, C. Y., A. Strom, V. B. Vega, S. L. Kong, A. L. Yeo, J. S. Thomsen, W. C. Chan, B. Doray, D. K. Bangarusamy, A. Ramasamy, L. A. Vergara, S. Tang, A. Chong, V. B. Bajic, L. D. Miller, J. A. Gustafsson and E. T. Liu. 2004. Discovery of estrogen receptor alpha target genes and response elements in breast tumor cells. *Genome Biol* 5:R66.

Lin, Y., P. Reynolds and E. Feingold. 2003. An empirical Bayesian method for differential expression studies using one-channel microarray data. *Stat Appl Genet Mol Biol* 2:1-16.

Lindroos, K., S. Sigurdsson, K. Johansson, L. Ronnblom and A. C. Syvanen. 2002. Multiplex SNP genotyping in pooled DNA samples by a four-colour microarray system. *Nucleic Acids Res* 30:E70.

Liu, E. T. 2005. Systems biology, integrative biology, predictive biology. *Cell* 121:505-506.

Lobry, J. R. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* 13:660-665.

Lönnstedt, I. and T. Speed. 2002. Replicated microarray data. *Stat Sinica* 12:31-46.

Luan, Y. and H. Li. 2004. Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics* 20:332-339.

Lynch, T. W., E. K. Read, A. N. Mattis, J. F. Gardner and P. A. Rice. 2003. Integration host factor: putting a twist on protein-DNA recognition. *J Mol Biol* 330:493-502.

Majander, K., L. Anton, J. Antikainen, H. Lang, M. Brummer, T. K. Korhonen and B. Westerlund-Wikstrom. 2005. Extracellular secretion of polypeptides using a modified *Escherichia coli* flagellar secretion apparatus. *Nat Biotechnol* 23:475-481.

Martin, O. 2002. Approches statistiques pour les données de puces à ADN. Pp. 147. Thèse de Doctorat. Université Joseph Fourier, Grenoble (France).

Martin-Magnette, M. L. and S. Robin. 2004. Techniques statistiques pour l'analyse du transcriptome. Boulicaut J.-F. et Gandrillon, O. Informatique pour l'analyse du transcriptome (traité IC2). Hermès sciences publication, Paris.

Mary-Huard, T., J. J. Daudin, S. Robin, F. Bitton, E. Cabannes and P. Hilson. 2004. Spotting effect in microarray experiments. *BMC Bioinformatics* 5:63.

Mazurie, A., S. Bottani and M. Vergassola. 2005. An evolutionary and functional assessment of regulatory network motifs. *Genome Biol* 6:R35.

McNulty, D. E., B. A. Claffee, M. J. Huddleston, M. L. Porter, K. M. Cavnar and J. F. Kane. 2003. Mistranslational errors associated with the rare arginine codon CGG in *Escherichia coli*. *Protein Express Purif* 27:365-374.

Minks, A. and P. Harrewijn. 1987. World Crop Pests. Elsevier.

Misra, R. V., R. S. Horler, W. Reindl, Goryanin, II and G. H. Thomas. 2005. EchoBASE: an integrated post-genomic database for *Escherichia coli*. *Nucleic Acids Res* 33:D329-333.

Moran, N. and A. Mira. 2001. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biol* 2:1-12.

Moran, N. A. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci USA* 93:2873-2878.

Moran, N. A. 2002. Microbial minimalism: genome reduction in bacterial pathogens. *Cell* 108:583-586.

Moran, N. A. 2003. Tracing the evolution of gene loss in obligate bacterial symbionts. *Curr Opin Microbiol* 6:512-518.

Moran, N. A., Dunbar, H. E. and Wilcox, J. L. 2005. Regulation of transcription in a reduced bacterial genome: nutrient-provisioning genes of the obligate symbiont *Buchnera aphidicola*. *J. Bacteriol.* 187, 4229-37.

Moran, N. A., C. D. Vondohlen and P. Baumann. 1995. Faster evolutionary rates in endosymbiotic bacteria than in cospeciating insect hosts. *J Mol Evol* 41:727-731.

Moreno-Hagelsieb, G., V. Trevino, E. Perez-Rueda, T. F. Smith and J. Collado-Vides. 2001. Transcription unit conservation in the three domains of life: a perspective from *Escherichia coli*. *Trends Genet* 17:175-177.

Morin, N. 2004. Analyse de l'expression des gènes chez *Buchnera*, bactérie symbiotique des pucerons, par l'utilisation d'une puce à ADN dédiée. Pp. 30. Rapport de DEA. INSA de Lyon.

Munson, M., P. Baumann and M. Kinsey. 1991. *Buchnera gen. nov*. and *Buchnera aphidicola sp. nov.*, a taxon consisting of the mycetocyte-associated, primary endosymbionts of aphids. *Int J Syst Bact* 41:566-568.

Nadon, R. and J. Shoemaker. 2002. Statistical issues with microarrays: processing and analysis. *Trends Genet* 18:265-271.

Nakabachi, A. and H. Ishikawa. 2000. Polyamine composition and expression of genes related to polyamine biosynthesis in an aphid endosymbiont, *Buchnera*. *Appl Environ Microbiol* 66:3305-3309.

Nardon, P. and H. Charles. 2001. Morphological aspects of symbiosis. Pp. 13-44 *in* J. Seckbach, ed. Cellular Origin and Life in Extreme Habitats. "Symbiosis, Mechanisms and Model Systems". Kluwer Academic Publishers, Dordrecht, Netherlands.

Nardon, P., H. Charles, B. Delobel, A. Lambert, C. Nardon and A. Heddi. 2000. Biodiversity of symbiosis in the *Rhynchophoridae* weevils. XVIIIth International Congress of Zoology, Athènes (Grèce), août-septembre 2000.

Nardon, P., H. Charles, B. Delobel, C. Lefèvre and A. Heddi. 2003. Symbiosis in the Dryophthoridae weevils (Coleoptera, Curculionoidea): morphological variability of symbiotic intracellular bacteria. *Symbiosis* 34:231-251.

Nardon, P. and A. M. Grenier. 1993. Symbiose et évolution. *Ann Soc Entomol Fr* 29:113-140.

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

Nitschke, P., P. Guerdoux-Jamet, H. Chiapello, G. Faroux, C. Henaut, A. Henaut and A. Danchin. 1998. Indigo: a World-Wide-Web review of genomes and gene functions. *FEMS Microbiol Rev* 22:207-227.

Occhialini, A., S. Cunnac, N. Reymond, S. Genin and C. Boucher. 2005. Genome-wide analysis of gene expression in *Ralstonia solanacearum* reveals that the *hrpB* gene acts as a regulatory switch controlling multiple virulence pathways. *Mol Plant Microbe Interact* 18:938-949.

Ochman, H., J. G. Lawrence and E. A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299-304.

Oliver, K. M., N. A. Moran and M. S. Hunter. 2005. Variation in resistance to parasitism in aphids is due to symbionts not host genotype. *Proc Natl Acad Sci USA* 102:12795-12800.

Paldi, A. 2003. Stochastic gene expression during cell differentiation: order from disorder? *Cell Mol Life Sci* **60**:1775-1778.

Pan, K. H., C. J. Lih and S. N. Cohen. 2002. Analysis of DNA microarrays using algorithms that employ rule-based expert knowledge. *Proc Natl Acad Sci U S A* 99:2118-2123.

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

Pan, W. 2003. On the use of permutation and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics* 19:1333-1340.

Pan, W., J. Lin and C. T. Le. 2003. A mixture model approach to detecting differentially expressed genes with microarray data. *Funct Integr Genomics* 3:117-124.

Panina, E. M., A. G. Vitreschak, A. A. Mironov and M. S. Gelfand. 2003. Regulation of biosynthesis and transport of aromatic amino acids in low-GC Gram-positive bacteria. *FEMS Microbiol Lett* 222:211-220.

Pavlidis, P., Q. Li and W. Noble. 2003. The effect of replication on gene expression microarray experiments. *Bioinformatics* 19:1620-1627.

Pe'er, D. 2005. Bayesian network analysis of signaling networks: a primer. *Sci STKE* 2005:pl4.

Peden, J. F. 1999. Analysis of codon usage. Pp. 215. Thèse de Doctorat. University of Nottingham, Nottingham (Angleterre).

Peeters, J. K. and P. J. Van der Spek. 2005. Growing applications and advancements in microarray technology and analysis tools. *Cell Biochem Biophys* 43:149-166.

Peng, X., R. K. Karuturi, L. D. Miller, K. Lin, Y. Jia, P. Kondu, L. Wang, L. S. Wong, E. T. Liu, M. K. Balasubramanian and J. Liu. 2005. Identification of cell cycle-regulated genes in fission yeast. *Mol Biol Cell* 16:1026-1042.

Perrin, B. E., L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet and F. D'Alche-Buc. 2003. Gene networks inference using dynamic Bayesian networks. *Bioinformatics* 19 Suppl 2:II138-II148.

Prickett, M.D., D. Page, A.E. Douglas, G.H. Thomas. 2006. BuchneraBASE: a post genomic resource for *Buchnera* sp. APS. Bioinformatics (sous presse).

Pritchard, C., L. Hsu, J. Delrow and P. Nelson. 2001. Project normal: Defining normal variance in mouse gene expression. *Proc Natl Acad Sci USA* 98:13266-13271.

Prosser, W. A., S. J. Simpson and A. E. Douglas. 1992. How an aphid symbiosis responds to variation in dietary nitrogen. *J Insect Physiol* 38:301-307.

Rahbé, Y., G. Febvay, B. Delobel and G. Bonnot. 2000. Amino acids and proteins as cues in the interactions of Aphids (Homoptera: Aphididae) and plants. Pp. 212-236 *in* G. P. Walker and E. A. Backus, eds. Principles and applications of electronic monitoring and other techniques in the study of Homopteran feeding behavior. Thomas Say Publ., USA.

Ramdas, L., K. R. Coombes, K. Baggerly, L. Abruzzo, W. E. Highsmith, T. Krogmann, S. R. Hamilton and W. Zhang. 2001. Sources of nonlinearity in cDNA microarray expression measurements. *Genome Biol* 2:RESEARCH0047.

Reiner, A., D. Yekutieli and Y. Benjamini. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19:368-375.

Reis, M., L. Wernisch and R. Savva. 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Bioinformatics* 31:6976-6985.

Reymond, N. 2004. Bioinformatique des puces à ADN et application à l'analyse du transcriptome de *Buchnera aphidicola*. Pp. 323. Thèse de Doctorat. INSA de Lyon.

Reymond, N., H. Charles, L. Duret, F. Calevro, G. Beslon and J. M. Fayard. 2004a. ROSO: optimizing oligonucleotide probes for microarrays. *Bioinformatics* 20:271-273.

Reymond, N., H. Charles, S. Rome and J. Marty. 2004b. Les données d'expression. Boulicaut J.-F. et Gandrillon, O. Informatique pour l'analyse du transcriptome (traité IC2). Hermès sciences publication, Paris.

Rhodes, J. D., P. C. Croghan and A. F. G. Dixon. 1997. Dietary sucrose and oligosaccharide synthesis in relation to osmoregulation in the pea aphid, *Acyrthosiphon pisum*. *Physiol Entomol* 22:373-379.

Rispe, C. 2005. Sexualité, clonalité et évolution du génome, chez les pucerons et leurs bactéries symbiotiques. Pp. 93. Université de Rennes 1, Rennes (France).

Rispe, C., F. Delmotte, R. C. van Ham and A. Moya. 2004. Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. *Genome Res* 14:44-53.

Rocha, E. and A. Danchin. 2003a. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res* 31:6570-6577.

Rocha, E. and A. Danchin. 2003b. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet* 34:377-378.

Rocha, E. P. and A. Danchin. 2002. Base composition bias might result from competition for metabolic resources. *Trends Genet* 18:291-294.

Rocke, D. M. and B. Durbin. 2001. A model for measurement error for gene expression arrays. *J Comput Biol* 8:557-569.

Rodionov, D. A., M. S. Gelfand and N. Hugouvieux-Cotte-Pattat. 2004. Comparative genomics of the KdgR regulon in *Erwinia chrysanthemi* 3937 and other gamma-proteobacteria. *Microbiology* 150:3571-3590.

Rosa, G. J. M., J. P. Steibel and R. J. Tempelman. 2005. Reassessing design and analysis of two-colour microarray experiments using mixed effects models. *Comp Funct Genomics* 6:123-131.

Sakurai, M., R. Koga, T. Tsuchida, X. Y. Meng and T. Fukatsu. 2005. Rickettsia symbiont in the pea aphid *Acyrthosiphon pisum*: Novel cellular tropism, effect on host fitness and interaction with the essential symbiont *Buchnera*. *Appl Environ Microbiol* 71:4069-4075.

Salgado, H., S. Gama-Castro, A. Martinez-Antonio, E. Diaz-Peredo, F. Sanchez-Solano, M. Peralta-Gil, D. Garcia-Alonso, V. Jimenez-Jacinto, A. Santos-Zavaleta, C. Bonavides-Martinez and J. Collado-Vides. 2004. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res* 32:D303-306.

Sandstrom, J., A. Telang and N. A. Moran. 2000. Nutritional enhancement of host plants by aphids - a comparison of three aphid species on grasses. *J Insect Physiol* 46:33-40.

Sandstrom, J. P. and N. A. Moran. 2001. Amino acid budgets in three aphid species using the same host plant. *Physiol Entomol* 26:202-211.

SantaLucia, J. 1998. A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci USA* 95:1460-1465.

Santos, J. M., M. Lobo, A. P. Matos, M. A. De Pedro and C. M. Arraiano. 2002. The gene *bolA* regulates *dacA* (PBP5), *dacC* (PBP6) and *ampC* (AmpC), promoting normal morphology in *Escherichia coli*. *Mol Microbiol* 45:1729-1740.

Sato, S. and H. Ishikawa. 1997. Expression and control of an operon from an intracellular symbiont which is homologous to the *groE* operon. *J Bacteriol* 179:2300-2304.

Scarborough, C. L., J. Ferrari and H. C. Godfray. 2005. Aphid protected from pathogen by endosymbiont. *Science* 310:1781.

Schadt, E. E., C. Li, B. Ellis and W. H. Wong. 2001. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J Cell Biochem* sup37:120-125.

Schena, M., D. Shalon, R. W. Davis and P. O. Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467-470.

Schena, M., D. Shalon, R. Heller, A. Chai, P. O. Brown and R. W. Davis. 1996. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci U S A* 93:10614-10619.

Schuchhardt, J., D. Beule, A. Malik, E. Wolski, H. Eickhoff, H. Lehrach and H. Herzel. 2000. Normalization strategies for cDNA microarrays. *Nucleic Acids Res* 28:E47.

Sharp, P. and W. Li. 1986. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for "rare" codons. *Nucleic Acids Res* 14:7737-7749.

Sharp, P. M., E. Bailes, R. J. Grocock, J. F. Peden and R. E. Sockett. 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res* 33:1141-1153.

Shigenobu, S., H. Watanabe, M. Hattori, Y. Sasaki and H. Ishikawa. 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* 407:81-86.

Silva, F. J., A. Latorre and A. Moya. 2001. Genome size reduction through multiple events of gene disintegration in *Buchnera* APS. *Trends Genet.* 17:615-618.

Silva, F. J., A. Latorre and A. Moya. 2003. Why are the genomes of endosymbiotic bacteria so stable? *Trends Genet* 19:176-180.

Smyth, G. and T. Speed. 2003. Normalization of cDNA microarray data. *Methods* 31:265-273.

Smyth, G., Y. H. Yang and T. Speed. 2002. Statistical issues in cDNA microarray data analysis. *www.stat.berkeley.edu/users/terry/zarray/Html/log.html*.

Soula, H., C. Robardet, F. Perrin, S. Gripon, G. Beslon and O. Gandrillon. 2005. Modeling the emergence of multi-protein dynamic structures by principles of self-organization through the use of 3DSpi, a multi-agent-based software. *BMC Bioinformatics* 6:228.

Southern, E., S. Case-Green, J. Elder, M. Johnson, K. Mir, L. Wang and J. Williams. 1994. Arrays of complementary oligonucleotides for analysing the hybridization behaviour of nucleic acids. *Nucleic Acids Res* 22:1368-1373.

Speed, T. and C. Yang. 2002. Direct versus indirect designs for cDNA Microarray Experiments. *The Indian Journal of Statistics* 64:706-720.

Spellman, P., G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein and B. Futcher. 1998. Comprehensive Identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol* 9:3273-3297.

Stoeckert, C., H. Causton and C. Ball. 2002. Microarray databases: standards and ontologies. *Nature Genetics* 32:469-473.

Storey, J. D. and R. Tibshirani. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 100:9440-9445.

Talaat, A., P. Hunter and S. Johnston. 2000. Genome-directed primers for selective labeling of bacterial transcripts for DNA microarray analysis. *Nat Biotechnol* 18:679-682.

Tamas, I., L. Klasson, B. Canback, A. K. Naslund, A. S. Eriksson, J. J. Wernegreen, J. P. Sandstrom, N. A. Moran and S. G. E. Andersson. 2002. 50 million years of genomic stasis in endosymbiotic bacteria. *Science* 296:2376-2379.

Tao, H., C. Bausch, C. Richmond, F. R. Blattner and T. Conway. 1999. Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. *J Bacteriol* 181:6425-6440.

Taylor, E., D. Cogdell, K. Coombes, L. Hu, L. Ramdas, A. Tabor, S. Hamilton and W. Zhang. 2001. Sequence verification as quality-control step for production of cDNA microarrays. *Biotechniques* 31:62-65.

Tempelman, R. J. 2005. Assessing statistical precision, power and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. *Vet Immunol Immunopathol* 105:175-186.

Thattai, M. and A. van Oudenaarden. 2004. Stochastic gene expression in fluctuating environments. *Genetics* 167:523-30.

Theilhaber, J., S. Bushnell, A. Jackson and R. Fuchs. 2001. Bayesian estimation of fold-changes in the analysis of gene expression: the PFOLD algorithm. *J Comput Biol* 8:585-614.

Tomassone, R., C. Dervin, J.P. Masson. 1993. Biométrie, modélisation de phénomènes biologiques. Masson eds. pp. 576.

Townsend, J. and D. Hartl. 2002. Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments. *Genome Biol* 3:research0071.

Tseng, G. C., M. K. Oh, L. Rohlin, J. C. Liao and W. H. Wong. 2001. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res* 29:2549-2557.

Tsuchida, T., R. Koga and T. Fukatsu. 2004. Host plant specialization governed by facultative symbiont. *Science* **303**:1989.

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

van de Peppel, J., P. Kemmeren, H. van Bakel, M. Radonjic, D. van Leenen and F. C. Holstege. 2003. Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Rep* 4:387-393.

van der Laan, M., S. Dudoit and K. S. Pollard. 2003a. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *U.C. Berkeley Division of Biostatistics working paper series (139)*.

van der Laan, M., S. Dudoit and K. S. Pollard. 2003b. Multiple testing. Part III. Procedures for control of the generalized family-wise error rate and proportion of false positives. *U.C. Berkeley Division of Biostatistics working paper series (141)*.

Van Ham, R., J. Kamerbeek, C. Palacios, C. Rausell, F. Abascal, U. Bastolla, J. Fernandez, L. Jiménez, M. Postigo, F. Silva, J. Tamames, E. Viguera, A. Latorre, A. Valencia, F. Moran and A. Moya. 2003. Reductive genome evolution in *Buchnera aphidicola*. *Proc Natl Acad Sci USA* 100:581-586.

Van Ham, R. C. H. J., D. Martinez-Torres, A. Moya and A. Latorre. 1999. Plasmid-encoded anthranilate synthase (TrpEG) in *Buchnera aphidicola* from aphids of the family Pemphigidae. *Appl Environ Microbiol* 65:117-125.

Van Ham, R. C. H. J., A. Moya and A. Latorre. 1997. Putative evolutionary origin of plasmids carrying the genes involved in leucine biosynthesis in *Buchnera aphidicola* (Endosymbiont of aphids). *J Bacteriol* 179:4768-4777.

Vert, J. P. and M. Kanehisa. 2003. Extracting active pathways from gene expression data. *Bioinformatics* 19 Suppl 2:II238-II244.

Vinuelas, J., F. Calevro, J. Bernillon, Y. Rahbé, G. Febvay, J. M. Fayard and H. Charles. 2005. Genomic DNA: an attractive candidate for microarray data normalization. Integrative Post-Genomics (IPG'05), Villeurbanne (France), novembre 2005.

Wernegreen, J. J. and N. A. Moran. 1999. Evidence for genetic drift in endosymbionts (*Buchnera*): Analyses of protein-coding genes. *Mol Biol Evol* 16:83-97.

Wernisch, L., S. L. Kendall, S. Soneji, A. Wietzorrek, T. Parish, J. Hinds, P. D. Butcher and N. G. Stoker. 2003. Analysis of whole-genome microarray replicates using mixed models. *Bioinformatics* 19:53-61.

Whitehead, L. F. and A. E. Douglas. 1993. A metabolic study of *Buchnera*, the intracellular bacterial symbionts of the pea aphid *Acyrthosiphon pisum*. *J Gen Microbiol* 139:821-826.

Wilcox, J. L., H. E. Dunbar, R. D. Wolfinger and N. A. Moran. 2003. Consequences of reductive evolution for gene expression in an obligate endosymbiont. *Mol Microbiol* 48:1491-1500.

Wilkinson, T. L., D. A. Ashford, J. Pritchard and A. E. Douglas. 1997. Honeydew sugars and osmoregulation in the pea aphid *Acyrthosiphon pisum*. *J Exp Biol* 200:2137-2143.

Wilkinson, T. L., T. Fukatsu, and H. Ishikawa. 2003. Transmission of symbiotic bacteria *Buchnera* to parthenogenetic embryos in the aphid *Acyrthosiphon pisum* (Hemiptera: Aphidoidea). *Arthropod Struct Dev* **32**:241-245.

Wilkinson, T. L. and H. Ishikawa. 1999. The assimilation and allocation of nutrients by symbiotic and aposymbiotic pea aphids, *Acyrthosiphon pisum*. *Entomol Exp Appl* 91:195-201.

Wolfinger, R. D., G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari and R. S. Paules. 2001. Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol* 8:625-637.

Workman, C., L. Jensen, H. Jarmer, R. Berka, L. Gautier, H. Nielser, H. H. Saxild, C. Nielsen, S. Brunak and S. Knudsen. 2002. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol* 3:RESEARCH0048.

Wright, F. and M. J. Bibb. 1992. Codon usage in the G+C-rich *Streptomyces* genome. *Gene* 113:55-65.

Wu, B. 2005. Differential gene expression detection using penalized linear regression models: the improved SAM statistics. *Bioinformatics* 21:1565-1571.

Wu, S., M. K. Kerr, X. Cui and G. A. Churchill. 2003. MAANOVA: a software package for the analysis of spotted cDNA microarray experiments. The analysis of gene expression data: methods and software. Springer.

Wu, W., S. E. Wildsmith, A. J. Winkley, R. Yallop, F. J. Elcock and P. J. Bugelski. 2001. Chemometric strategies for normalisation of gene expression data obtained from cDNA microarrays. *Anal Chim Acta* 446:451-466.

Wyatt, I. J. and P. F. White. 1977. Simple estimation of intrinsic increase rates for aphids and tetranychid mites. *J Appl Zool* 14:757-766.

Xie, Y., W. Pan and A. B. Khodursky. 2005. A note on using permutation based false discovery rate estimate to compare different analysis methods for microarray data. *Bioinformatics*. 21:4280-4288.

Yamanishi, Y., J. P. Vert and M. Kanehisa. 2004. Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics* 20 Suppl 1:I363-I370.

Yamanishi, Y., J. P. Vert, A. Nakaya and M. Kanehisa. 2003. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics* 19 Suppl 1:i323-330.

Yang, M. C., Q. G. Ruan, J. J. Yang, S. Eckenrode, S. Wu, R. A. McIndoe and J. X. She. 2001a. A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. *Physiol Genomics* 7:45-53.

Yang, M. C., J. J. Yang, R. A. McIndoe and J. X. She. 2003. Microarray experimental design: power and sample size considerations. *Physiol Genomics* 16:24-28.

Yang, Y., M. J. Buckley and T. P. Speed. 2001b. Analysis of cDNA microarray images. *Brief Bioinform* 2:341-349.

Yang, Y. and T. Speed. 2002. Design issues for cDNA microarray experiments. *Nature Reviews Genetics* 3:579-588.

Yang, Y. and T. Speed. 2003. Design and analysis of comparative microarray experiments. Pp. 240 *in* T. Speed, ed. Statistical analysis of gene expression microarray data. Chapman and Hall.

Yang, Y. H., S. Dudoit, P. Luu, D. Lin, V. Peng, J. Ngai and T. Speed. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 30:1-11.

Yue, H., P. S. Eastman, B. B. Wang, J. Minor, M. H. Doctolero, R. L. Nuttall, R. Stack, J. W. Becker, J. R. Montgomery, M. Vainer and R. Johnston. 2001. An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Res* 29:E41.

Zeilstra-Ryalls, J., O. Fayet and C. Georgopoulos. 1991. The universally conserved GroE (Hsp60) chaperonins. *Annu Rev Microbiol* 45:301-325.

Zuker, M. 2003. Mfold web server for nucleic acid folding hybridization prediction. *Nucleic Acids Res* 31:3406-3415.

ANNEXES

Présentation du modèle additif d'analyse de la variance pour le gène *g* de l'expérience AAS (l'indice *g* a été omis pour simplifier l'expression du modèle)

$$Y_{ldas} = \mu + \lambda_l + \delta_d + \alpha_a + \varphi_s + \gamma_{as} + \varepsilon_{ldas} \qquad \text{Eq. A.1}$$

Avec $Y_{g,\,ldas}$ le vecteur des logarithmes de fluorescence normalisés du gène *g*, μ la moyenne de fluorescence du gène *g*, λ_l l'effet lame (L niveaux), δ_d l'effet fluorochrome (D = 2 niveaux), α_q l'effet acide aminé (A = 2 niveaux), φ_s l'effet saccharose (S = 2 niveaux), γ_{as} l'interaction entre les effets acides aminés et saccharose (AS = 1 niveau) et ε_{ldas} l'effet aléatoire. Pour déterminer les coefficients du modèle, il faut définir les contrastes associés à chacun des facteurs. Pour chaque facteur, nous choisissons le contraste correspondant à la comparaison à l'effet moyen (somme des effets nulle). Le tableau A.1 présente la matrice centrée réduite de projection du modèle *Xg*. Elle est de dimension 36 x 24. Le bloc du bas du tableau correspond aux contrastes.

μ	**L1**	**L2**	**L3**	**L4**	**L5**	**L6**	**L7**	**L8**	**…**	**D1**	**D2**	**A1**	**A2**	**S1**	**S2**	**AS**
1	1	0	0	0	0	0	0	0	…	1	0	0	1	1	0	0
1	1	0	0	0	0	0	0	0	…	0	1	1	0	1	0	1
1	0	1	0	0	0	0	0	0	…	1	0	1	0	1	0	1
1	0	1	0	0	0	0	0	0	…	0	1	0	1	0	1	1
1	0	0	1	0	0	0	0	0	…	1	0	0	1	0	1	1
1	0	0	1	0	0	0	0	0	…	0	1	1	0	0	1	0
1	0	0	0	1	0	0	0	0	…	1	0	1	0	0	1	0
1	0	0	0	1	0	0	0	0	…	0	1	0	1	1	0	0
1	0	0	0	0	1	0	0	0	…	1	0	0	1	1	0	0
1	0	0	0	0	1	0	0	0	…	0	1	0	1	0	1	1
1	0	0	0	0	0	1	0	0	…	1	0	1	0	1	0	1
1	0	0	0	0	0	1	0	0	…	0	1	1	0	0	1	0
1	0	0	0	0	0	0	1	0	…	1	0	1	0	0	1	0
1	0	0	0	0	0	0	1	0	…	0	1	1	0	1	0	1
1	0	0	0	0	0	0	0	1	…	1	0	0	1	0	1	1
1	0	0	0	0	0	0	0	1	…	0	1	0	1	1	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
0	1	1	1	1	1	1	1	1	…	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	…	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	…	0	0	1	1	0	0	0
0	0	0	0	0	0	0	0	0	…	0	0	0	0	1	1	0

Tableau A.1. Matrice *Xg*, base de la projection du vecteur des observations de l'expérience AAS pour le gène g. Les lames 9 à 16 ont été éliminées du tableau pour faciliter la lecture.

A partir de la matrice Xg, il est possible d'estimer la matrice des variances / covariances associée à l'expérience AAS pour le gène *g* (Cg). Cette matrice est de dimension (32x32). Elle est donnée par la formule suivante :

$$C_g = \left(X_g^t \times X_g \right)^{-1} \qquad \text{Eq. A.2.}$$

La matrice est symétrique, mais non diagonale (tableau A.2). Toutes les covariances entre les paramètres ne sont pas nulles, car le plan n'est pas entièrement orthogonal. On distingue dans cette matrice le bloc des lames 1 à 4 (boucle sans « dye swap ») et le bloc des lames 5 à 8 en « dye swap » complet. On note également que l'effet fluorochrome (D) est diagonal, contrairement aux effets A, S et AS. Enfin, on note une variance plus forte associée au facteur acide aminé (huit lames en comparaison directe), par rapport au facteur sucre (12 lames en comparaison directe). Le plan est donc un peu plus puissant pour détecter les effets du facteur sucre.

	mu	**L1**	**L2**	**L3**	**L4**	**L5**	**L6**	**L7**	**L8**	**…**	**D1**	**D2**	**A1**	**A2**	**S1**	**S2**	**AS**
mu	**0,827**	-0,004	0,038	-0,004	-0,046	-0,004	-0,004	-0,004	-0,004	…	-0,25	-0,25	-0,25	-0,25	-0,25	-0,25	-0,083
L1	-0,004	**0,514**	-0,027	-0,069	-0,027	-0,027	-0,027	-0,027	-0,027	…	0	0	0	0	-0,042	0,042	0
L2	0,038	-0,027	**0,514**	-0,027	-0,069	-0,027	-0,027	-0,027	-0,027	…	0	0	0	0	0	0	-0,083
L3	-0,004	-0,069	-0,027	**0,514**	-0,027	-0,027	-0,027	-0,027	-0,027	…	0	0	0	0	0,042	-0,042	0
L4	-0,046	-0,027	-0,069	-0,027	**0,514**	-0,027	-0,027	-0,027	-0,027	…	0	0	0	0	0	0	0,083
L5	-0,004	-0,027	-0,027	-0,027	-0,027	**0,535**	-0,090	-0,090	0,035	…	0	0	0,063	-0,063	0	0	0
L6	-0,004	-0,027	-0,027	-0,027	-0,027	-0,090	**0,535**	0,035	-0,090	…	0	0	-0,063	0,063	0	0	0
L7	-0,004	-0,027	-0,027	-0,027	-0,027	-0,090	0,035	**0,535**	-0,090	…	0	0	-0,063	0,063	0	0	0
L8	-0,004	-0,027	-0,027	-0,027	-0,027	0,035	-0,090	-0,090	**0,535**	…	0	0	0,063	-0,063	0	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
D1	-0,250	0	0	0	0	0	0	0	0	…	**0,281**	0,219	0	0	0	0	0
D2	-0,250	0	0	0	0	0	0	0	0	…	0,219	**0,281**	0	0	0	0	0
A1	-0,250	0	0	0	0	0,063	-0,063	-0,063	0,063	…	0	0	**0,313**	0,188	0	0	0
A2	-0,250	0	0	0	0	-0,063	0,063	0,063	-0,063	…	0	0	0,188	**0,313**	0	0	0
S1	-0,250	-0,042	0	0,042	0	0	0	0	0	…	0	0	0	0	**0,292**	0,208	0
S2	-0,250	0,042	0	-0,042	0	0	0	0	0	…	0	0	0	0	0,208	**0,292**	0
AS	-0,083	0	-0,083	0	0,083	0	0	0	0	…	0	0	0	0	0	0	**0,167**

Tableau A.2. Matrice des variances covariances de l'expérience AAS. Les lames 9 à 16 ont été éliminées de la matrice pour faciliter la lecture.

L'ajustement du modèle linéaire est réalisé selon la méthode des moindres carrés. Le vecteur des paramètres θ est alors donné par l'équation : $\theta = \left(X_g^t X_g \right)^{-1} X_g^t Y$. Les coefficients du modèle (Eq. A.1) sont calculés : λ_l est la moyenne des deux canaux de la lame *l* avec un faible effet additif lié aux covariances des autres lames. Le paramètre δ_d est la différence de moyennes entre les deux fluorochromes, les paramètres α_a, φ_s *et* γ_{as} sont les différences de moyennes entre les deux niveaux d'acides aminés, de saccharose et d'interactions sur les lames ou la comparaison est réalisée.

BU	Noms	A	M	Exp.	+/-
BU017	tRNA-Phe-GAA	10,73	-0,175	0,885	0
BU041	tRNA-Thr-GGU	10,80	-0,030	0,979	0
BU042	tRNA-Gly-GCC	10,29	0,646	1,565	+
BU043	tRNA-Tyr-GUA	10,71	0,244	1,184	0
BU044	tRNA-Thr-TGT	10,23	0,406	1,325	+*
BU068	tRNA-Lys-UUU	10,69	0,154	1,113	0
BU069	tRNA-Val-GAC	9,84	0,231	1,173	0
BU071	tRNA-Ala-GGC	10,70	-0,028	0,980	0
BU111	tRNA-Val-UAC	11,16	0,176	1,130	0
BU244	tRNA-Ile-GAU	12,66	0,490	1,405	+
BU245	tRNA-Ala-UGC	13,47	0,026	1,018	0
BU249	tRNA-Asp-GUC	13,04	0,013	1,009	0
BU329	tRNA-Leu-GAG	10,55	0,851	1,804	+**
BU330	tRNA-Cys-PseudoGCA	11,87	0,258	1,200	+*
BU331	tRNA-Ser-GCU	7,55	0,934	1,911	+
BU379	tRNA-Leu-UAA	9,86	-0,276	0,825	-
BU405	tRNA-Ser-GGA	11,02	0,373	1,295	+
BU406	tRNA-Arg-ACG	11,33	0,115	1,083	0
BU412	tRNA-Gln-UUG	11,39	0,297	1,228	+
BU413	tRNA-Leu-UAG	11,72	0,510	1,424	+*
BU414	tRNA-Met-CAU	10,99	0,072	1,051	0
BU457	tRNA-Met-CAU	11,29	0,193	1,143	0
BU485	tRNA-Arg-CCG	8,34	0,404	1,324	+
BU492	tRNA-Glu-UUC	13,58	-0,456	0,728	-
BU540	tRNA-Ser-UGA	10,21	0,967	1,955	+
BU557	tRNA-Asn-GUU	12,32	-0,016	0,988	0
BU558	tRNA-Met-CAU	10,11	0,417	1,335	+
BU575	tRNA-Gly-UUC	10,74	0,271	1,207	+
BU593	tRNA-Pro-UGG	9,71	0,923	1,896	+
BU594	tRNA-His-GUG	13,04	0,179	1,132	0
BU595	tRNA-Arg-UCU	8,79	0,226	1,169	0*
BU601	tRNA-Trp-CCA	10,18	0,246	1,186	0

Tableau A.3. Expression des ARNt isoaccepteurs chez *Buchnera* dans l'expérience YF. La colonne A représente l'expression moyenne $A = \log 2\left(\sqrt{YF° \times AP3}\right)$ et M l'expression relative $M = \log 2\left(\frac{YF°}{AP3}\right)$. La colonne Exp. présente les rapports d'expression. La colonne +/- donne le sens de la régulation (** : probabilité < 0,05, * : probabilité < 0,1). Les lignes pointillées correspondent aux différents contigs sur le chromosome.

BU	Gène	Produit du gène	RE	p-value	test
BU592	*hemD*	uroporphyrinogen-iii synthase, putative	0,318	0,022	a
BU333	*mviN*	Virulence factor mviN homolog	0,332	0,012	t, a
BU478	*ppiD*	Peptidyl-prolyl cis-trans isomerase D	0,411	0,045	a
BU243	*rrs*	16S rRNA	0,426	0,001	a
BU265	*fabI*	Enoyl-[acyl-carrier-protein] reductase [NADH]	0,448	0,020	a
BU130	*pheT*	Phenylalanyl-tRNA synthetase beta chain	0,460	0,028	a
BU142	*apaH*	Bis(5'-nucleosyl)-tetraphosphatase	0,465	0,034	a
BU490	*rrf*	5S rRNA	0,475	0,016	a
BU087	*ytfN*	conserved hypothetical protein	0,494	0,025	a
BU432	*yihA*	Probable GTP-binding protein engB.	0,507	0,021	t, a
BUpT01	*trpE*	Anthranilate synthase component I	0,673	0,015	t, a
BU089	*pmbA*	PmbA protein homolog	0,678	0,048	t, a
BU407	*gshA*	Glutamate--cysteine ligase	0,691	0,040	t
BU521	*rplB*	50S ribosomal protein L2	0,699	0,038	t
BU207	*lpdA*	Dihydrolipoyl dehydrogenase	0,704	0,040	t
BU099	*hisG*	ATP phosphoribosyltransferase	0,707	0,046	t, a
BU101	*hisC*	Histidinol-phosphate aminotransferase	0,718	0,009	t, a
BU172	*hemK*	HemK protein homolog	0,722	0,037	t, a
BU471	*cyoB*	Ubiquinol oxidase polypeptide I	0,728	0,026	t, a
BU465	*ispA*	Geranyltranstransferase	0,728	0,016	t
BU144	*carB*	Carbamoyl-phosphate synthase large chain	0,764	0,042	t, a
BU039	*nusG*	Transcription antitermination protein nusG	1,205	0,008	a
BU065	*ptsH*	Phosphocarrier protein HPr (His-containing prot.)	1,216	0,030	t
BU576	*amiB*	Putative N-acetylmuramoyl-L-alanine amidase	1,247	0,043	t
BU067	*lig*	DNA ligase	1,272	0,036	t
BU035	*rplL*	50S ribosomal protein L7/L12	1,281	0,017	t
BU016	*thdF*	Probable tRNA modification GTPase trmE	1,286	0,045	t
BUpL01	*repA1*	Probable replication-associated protein repA1	1,292	0,003	t, a
BU488	*ybeD*	conserved hypothetical protein	1,302	0,049	a
BU396	*trmD*	tRNA (Guanine-N(1)-)-methyltransferase	1,340	0,014	t
BU384	*greA*	Transcription elongation factor greA	1,410	0,009	t, a
BU585	*yba4*	conserved hypothetical protein	1,464	0,030	t, a
BU153	*dnaK*	Chaperone protein dnaK (HSP 70 kDa protein)	1,502	0,013	t
BU215	*murC*	UDP-N-acetylmuramate--L-alanine ligase	1,531	0,037	t, a
BU586	*yhiQ*	conserved hypothetical protein	1,544	0,019	t, a
BU177	*yfaE*	2Fe-2S iron-sulfur cluster binding domains, putative	1,561	0,031	t, a
BU187	*ydhD*	Hypothetical monothiol glutaredoxin BU187	1,568	0,003	t, a
BU580	*ibpA*	Small heat shock protein ibp	1,581	0,022	a
BU447	*yhhP*	SirA protein homolog.	1,590	0,015	t, a
BU397	*rplS*	50S ribosomal protein L19.	1,649	0,016	t
BUpL07	*leuD*	3-isopropylmalate dehydratase small subunit	1,650	0,033	a
BU329	*tRNA-Leu*	TRNA-Leu-GAG	1,805	0,040	t
BU574	*orn*	Oligoribonuclease	1,858	0,014	a
BU112	*ribE*	Riboflavin synthase alpha chain	1,943	0,004	a
BU135	*glyS*	Glycyl-tRNA synthetase beta chain	2,177	0,023	a
BU336	*flgA*	Flagella basal body P-ring formation	2,291	0,011	a
BU140	*surA*	Survival protein SurA homolog precursor.	2,971	0,001	a
BU070	*gltX*	Glutamyl-tRNA synthetase	3,283	0,003	a
BU103	*hisH*	Imidazole glycerol phosphate synthase subunit hisH	3,338	0,001	a
BU062	*bacA*	Undecaprenyl-diphosphatase	4,208	0,001	a
BU182	*ahpC*	Probable peroxiredoxin	4,763	0,001	a
BU608	*yfgM*	conserved hypothetical protein	10,448	0,001	a

Tableau A.4. Liste des gènes différentiels dans l'expérience YF. (t, a) correspondant aux gènes significatifs dans le test de t ou dans l'ANOVA (lorsque les deux tests sont significatifs, la plus faible des probabilités est indiquée). RE : rapports d'expression.

BU labels	Names	RE	p-value	$M_{0.5}$ M_1
BU337	*flgB*	1,65[2]	0,001	+ +[1]
BU172	*hemK*	1,44	0,034	+ +
BU092	*fabB*	1,41	0,029	+ +
BU209	*speE*	1,24	0,002	+ +
BU131	*himA*	1,16	0,023	+ +
BU349	*rpmF*	1,63	0,025	0 +
BU534	*argD*	1,47	10^{-4}	0 +
BU144	*carB*	1,42	0,045	0 +
BU066	*cysK*	1,28	0,023	0 +
BU280	*trpD*	1,18	0,011	0 +
BU181	*yba2*	1,25	0,001	0 +
BU085	*rpmG*	1,24	10^{-4}	0 +
BU339	*flgD*	1,23	0,008	0 +
BU521	*rplB*	1,80	0,015	+ 0
BU292	*bioA*	1,45	0,042	+ 0
BU605	*hscA*	1,33	10^{-5}	+ 0
BU168	*cvpA*	1,28	10^{-5}	+ 0
BU304	*gpmA*	1,27	0,037	+ 0
BU382	*hflB*	1,21	0,001	+ 0
BU299	*fldA*	0,44	0,001	- -
BU010	*gyrB*	0,71	0,046	- -
BU341	*flgF*	0,75	0,001	- -
BU126	*infC*	0,77	0,001	- -
BU417	*eno*	0,78	0,005	- -
BU073	*fliF*	0,79	0,04	- -
BU234	*frr*	0,80	0,04	- -
BU513	*rplX*	0,82	0,011	- -
BU235	*dxr*	0,83	0,029	- -
BU039	*nusG*	0,71	10^{-5}	0 -
BU359	*ompF*	0,74	10^{-5}	0 -
BU472	*cyoA*	0,74	0,001	0 -
BU124	*aroH*	0,79	10^{-5}	0 -

Tableau A.5. Liste des gènes différentiels en condition de stress en acides aminés de l'expérience AAS. [(1)] : signes des log(rapport d'expression) à 0,5 et 1 M de saccharose ; [(2)] : La valeur moyenne du rapport d'expression (RE) est reportée lorsque le rapport est significatif dans les deux conditions.

BU labels	Names	FC	p-value	M_{50} M_{25}
BU419	*ygbB*	1,62	0,013	++
BU168	*cvpA*	1,47	10^{-4}	++
BUpL02	*yqhA*	1,34	10^{-5}	++
BU286	*yfgB*	1,20	0,001	++
BU252	*grpE*	1,69	0,043	+ 0
BU452	*yggB*	1,48	0,015	+ 0
BU092	*fabB*	1,41	0,029	+ 0
BUpL07	*leuD*	1,34	0,016	+ 0
BU300	*phrB*	1,29	0,035	+ 0
BU051	*argH*	1,29	0,032	+ 0
BU603	*iscU*	1,29	0,032	+ 0
BU377	*infB*	1,27	0,009	+ 0
BU110	*mesJ*	1,26	0,001	+ 0
BU123	*yb1688*	1,26	0,003	+ 0
BU509	*rplF*	1,25	0,011	+ 0
BU108	*dcd*	1,25	0,031	+ 0
BU541	*deoD*	1,25	0,033	+ 0
BU531	*yheM*	1,25	0,003	+ 0
BU019	*mopA*	1,23	0,001	+ 0
BU570	*mutL*	1,23	0,013	+ 0
BU359	*ompF*	1,20	10^{-5}	+ 0
BU554	*murI*	0,60	0,045	- -
BU204	*guaC*	0,72	0,018	- -
BU045	*murB*	0,79	10^{-5}	- -
BU119	*nth*	0,82	0,018	- -
BU152	*dnaJ*	0,82	10^{-5}	- -
BU275	*yciB*	0,83	0,001	- -
BU576	*amiB*	0,83	0,042	- -
BU464	*dxs*	0,81	0,014	0 -
BU337	*flgB*	0,63	0,001	- 0
BU272	*hns*	0,67	0,005	- 0
BU268	*lipB*	0,69	0,004	- 0
BU086	*rpmB*	0,71	10^{-4}	- 0
BU085	*rpmG*	0,71	10^{-4}	- 0
BU356	*ptsG*	0,72	0,038	- 0
BU282	*yciL*	0,74	0,016	- 0
BU116	*ydgO*	0,74	0,03	- 0
BU585	*yba4*	0,75	0,01	- 0
BU181	*yba2*	0,76	0,001	- 0

BU349	*rpmF*	0,76	0,011	- 0
BU338	*flgC*	0,76	0,009	- 0
BU417	*eno*	0,77	0,005	- 0
BU255	*yfhC*	0,78	0,046	- 0
BU066	*cysK*	0,78	0,001	- 0
BU460	*thiL*	0,79	0,021	- 0
BU397	*rplS*	0,79	0,016	- 0
BU278	*trpB*	0,79	0,035	- 0
BU314	*trxB*	0,79	0,03	- 0
BU144	*carB*	0,80	0,044	- 0
BU597	*trxA*	0,81	0,012	- 0
BU532	*yheN*	0,81	0,008	- 0
BU065	*ptsH*	0,81	0,007	- 0
BU400	*fis*	0,81	0,012	- 0
BU107	*gnd*	0,82	0,001	- 0
BU387	*rplU*	0,82	0,001	- 0
BU315	*infA*	0,83	0,028	- 0
BU187	*ydhD*	0,83	0,027	- 0
BU121	*tyrS*	0,84	0,002	- 0
BU194	*thrA*	0,84	0,035	- 0
BU425	*cysG*	0,84	0,007	- 0
BU299	*fldA*	1,46-0,44	0,001	+ -
BU605	*hscA*	1,33-0,81	10^{-5}	+ -

Tableau A.6. Liste des gènes différentiels sous l'effet du stress osmotique dans les deux conditions d'acides aminés de l'expérience AAS. (1) : signes des log(rapport d'expression) à 0,5 et 1 M de saccharose ; (2) : la valeur moyenne du rapport d'expression (RE) est relevée lorsque le rapport est significatif dans les deux conditions.

Gene	amorces (5'_3')	P	L	T	Type	q-RT-PCR$_{RD}$	Puce$_{RD}$
atpA	For: AAATGATTTTATTGCCTAA Rev: CAGAAATATGAATATAGGGA	152	99	47	control	1,00	-
rpsF	For: ATCAAAGCGACAAAATACC Rev: CCCCAATCTTCTAATCTATGTA	38	91	47	control	1,05	-
fliJ	For: AAATTAATATTAGGTGTGTCGG Rev: AATTTGATTTTTCGACCATA	163	153	47	target	0,93	0,46
leuD	For: GTAATAATTGCACCTAGTTTTG Rev: CCTTATTTACAGTAACCGTTT	295	175	47	target	1,40	1,65
pheA	For: CGTGACGAAATTAATAACATT Rev: CTATATCTCTTATCGCCTGATT	31	118	47	target	0,97	1,10
pheT	For: TTATAAAGAATCGCATAATCAT Rev: CCTTAAAACGGCTTGAA	2235	148	47	target	0,75	0,46
yfgM	For: TGAAGTAGAAAATTTTATCGTT Rev: GTTTATTTTCAAAAGGTTCTT	174	172	47	target	2,11	10,45
atpA	For: CCATTACAGCTTTGCCTAT Rev: CACGAGAAACAGATATACCG	968	162	51	control	1,00	-
rplO	For: AAACTTCTGGTCGAGGTC Rev: GTCTTCTGTATAACGGCATT	86	96	51	control	1,04	-
rpsF	For: ATCCTGATCAAAGCGACAAAAT Rev: ACGTCTCCCCCAATCTTCTAAT	32	104	51	control	1,05	-
aroH	For: CCATATGGGCATATAATCAT Rev: ATTACCGTGACTAAAATCAATC	676	132	51	target	**1,61**	**0,9**
dnaK	For: AATTATCATCCGCACAACAA Rev: TTTTAATGCAACTTTTAGTGGC	815	161	51	target	1,46	1,50
flgA	For: GTTTAGATGTACTGCCGAACAAT Rev: TTAACTAACCAAAATGGACGTG	383	127	51	target	1,35	2,29
hisC	For: TTTGCAATCCCAATAATCCTAC Rev: GGCAAAAGCTTTAGACAAGGTT	458	197	51	target	0,67	0,72
hisG	For: AGCAGGTGATGATAATCGAGT Rev: GCGCCTAATAATTTTAGCTTTT	768	90	51	target	0,86	0,71
hisH	For: TTACCAGGAGTAGGAACTGC Rev: GTTTTTACGCCATTAGATTCTT	130	155	51	target	**0,91**	**3,34**
yfgK	For: CGGTCGTACTAACGTAGGAA Rev: TGCTTGTTTTTCTATTTCGTTT	24	193	51	target	1,12	1,66

Tableau A7. Résultats et amorces utilisées pour la RT-PCR quantitative. P : position de l'amorce à partir du codon start ; L : longueur du produit amplifié en pb ; T : température d'hybridation en °C ; Rapports d'expression différentielle estimés par PCR (qRT-PCR$_{RD}$) ou par puce à ADN dans l'expérience YF (Puce$_{RD}$). Les valeurs en gras sont les rapports d'expression différentielle incohérents entre les deux techniques.