HAL
open science

Bioinformatique des puces à ADN et application à l'analyse du transcriptome de Buchnera aphidicola

Nancie Reymond

To cite this version:

Nancie Reymond. Bioinformatique des puces à ADN et application à l'analyse du transcriptome de Buchnera aphidicola. Sciences du Vivant [q-bio]. INSA de Lyon, 2004. Français. NNT : . tel-00008630v3

HAL Id: tel-00008630

https://theses.hal.science/tel-00008630v3

Submitted on 20 Apr 2005

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre 04 ISAL 0100
Année 2004

Thèse

Bioinformatique des puces à ADN et application à l'analyse du transcriptome de *Buchnera aphidicola*

Présentée devant
L'institut national des sciences appliquées de Lyon

Pour obtenir
Le grade de docteur

Formation doctorale
Analyse et Modélisation des Systèmes Biologiques
Bioinformatique
École doctorale Évolution, Écosystèmes, Microbiologie, Modélisation (E2M2)

Par
Nancie Reymond
(Ingénieur)

Soutenue le 16 décembre 2004 devant la Commission d'examen

Jury MM.

H. Charles	Maître de Conférences (INSA de Lyon) – Directeur
J.-M. Fayard	Professeur (INSA de Lyon) – Directeur
G. Febvay	Directeur de Recherches (INRA)
C. Gautier	Professeur (UCBL) – Président
M.-C. Potier	Chargée de Recherches (CNRS) – Rapporteur
D. Tagu	Directeur de Recherches (INRA) – Rapporteur
A. Trubuil	Ingénieur de Recherches (INRA)

Écoles Doctorales

CHIMIE DE LYON

Responsable : M. Denis SINOU

Université Claude Bernard Lyon 1
Lab Synthèse Asymétrique UMR UCB/CNRS 5622
Bât 308
2ème étage
43 bd du 11 novembre 1918
69622 VILLEURBANNE Cedex
Tél : 04.72.44.81.83
sinou@univ-lyon1.fr

ECONOMIE, ESPACE ET MODELISATION DES COMPORTEMENTS (E2MC)

Responsable : M. Alain BONNAFOUS

Université Lyon 2
14 avenue Berthelot
MRASH
Laboratoire d'Economie des Transports
69363 LYON Cedex 07
Tél : 04.78.69.72.76
Alain.Bonnafous@mrash.fr

ELECTRONIQUE, ELECTROTECHNIQUE, AUTOMATIQUE (EEA)

Responsable : M. Daniel BARBIER

INSA DE LYON
Laboratoire Physique de la Matière
Bâtiment Blaise Pascal
69621 VILLEURBANNE Cedex
Tél : 04.72.43.64.43
Daniel.Barbier@insa-lyon.fr

EVOLUTION, ECOSYSTEMES, MICROBIOLOGIE, MODELISATION (E2M2)

http://biomserv.univ-lyon1.fr/E2M2

Responsable : M. Jean-Pierre FLANDROIS

UMR 5558 Biométrie et Biologie Evolutive
Equipe Dynamique des Populations Bactériennes
Faculté de Médecine Lyon-Sud Laboratoire de Bactériologie BP 12
69600 OULLINS
Tél : 04.78.86.31.50
Jean-Pierre.Flandrois@biomserv.univ-lyon1.fr

INFORMATIQUE ET INFORMATION POUR LA SOCIETE (EDIIS)

http://www.insa-lyon.fr/ediis

Responsable : M. Lionel BRUNIE

INSA DE LYON
EDIIS
Bâtiment Blaise Pascal
69621 VILLEURBANNE Cedex
Tél : 04.72.43.60.55
lbrunie@if.insa-lyon.fr

INTERDISCIPLINAIRE SCIENCES-SANTE (EDISS)

http://www.ibcp.fr/ediss
Responsable : M. Alain Jean COZZONE

IBCP (UCBL1)
7 passage du Vercors
69367 LYON Cedex 07
Tél : 04.72.72.26.75
cozzone@ibcp.fr

MATERIAUX DE LYON

http://www.ec-lyon.fr/sites/edml
Responsable : M. Jacques JOSEPH

Ecole Centrale de Lyon
Bât F7 Lab. Sciences et Techniques des Matériaux et des Surfaces
36 Avenue Guy de Collongue BP 163
69131 ECULLY Cedex
Tél : 04.72.18.62.51
Jacques.Joseph@ec-lyon.fr

MATHEMATIQUES ET INFORMATIQUE FONDAMENTALE (Math IF)

http://www.ens-lyon.fr/MathIS
Responsable : M. Franck WAGNER

Université Claude Bernard Lyon1
Institut Girard Desargues
UMR 5028 MATHEMATIQUES
Bâtiment Doyen Jean Braconnier
Bureau 101 Bis, 1er étage
69622 VILLEURBANNE Cedex
Tél : 04.72.43.27.86
wagner@desargues.univ-lyon1.fr

MECANIQUE, ENERGETIQUE, GENIE CIVIL, ACOUSTIQUE (MEGA)

http://www.lmfa.ec-lyon.fr/autres/MEGA/index.html

Responsable : M. François SIDOROFF

Ecole Centrale de Lyon
Lab. Tribologie et Dynamique des Systêmes Bât G8
36 avenue Guy de Collongue
BP 163
69131 ECULLY Cedex
Tél :04.72.18.62.14
Francois.Sidoroff@ec-lyon.fr

Novembre 2003

INSTITUT NATIONAL DES SCIENCES APPLIQUEES DE LYON

Directeur : STORCK A.

Professeurs

AMGHAR Y.	LIRIS
AUDISIO S.	PHYSICOCHIMIE INDUSTRIELLE
BABOT D.	CONT. NON DESTR. PAR RAYONNEMENTS IONISANTS
BABOUX J.C.	GEMPPM***
BALLAND B.	PHYSIQUE DE LA MATIERE
BAPTISTE P.	PRODUCTIQUE ET INFORMATIQUE DES SYSTEMES MANUFACTURIERS
BARBIER D.	PHYSIQUE DE LA MATIERE
BASKURT A.	LIRIS
BASTIDE J.P.	LAEPSI****
BAYADA G.	MECANIQUE DES CONTACTS
BENADDA B.	LAEPSI****
BETEMPS M.	AUTOMATIQUE INDUSTRIELLE
BIENNIER F.	PRODUCTIQUE ET INFORMATIQUE DES SYSTEMES MANUFACTURIERS
BLANCHARD J.M.	LAEPSI****
BOISSE P.	LAMCOS
BOISSON C.	VIBRATIONS-ACOUSTIQUE
BOIVIN M. (Prof. émérite)	MECANIQUE DES SOLIDES
BOTTA H.	UNITE DE RECHERCHE EN GENIE CIVIL - Développement Urbain

BOTTA-ZIMMERMANN M. (Mme)	UNITE DE RECHERCHE EN GENIE CIVIL - Développement Urbain
BOULAYE G. (Prof. émérite)	INFORMATIQUE
BOYER J.C.	MECANIQUE DES SOLIDES
BRAU J.	CENTRE DE THERMIQUE DE LYON - Thermique du bâtiment
BREMOND G.	PHYSIQUE DE LA MATIERE
BRISSAUD M.	GENIE ELECTRIQUE ET FERROELECTRICITE
BRUNET M.	MECANIQUE DES SOLIDES
BRUNIE L.	INGENIERIE DES SYSTEMES D'INFORMATION
BUFFIERE J-Y.	GEMPPM***
BUREAU J.C.	CEGELY*
CAMPAGNE J-P.	PRISMA
CAVAILLE J.Y.	GEMPPM***
CHAMPAGNE J-Y.	LMFA
CHANTE J.P.	CEGELY*- Composants de puissance et applications
CHOCAT B.	UNITE DE RECHERCHE EN GENIE CIVIL - Hydrologie urbaine
COMBESCURE A.	MECANIQUE DES CONTACTS
COURBON	GEMPPM
COUSIN M.	UNITE DE RECHERCHE EN GENIE CIVIL - Structures
DAUMAS F. (Mme)	CENTRE DE THERMIQUE DE LYON - Energétique et Thermique
DJERAN-MAIGRE I.	UNITE DE RECHERCHE EN GENIE CIVIL
DOUTHEAU A.	CHIMIE ORGANIQUE
DUBUY-MASSARD N.	ESCHIL
DUFOUR R.	MECANIQUE DES STRUCTURES
DUPUY J.C.	PHYSIQUE DE LA MATIERE
EMPTOZ H.	RECONNAISSANCE DE FORMES ET VISION
ESNOUF C.	GEMPPM***
EYRAUD L. (Prof. émérite)	GENIE ELECTRIQUE ET FERROELECTRICITE
FANTOZZI G.	GEMPPM***
FAVREL J.	PRODUCTIQUE ET INFORMATIQUE DES SYSTEMES MANUFACTURIERS

FAYARD J.M.	BIOLOGIE FONCTIONNELLE, INSECTES ET INTERACTIONS
FAYET M. (Prof. émérite)	MECANIQUE DES SOLIDES
FAZEKAS A.	GEMPPM
FERRARIS-BESSO G.	MECANIQUE DES STRUCTURES
FLAMAND L.	MECANIQUE DES CONTACTS
FLEURY E.	CITI
FLORY A.	INGENIERIE DES SYSTEMES D'INFORMATIONS
FOUGERES R.	GEMPPM***
FOUQUET F.	GEMPPM***
FRECON L. (Prof. émérite)	REGROUPEMENT DES ENSEIGNANTS CHERCHEURS ISOLES
GERARD J.F.	INGENIERIE DES MATERIAUX POLYMERES
GERMAIN P.	LAEPSI****
GIMENEZ G.	CREATIS**
GOBIN P.F. (Prof. émérite)	GEMPPM***
GONNARD P.	GENIE ELECTRIQUE ET FERROELECTRICITE
GONTRAND M.	PHYSIQUE DE LA MATIERE
GOUTTE R. (Prof. émérite)	CREATIS**
GOUJON L.	GEMPPM***
GOURDON R.	LAEPSI****.
GRANGE G. (Prof. émérite)	GENIE ELECTRIQUE ET FERROELECTRICITE
GUENIN G.	GEMPPM***
GUICHARDANT M.	BIOCHIMIE ET PHARMACOLOGIE
GUILLOT G.	PHYSIQUE DE LA MATIERE
GUINET A.	PRODUCTIQUE ET INFORMATIQUE DES SYSTEMES MANUFACTURIERS
GUYADER J.L.	VIBRATIONS-ACOUSTIQUE
GUYOMAR D.	GENIE ELECTRIQUE ET FERROELECTRICITE
HEIBIG A.	MATHEMATIQUE APPLIQUEES DE LYON
JACQUET-RICHARDET G.	MECANIQUE DES STRUCTURES
JAYET Y.	GEMPPM***
JOLION J.M.	RECONNAISSANCE DE FORMES ET VISION
JULLIEN J.F.	UNITE DE RECHERCHE EN GENIE CIVIL - Structures

JUTARD A. (Prof. émérite)	AUTOMATIQUE INDUSTRIELLE
KASTNER R.	UNITE DE RECHERCHE EN GENIE CIVIL - Géotechnique
KOULOUMDJIAN J. (Prof. émérite)	INGENIERIE DES SYSTEMES D'INFORMATION
LAGARDE M.	BIOCHIMIE ET PHARMACOLOGIE
LALANNE M. (Prof. émérite)	MECANIQUE DES STRUCTURES
LALLEMAND A.	CENTRE DE THERMIQUE DE LYON - Energétique et thermique
LALLEMAND M. (Mme)	CENTRE DE THERMIQUE DE LYON - Energétique et thermique
LAREAL P (Prof. émérite)	UNITE DE RECHERCHE EN GENIE CIVIL - Géotechnique
LAUGIER A. (Prof. émérite)	PHYSIQUE DE LA MATIERE
LAUGIER C.	BIOCHIMIE ET PHARMACOLOGIE
LAURINI R.	INFORMATIQUE EN IMAGE ET SYSTEMES D'INFORMATION
LEJEUNE P.	UNITE MICROBIOLOGIE ET GENETIQUE
LUBRECHT A.	MECANIQUE DES CONTACTS
MASSARD N.	INTERACTION COLLABORATIVE TELE-FORMATION TELEACTIVITE
MAZILLE H. (Prof. émérite)	PHYSICOCHIMIE INDUSTRIELLE
MERLE P.	GEMPPM***
MERLIN J.	GEMPPM***
MIGNOTTE A. (Mle)	INGENIERIE, INFORMATIQUE INDUS-TRIELLE
MILLET J.P.	PHYSICOCHIMIE INDUSTRIELLE
MIRAMOND M.	UNITE DE RECHERCHE EN GENIE CIVIL - Hydrologie urbaine
MOREL R. (Prof. émérite)	MECANIQUE DES FLUIDES ET D'ACOUSTIQUES
MOSZKOWICZ P.	LAEPSI****
NARDON P. (Prof. émérite)	BIOLOGIE FONCTIONNELLE, INSECTES ET INTERACTIONS
NAVARRO Alain (Prof. émérite)	LAEPSI****
NELIAS D.	LAMCOS
NIEL E.	AUTOMATIQUE INDUSTRIELLE
NORMAND B.	GEMPPM
NORTIER P.	DREP
ODET C.	CREATIS**
OTTERBEIN M. (Prof. émérite)	LAEPSI****

PARIZET E.	VIBRATIONS-ACOUSTIQUE
PASCAULT J.P.	INGENIERIE DES MATERIAUX POLYMERES
PAVIC G.	VIBRATIONS-ACOUSTIQUE
PECORARO S.	GEMPPM
PELLETIER J.M.	GEMPPM***
PERA J.	UNITE DE RECHERCHE EN GENIE CIVIL - Matériaux
PERRIAT P.	GEMPPM***
PERRIN J.	INTERACTION COLLABORATIVE TELEFORMATION TELEACTIVITE
PINARD P. (Prof. émérite)	PHYSIQUE DE LA MATIERE
PINON J.M.	INGENIERIE DES SYSTEMES D'INFORMATION
PONCET A.	PHYSIQUE DE LA MATIERE
POUSIN J.	MODELISATION MATHEMATIQUE ET CALCUL SCIENTIFIQUE
PREVOT P.	INTERACTION COLLABORATIVE TELEFORMATION TELEACTIVITE
PROST R.	CREATIS**
RAYNAUD M.	CENTRE DE THERMIQUE DE LYON - Transferts Interfaces et Matériaux
REDARCE H.	AUTOMATIQUE INDUSTRIELLE
RETIF J-M.	CEGELY*
REYNOUARD J.M.	UNITE DE RECHERCHE EN GENIE CIVIL - Structures
RICHARD C.	LGEF
RIGAL J.F.	MECANIQUE DES SOLIDES
RIEUTORD E. (Prof. émérite)	MECANIQUE DES FLUIDES
ROBERT-BAUDOUY J. (Mme) (Prof. émérite)	GENETIQUE MOLECULAIRE DES MICRO ORGANISMES
ROUBY D.	GEMPPM***
ROUX J.J.	CENTRE DE THERMIQUE DE LYON – Thermique de l'Habitat
RUBEL P.	INGENIERIE DES SYSTEMES D'INFORMATION
SACADURA J.F.	CENTRE DE THERMIQUE DE LYON - Transferts Interfaces et Matériaux
SAUTEREAU H.	INGENIERIE DES MATERIAUX POLYMERES
SCAVARDA S. (Prof. émérite)	AUTOMATIQUE INDUSTRIELLE

SOUIFI A.	PHYSIQUE DE LA MATIERE
SOUROUILLE J.L.	INGENIERIE INFORMATIQUE INDUSTRIELLE
THOMASSET D.	AUTOMATIQUE INDUSTRIELLE
THUDEROZ C.	ESCHIL – Equipe Sciences Humaines de l'Insa de Lyon
UBEDA S.	CENTRE D'INNOV. EN TELECOM ET INTEGRATION DE SERVICES
VELEX P.	MECANIQUE DES CONTACTS
VERMANDE P. (Prof émérite)	LAEPSI
VIGIER G.	GEMPPM***
VINCENT A.	GEMPPM***
VRAY D.	CREATIS**
VUILLERMOZ P.L. (Prof. émérite)	PHYSIQUE DE LA MATIERE

Directeurs de recherche C.N.R.S.

BERTHIER Y.	MECANIQUE DES CONTACTS
CONDEMINE G.	UNITE MICROBIOLOGIE ET GENETIQUE
COTTE-PATAT N. (Mme)	UNITE MICROBIOLOGIE ET GENETIQUE
ESCUDIE D. (Mme)	CENTRE DE THERMIQUE DE LYON
FRANCIOSI P.	GEMPPM***
MANDRAND M.A. (Mme)	UNITE MICROBIOLOGIE ET GENETIQUE
POUSIN G.	BIOLOGIE ET PHARMACOLOGIE
ROCHE A.	INGENIERIE DES MATERIAUX POLYMERES
SEGUELA A.	GEMPPM***
VERGNE P.	LaMcos

Directeurs de recherche I.N.R.A.

FEBVAY G.	BIOLOGIE FONCTIONNELLE, INSECTES ET INTERACTIONS
GRENIER S.	BIOLOGIE FONCTIONNELLE, INSECTES ET INTERACTIONS
RAHBE Y.	BIOLOGIE FONCTIONNELLE, INSECTES ET INTERACTIONS

Directeurs de recherche I.N.S.E.R.M.

KOBAYASHI T.	PLM
PRIGENT A.F. (Mme)	BIOLOGIE ET PHARMACOLOGIE
MAGNIN I. (Mme)	CREATIS**

---

* CEGELY CENTRE DE GENIE ELECTRIQUE DE LYON
** CREATIS CENTRE DE RECHERCHE ET D'APPLICATIONS EN TRAITEMENT DE L'IMAGE ET DU SIGNAL
***GEMPPM GROUPE D'ETUDE METALLURGIE PHYSIQUE ET PHYSIQUE DES MATERIAUX
****LAEPSI LABORATOIRE D'ANALYSE ENVIRONNEMENTALE DES PROCEDES ET SYSTEMES INDUSTRIELS

Remerciements

Cette thèse m'a donnée l'occasion de rencontrer et de travailler avec des personnes absolument épatantes. Il est difficile de leur dire ici à quel point j'ai été touchée par tout ce qu'ils ont fait pour moi. Pour les remercier du fond du cœur, je profite de cette page pour leur offrir à chacun une citation.

À Jean-Michel Fayard, sans qui je ne me serai jamais lancée dans la fantastique aventure de cette thèse :

« On rencontre sa destinée, souvent par les chemins qu'on prend pour l'éviter » Jean de la Fontaine

À Hubert Charles, qui a toujours été là pour moi, même les jours de plus grand doute :

« Ce n'est pas parce que les choses sont difficiles que nous n'osons pas, c'est parce que nous n'osons pas, qu'elles sont difficiles. » Sénèque

À Gérard Febvay qui m'a accueillie chaleureusement dans son laboratoire :

« Celui qui sourit au lieu de se mettre en colère est toujours le plus fort. » Sagesse japonaise

À Federica Calevro, sans qui les expérimentations n'auraient certainement pas été aussi réussies :

« Apprends à écrire tes blessures dans le sable et à graver tes joies dans la pierre. » En souvenir de l'exposition Sable du muséum d'historie naturelle de Lyon.

À Yvan Rahbé, toujours prêt à répondre à mes nombreuses et incessantes questions :

« La gentillesse est le langage qu'un sourd peut entendre et qu'un aveugle peut voir. » Mark Twain

À Gaby, qui a élevé beaucoup des pucerons… travail ô combien fastidieux :

« Choisissez un travail que vous aimez et vous n'aurez pas à travailler un seul jour de votre vie. » Confucius

À Nicolas, compagnon d'infortune des analyses statistiques :

« S'il n'y a pas de solution c'est qu'il n'y a pas de problème. » Jacques Rouxel

À Christian Laugier, pour la pertinence de ses corrections sur le manuscript de cette thèse :

« La science est un jeu dont la règle du jeu consiste à trouver quelle est la règle du jeu. » François Cavanna

À Sandrine, Olivier, Cédric, Caroline, Alexandre, Cyril, Zeina, Florence, Marie et Luca pour tous les bons moments passés ensemble dans la salle des stagiaires :

« Le futur appartient à ceux qui croient à la beauté de leurs rêves. » Eleanor Roosevelt

À Laurent Duret et Bruno Spataro pour leur aide précieuse lors de l'installation parfois problématique de ROSO sur le site du PBIL :

« Si tu comprends, les choses sont comme elles sont, Si tu ne comprends pas, les choses sont comme elles sont. » Proverbe Zen

À Jacques Bernillon pour le spottage de toutes les lames autant que pour ses blagues et à Nathalie Allioli pour toutes les manips que nous avons réalisées côte à côte :

«Pour réaliser une chose vraiment extraordinaire, commencez par la rêver.» Walt Disney

Aux membres du groupe « Biologie et Modélisation des Systèmes Biologiques » qui m'ont donnée l'occasion de murir mes réflexions sur la biologie des systèmes :

« Aucun de nous ne sait ce que nous savons tous, ensemble. » Lao-Tseu

À Angela Douglas pour son accueil au sein du laboratoire de l'Université d'York, sans oublier bien sûr Claire, Tsumoto, et Camille :

L'anglais, ce n'est jamais que du français mal prononcé. » Georges Clemenceau.

Aux rapporteurs Marie-Claude Potier et Denis Tagu et aux membres du jury Christian Gautier et Alain Trubuil qui ont si gentiment accepté de juger ce travail :

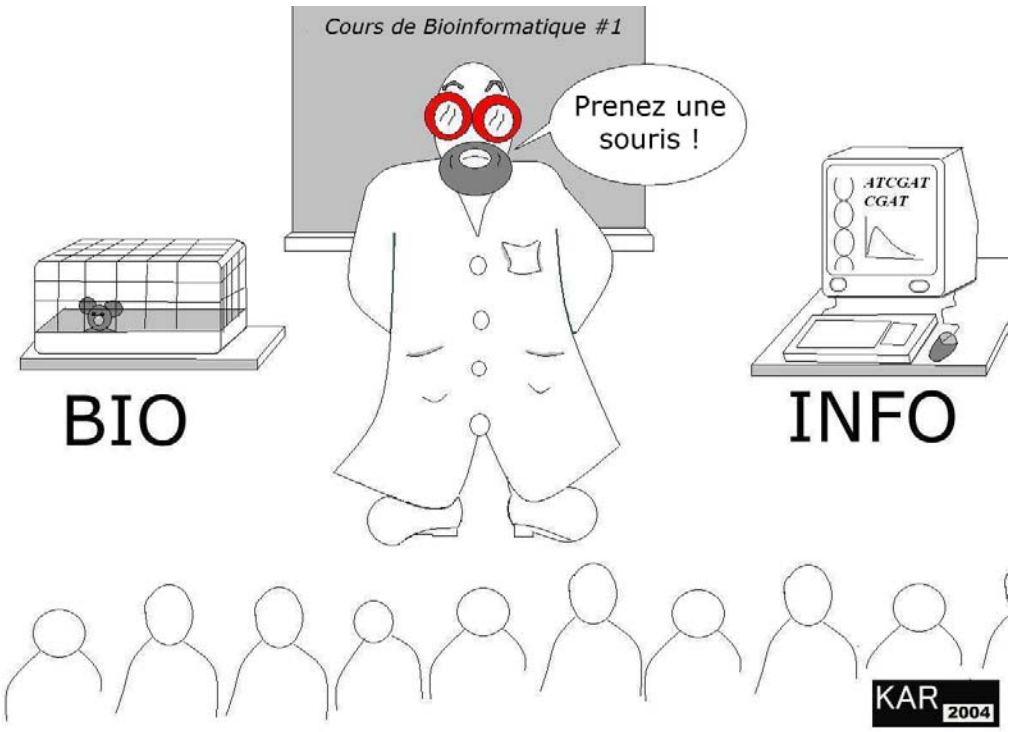
« Ce qui a échappé aux spectateurs pourra être remarqué par les lecteurs. » Jean Racine

Et bien sûr à tous les membres du labo pour leur sourire :

« Sourire est une lumière à la fenêtre d'un visage qui montre que le cœur est à l'intérieur. » Marc Twain

Cette thèse doit également beaucoup à mes proches et à ma famille qui ont toujours été là quand ça n'allait pas. Cette page m'offre l'occasion de dire à mes parents, à mon frère et à Karim à quel point leur soutien et leur amour ont été précieux. Enfin mes plus tendres pensées vont à mon grand-père qui aurait été tellement fier…

À Karim.


Cours de Bioinformatique #1
Prenez une souris !
ATCGAT
CGAT
BIO
INFO
KAR 2004

Sommaire

Avant-propos.....39

Partie A Étude bibliographique.....43

1 Contexte biologique 45

1.1 La symbiose.....45
1.1.1 Généralités..... 45
1.1.2 Une origine endocytobiotique pour la cellule eucaryote 47
1.1.3 Endocytobiose et unité de sélection 49
1.1.31 Impact sur l'hôte..... 49
1.1.32 Impact sur le symbiote..... 49
1.1.33 Impact sur le symbiocosme..... 50
1.1.4 Le concept d'individu..... 51

1.2 Les deux partenaires de l'association puceron - bactérie.52
1.2.1 Du côté du puceron *Acyrthosiphon pisum* 53
1.2.11 Caractéristiques biologiques..... 53
1.2.12 Nutrition 54
1.2.121 Composition du phloème 54
1.2.122 Prise alimentaire 54
1.2.123 Utilisation du saccharose 55
1.2.124 Utilisation des acides aminés..... 55
1.2.13 Localisation des symbiotes..... 56
1.2.2 Du côté de la bactérie *Buchnera aphidicola* 58

1.2.21 Un génome avec une organisation originale 58
1.2.211 Au niveau global 59
Une taille réduite..... 59
Une forte polyploïdie..... 60
La présence de deux plasmides 61
1.2.212 Au niveau des gènes..... 61
Répartition des gènes sur les brins précoce et tardif du chromosome 62
1.2.213 Au niveau des séquences..... 63
1.2.22 Un métabolisme essentiellement dédié à la biosynthèse des acides aminés 64
1.2.221 Les voies de biosynthèse des acides aminés . 64
1.2.222 Le métabolisme énergétique 67
1.2.223 Que reste-t-il des autres voies ?..... 68
1.2.224 Le transport des métabolites..... 68
1.2.23 Une régulation de l'expression des gènes plutôt énigmatique 70
1.2.231 De rares éléments de régulation 70
1.2.232 Régulation du choc thermique 72
1.2.233 Régulation de la biosynthèse des acides aminés 75
1.2.234 Et si la réponse était ailleurs ? 77
1.2.3 Le couple du point de vue de la problématique du métabolisme des acides aminés..... 81
1.2.31 Des études nutritionnelles 81
1.2.32 Des études métaboliques..... 82
1.2.4 Un couple ou un mariage à trois ? 84

1.3 Les objectifs de la thèse du point de vue biologique.....86

2 Contexte méthodologique Les puces à ADN de A à Z (ou presque) 89

2.1 Introduction 89
2.1.1 Historique..... 89
2.1.2 Principe des puces à ADN et analyse du transcriptome..... 91
2.1.3 Des applications variées pour les puces..... 94
2.1.31 Des puces à ADN..... 94

2.1.311 Criblage de mutations 94
2.1.312 Séquençage par hybridation 94
2.1.313 Étude phylogénétique 94
2.1.314 Applications industrielles 95
2.1.315 Caractérisation physique du génome 95
2.1.32 Et même des puces sans ADN ! 95
Étude du protéome 95
Étude des phénotypes 96

2.2 Conception des puces à ADN 96
2.2.1 Le support d'hybridation 96
2.2.11 Les supports dits neutres 97
2.2.12 Les supports dits actifs 98
2.2.2 Les sondes 98
2.2.21 Différents type de sondes 98
2.2.22 Les sondes de contrôle 99
2.2.3 Le plan de dépôt 100
2.2.4 L'adressage des sondes 100
2.2.41 La fixation de sondes synthétisées 101
2.2.42 La synthèse in situ *..... 101*

2.3 Hybridation des puces à ADN 102
2.3.1 Préparation du matériel biologique 102
2.3.11 Choix et préparation des échantillons 102
2.3.12 Choix des cibles de référence et de contrôle 103
2.3.2 Marquage 104
2.3.21 Marquage par synthèse d'ADNc 105
2.3.22 Marquage direct de l'ARN total 106
2.3.23 Marquage direct des ARNm 106
2.3.3 Hybridation 107
2.3.31 Principe 107
2.3.32 Choix des conditions d'hybridation 107

2.4 Acquisition des données 109
2.4.1 Acquisition des images 109
2.4.2 Analyse des images 111
2.4.3 Filtration des données 112

2.5 Transformation des données 113
2.5.1 Notation et représentation des données 113
2.5.2 Correction du bruit de fond 115

2.5.3 Normalisation des données 115
2.5.31 Introduction 115
2.5.32 Choix des gènes de normalisation 116
2.5.33 Normalisation globale 117
2.5.34 Normalisation intensité-dépendante 118
2.5.35 Normalisation intensité-dépendante par aiguille 119
2.5.36 Normalisation itérative 119
2.5.37 Normalisation mixte 120
2.5.38 Normalisation inter-lames 120

2.6 Analyse statistique des données 121
2.6.1 Des données complexes 121
2.6.2 Importance du plan expérimental 122
2.6.21 Unité expérimentale et répétitions 123
2.6.22 Associer les échantillons 124
2.6.221 Plan en flip-flop 124
2.6.222 Plan avec référence 125
2.6.223 Plan en boucle 125
2.6.3 Détection des gènes différentiellement exprimés 126
2.6.31 Comparaison de deux conditions 126
2.6.311 Comparaison intra-lame 126
2.6.312 Comparaison inter-lames 128
Approche classique : test de t 128
Approche bayésienne 129
2.6.32 Comparaison de plus de deux conditions 130
2.6.321 Modèle global d'analyse de variance 130
2.6.322 Modèle d'analyse de variance gène-spécifique 131
2.6.323 Tests de F 132
2.6.4 Choix d'un seuil de significativité 133
2.6.41 Contrôle de l'erreur FWER 133
2.6.42 Contrôle de l'erreur FDR 134
2.6.5 Détermination de profils d'expression 134
2.6.51 Classification 135
2.6.52 Analyse en composante principale 137

2.7 Relier les profils d'expression aux voies métaboliques .. 138

2.8 Et après ? 139
2.8.1 Valider les données 139
2.8.2 Formaliser et stocker les données 140

2.8.21 Formaliser les connaissances associées au transcriptome... 140
2.8.22 Stocker les données..... 142
2.8.3 Intégrer des données de nature différente 143
2.8.4 Puces et art !.... 144

2.9 Les objectifs de la thèse du point de vue méthodologique.... 144

Partie B Développements méthodologiques..... 147

1 Le logiciel ROSO 149

1.1 De la problématique du choix des sondes au développement du logiciel ROSO 150
1.1.1 Matériel et méthodes : comment choisir une « bonne sonde » ?..... 150
1.1.11 Étude de la spécificité 150
1.1.111 Présentation des logiciels BLAST et DUBLASTN 151
1.1.112 Détermination des paramètres d'utilisation de BLASTN et DUBLASTN..... 154
1.1.12 Étude du rendement d'hybridation entre deux brins d'ADN..... 157
1.1.121 Détermination de la température de fusion. 157
1.1.122 Énergie libre de formation des structures secondaires..... 159
1.1.123 Autres critères thermodynamiques d'intérêt pour le choix des sondes 160
Énergie libre de formation des pentamères situés aux extrémités des sondes. 160
Taux de GC 160
Nature des bases situées aux extrémités des sondes..... 161
Absence de séquences contenant trois G ou trois C successifs..... 161

1.1.13 Choix informatiques 161
1.1.2 Résultats : la conception 162
1.1.21 Les fichiers d'entrée 162
1.1.22 Préparation des jeux de données 163
1.1.23 Recherche de sondes spécifiques 164
1.1.24 Recherche de sondes optimales 166
1.1.241 Étude de la température de fusion des sondes 167
1.1.242 Étude des structures secondaires des sondes 167
1.1.243 Homogénéisation des températures de fusion des sondes 169
1.1.244 Choix final des sondes 169
1.1.245 Fonctionnalité annexe : insertion manuelle de mutations dans les sondes 170
1.1.25 Les fichiers de sortie 170
1.1.26 L'interface du Pôle Bioinformatique Lyonnais (PBIL) .. 171
1.1.3 Discussion 172
1.1.31 Validations et performance 172
1.1.311 Validation sur des données simulées 172
1.1.312 Validation sur des données réelles 173
1.1.32 Choix de l'utilisation de BLAST 174
1.1.33 Choix de l'utilisation du modèle thermodynamique du plus proche voisin 176
1.1.331 Calcul du Tm 177
1.1.332 Étude des structures secondaires 178
1.1.34 Un problème d'optimisation multicritère 180
1.1.35 Extensions possibles 180
1.1.351 Application aux sondes d'ADNc 180
1.1.352 Application aux sondes destinées aux génotypage 181
1.1.353 Application aux sondes multi-transcriptomes 182
1.1.354 Application aux autres acides nucléiques 182
1.1.36 Comparaison avec d'autres logiciels 183
1.1.4 Conclusion 186

1.2 Recherche d'un jeu de sondes optimales pour l'étude du transcriptome de *Buchnera* 186
1.2.1 Les séquences d'entrée 186
1.2.2 Les spécifications des sondes 187

1.2.3 La démarche d'optimisation 189
1.2.4 Résultats..... 190
1.2.41 Caractéristiques du jeu de sondes pour Buchnera 190
1.2.42 Caractéristiques des sondes de contrôle..... 191
1.2.5 Discussion..... 192
1.2.6 Conclusion..... 193

2 La puce *Buchnera* 195

2.1 Une mini-puce devenue grande 195

2.2 La puce complète 197

2.3 Conclusion 199

Partie C Analyse du transcriptome de *Buchnera aphidicola*..... 201

1 Introduction..... 203

1.1 Contexte 203

1.2 Problématique 204

1.3 Plan expérimental 205

2 Matériel et méthodes 207

2.1 Préparation des lames..... 207
2.1.1 Dépôt des sondes 207
2.1.2 Traitement des lames 209

2.2 Préparation du matériel biologique..... 209

2.2.1 Élevage des pucerons 209
2.2.2 Purification des bactéries par filtration 210
2.2.3 Extraction des ARN 210
2.2.4 Préparation des cibles de contrôle 211

2.3 Marquage 211

2.4 Hybridation et lavages 213

2.5 Acquisition des images 213

2.6 Acquisition et de filtration des données 214

2.7 Analyse statistique 215
2.7.1 Étude du bruit de fond 215
2.7.2 Normalisation des données 216
2.7.3 Détection des gène exprimés de façon différentielle 217
2.7.31 Obtention des données moyennées 217
2.7.32 Analyse de variance 217
2.7.33 Test de F 218
2.7.34 Classification K-means *218*

3 Résultats 219

3.1 Étude physiologique de l'impact de la composition des milieux nutritionnels sur les pucerons 219

3.2 Acquisition et normalisation des données d'expression . 221

3.3 Analyse statistique 226

3.4 Interprétation biologique 229
3.4.1 Relation entre niveaux d'expression et organisation du génome 229
3.4.2 Relation entre niveaux d'expression et fonction des gènes 235
3.4.3 Relation entre niveaux d'expression et métabolisme 240

4 Discussion 245

4.1 Aspects statistiques 245

4.2 Aspects biologiques 249

5 Conclusion 255

6 Vers une biologie des systèmes... 257

Partie D Conclusion générale et perspectives 261

Publications et communications 269

Partie E Bibliographie 271

Partie F Annexes 307

1 Paramètres thermodynamiques du modèle du plus proche voisin 309

1.1 Termes d'initiation et de symétrie 309

1.2 Cas des hybrides homologues.....309

1.3 Cas des hybrides avec mésappariements 310

1.4 Paramètres d'énergie libre de formation des boucles....312

2 Protocoles d'utilisation des lames QUANTIFOIL® et ROSA® en hybridation manuelle 313

2.1 Traitement des lames.....313

2.2 Pré-hybridation.....313

2.3 Hybridation et lavages 314

3 Préparation des milieux artificiels pour l'élevage des pucerons.....315

3.1 Milieu AP3 315

3.2 Milieux utilisés à l'Université d'York.....317

4 Listes des gènes exprimés de façon différentielle 319

Liste des figures

FIGURE A.1.1 LARVES DE PUCERONS SUR LEUR PLANTE HÔTE (À GAUCHE) ET PUCERON ADULTE AVEC SES LARVES EN VUE RAPPROCHÉE (À DROITE) (RAHBÉ, Y., COMMUNICATION PERSONNELLE). 54
FIGURE A.1.2 DES BACTÉRIOCYTES DE PUCERONS DU POIS ADULTES (OBSERVATION PAR MICROSCOPIE ÉLECTRONIQUE À TRANSMISSION) (RAHBÉ, Y., COMMUNICATION PERSONNELLE). 57
FIGURE A.1.3 UNE BACTÉRIE *BUCHNERA* ENTOURÉE DE SES DEUX MEMBRANES (COMME TOUTES LES BACTÉRIES GRAM NÉGATIVE) ET D'UNE TROISIÈME MEMBRANE D'ORIGINE EUCARYOTE : LA MEMBRANE SYMBIOSOMALE (OBSERVATION EN MICROSCOPIE ÉLECTRONIQUE À TRANSMISSION) (RAHBÉ, Y., COMMUNICATION PERSONNELLE). 58
FIGURE A.1.4 REPRÉSENTATION DES VOIES DE BIOSYNTHÈSE DES ACIDES AMINÉS ESSENTIELS (À GAUCHE) ET NON ESSENTIELS (À DROITE) DÉDUITES DE L'ANNOTATION DU GÉNOME DE *BUCHNERA* (D'APRÈS SHIGENOBU *ET AL.*, 2000). LES ÉTAPES POUR LESQUELLES AUCUN GÈNE N'A ÉTÉ ANNOTÉ SONT INDIQUÉES EN ROSE. 66
FIGURE A.1.6 MODÈLE DES INTERACTIONS EXISTANT ENTRE *BUCHNERA* ET SON HÔTE POUR LE MÉTABOLISME DES ACIDES AMINÉS. *BUCHNERA* EST CAPABLE DE SYNTHÉTISER LES ACIDES AMINÉS QUI SONT ESSENTIELS POUR SON HÔTE (EN ROUGE) ET SEMBLE AVOIR BESOIN DE PLUSIEURS ACIDES AMINÉS NON ESSENTIELS (EN BLEU). PARMI CES ACIDES AMINÉS, LES DEUX PRINCIPAUX SONT L'ASPARTATE ET LE GLUTAMATE (EN ROSE) (D'APRÈS ZIENTZ *ET AL.*, 2001). 84
FIGURE A.1.7 DISTRIBUTION DES SYMBIOTES PRIMAIRES PASS (EN ROUGE) ET SECONDAIRES (EN VERT) CHEZ UN EMBRYON DE PUCERON DU POIS (IMAGE OBTENUE PAR FISH ET OBSERVATION AU MICROSCOPE À TRANSMISSION), (KOGA, R., COMMUNICATION PERSONNELLE). 85
FIGURE A.2.1 NOMBRE DE PUBLICATIONS CONCERNANT LES PUCES À ADN DE 1994 À 2004 (D'APRÈS LA BASE ENTREZ-PUBMED, MOTS-CLÉS : MICROARRAY* OU DNA CHIP*). 90
FIGURE A.2.2 PRINCIPE DE LA TECHNOLOGIE DES À PUCES À ADN. (A) LES SÉQUENCES DES SONDES SONT DÉTERMINÉES DE FAÇON À OPTIMISER LEUR SPÉCIFICITÉ ET LEUR SENSIBILITÉ. LES SONDES SYNTHÉTISÉES SONT DÉPOSÉES PAR UN ROBOT SUR LA SURFACE DE LA LAME SELON UN PLAN DÉFINI. (B) LES ARNM SONT EXTRAITS DES ÉCHANTILLONS BIOLOGIQUES À COMPARER, MARQUÉS AVEC DEUX FLUOROCHROMES DIFFÉRENTS PUIS MÉLANGÉS AVANT HYBRIDATION. (C) LA LECTURE DES LAMES EST RÉALISÉE AVEC UN SCANNER (MICROSCOPE À FLUORESCENCE) COUPLÉ À UN

PHOTOMULTIPLICATEUR (PMT). (D) L'IMAGE EST ALORS ANALYSÉE DE FAÇON À QUANTIFIER LE SIGNAL. LES DONNÉES SONT ENSUITE NORMALISÉES, ANALYSÉES ET INTERPRÉTÉES. 93
FIGURE A.2.3 VUE RAPPROCHÉE DES AIGUILLES D'UN ROBOT DE DÉPÔT. 101
FIGURE A.2.4 DESCRIPTION DES PROTOCOLES DE MARQUAGE DIRECT (À GAUCHE) ET INDIRECT (À DROITE) POUR L'INCORPORATION DES FLUOROCHROMES CY3 ET CY5 (D'APRÈS YU *ET AL.*, 2002). 106
FIGURE A.2.5 PRINCIPE DE LA SEGMENTATION DU SIGNAL (EN GRIS FONCÉ) ET DU BRUIT DE FOND (EN NOIR) RÉALISÉE PAR LE LOGICIEL *GENEPIX*. UNE RÉGION D'EXCLUSION LOCALISÉE ENTRE LES ZONES CORRESPONDANT AU SIGNAL ET AU BRUIT DE FOND N 'EST PAS PRISE EN COMPTE DANS LE CALCUL DES INTENSITÉS (D'APRÈS LE MANUEL *« GENEPIX PRO 4.0 ARRAY ACQUISITION AND ANALYSIS SOFTWARE FOR THE GENEPIX 4000B, REVIEW E »*). 112
FIGURE A.2.6 EXEMPLE DE REPRÉSENTATION MA POUR L'ENSEMBLE DES INTENSITÉS OBSERVÉES SUR UNE LAME.....114
FIGURE A.2.7 DISTRIBUTION DES NIVEAUX D'EXPRESSION (UNITÉ RELATIVE DE FLUORESCENCE EN ABSCISSE) CHEZ LA BACTÉRIE *ESCHERICHIA COLI* (5522 GÈNES) ET CHEZ L'HOMME (12 632 GÈNES). LES DIAGRAMMES (EN HAUT) MONTRENT L'ÉTIREMENT DE LA DISTRIBUTION VERS LES FORTES VALEURS D'EXPRESSION (CE QUI SIGNIFIE QUE SEUL UN FAIBLE NOMBRE DE GÈNES EST FORTEMENT EXPRIMÉ.) LA COURBE (EN BAS) REPRÉSENTE LA DISTRIBUTION NORMALE THÉORIQUE. LES DONNÉES SONT ISSUES DE LA BASE D'EXPRESSION GEO (NUMÉROS D'ACCESSION : GSE33 ET GSE516)....122
FIGURE A.2.8 EXEMPLE DE PLAN DE COMPARAISON DIRECT DE DEUX ÉCHANTILLONS (A ET B). PLUSIEURS ASSOCIATIONS SONT POSSIBLES : EN FLIP-FLOP (A), EN FLIP-FLOP AVEC RÉPÉTITIONS TECHNIQUES, EN FLIP-FLOP AVEC RÉPÉTITIONS BIOLOGIQUES (C) ET EN BOUCLE (D) (D'APRÈS CHURCHILL, 2002). 125
FIGURE A.2.9 EXEMPLES DE PLAN AVEC RÉFÉRENCE. LE PLAN CLASSIQUE UTILISE UNE LAME POUR CHAQUE ÉCHANTILLON (A,B, C…Z) (A). UNE VARIATION POSSIBLE EST DE COUPLER CE PLAN AVEC DES FLIP-FLOP POUR CHAQUE ÉCHANTILLON (B) (D'APRÈS CHURCHILL, 2002). 125
FIGURE A.2.10 EXEMPLE DE PLAN EN BOUCLE POUR TROIS ÉCHANTILLONS DIFFÉRENTS (N,S ET G) (D'APRÈS CHURCHILL, 2002). 126
FIGURE A.2.11 EXEMPLE DU GRAPHE OBTENU POUR UNE ANALYSE DE NEWTON. LES POINTS EN ROUGE SONT LES POINTS CONSIDÉRÉS COMME DIFFÉRENTIELLEMENT EXPRIMÉS. 128
FIGURE A.2.12 REPRÉSENTATION GRAPHIQUE ISSUE D'UNE ANALYSE PAR CLASSIFICATION HIÉRARCHIQUE. À PARTIR DES TABLEAUX DE MESURES DANS LESQUELS CHAQUE COLONNE CORRESPOND À UNE EXPÉRIENCE ET CHAQUE LIGNE À UN GÈNE (À GAUCHE), CHAQUE VALEUR EST REPRÉSENTÉE PAR UNE COULEUR QUI EST LE REFLET QUALITATIF ET QUANTITATIF DU RAPPORT DES FLUORESCENCES (AU CENTRE). LES GÈNES DONT LE NIVEAU D'EXPRESSION EST INCHANGÉ SONT REPRÉSENTÉS EN NOIR, CEUX DONT LE NIVEAU AUGMENTE EN ROUGE ET LES AUTRES EN VERT. UNE ÉCHELLE DE COULEURS PERMET DE QUANTIFIER LES NUANCES DE VERT ET DE ROUGE. FINALEMENT, LES GÈNES AYANT LE MÊME PROFIL D'EXPRESSION SUR PLUSIEURS EXPÉRIENCES SONT REGROUPÉS (À DROITE) (D'APRÈS BERTUCCI *ET AL.*, 2002).....136

FIGURE A.2.13 UN EXEMPLE DE L'*ART ARRAYS* INTITULÉ « COSMIQLOW » MAIS QUI POURRAIT ÉGALEMENT S'APPELER « LA BEAUTÉ DE L'ERREUR ». 144
FIGURE B.1.1 EXEMPLE DE DEUX SÉQUENCES AU FORMAT FASTA. 152
FIGURE B.1.2 EXEMPLE D'UN FICHIER DE SORTIE DE *DUBLASTN*.....154
FIGURE B.1.3 EXEMPLE DE LA COURBE DE FUSION OBTENUE POUR UN OLIGONUCLÉOTIDE (D'APRÈS MERGNY ET LACROIX, 2002). LA DÉNATURATION D'UNE MOLÉCULE D'ADN S'ACCOMPAGNE D'UNE AUGMENTATION DE L'ABSORPTION LUMINEUSE À 260 NM APPELÉE AUSSI EFFET HYPERCHROME (CHESTER ET MARSHAK, 1993). CETTE DÉNATURATION S'EFFECTUE GÉNÉRALEMENT DANS UNE ZONE DE TEMPÉRATURE RESTREINTE DONT LE POINT MÉDIAN CORRESPOND À LA TEMPÉRATURE DE FUSION OU TM. 158
FIGURE B1.4 EXEMPLE DE FORMATION D'UN HOMODIMÈRE ET D'UNE ÉPINGLE À CHEVEUX. 160
FIGURE B.1.5 PRÉSENTATION DES DIFFÉRENTS FICHIERS UTILISÉS AU COURS DU DÉVELOPPEMENT. ILS SONT NÉCESSAIRES À LA CRÉATION DE L'EXÉCUTABLE *ROSO*. 162
FIGURE B.1.6 ORGANIGRAMME DE L'ÉTAPE PRÉLIMINAIRE. LES FICHIERS EN BLEUS SONT LES TROIS FICHIERS D'ENTRÉE DES ÉTAPES DE RECHERCHE DE SONDES À PROPREMENT PARLER. 164
FIGURE B.1.7 EXEMPLE DE DÉCOUPAGE EN RÉGIONS POUR LE GÈNE 1 EN FONCTION DES TAUX D'IDENTITÉ DES DIFFÉRENTS HITS.....165
FIGURE B.1.8 ORGANIGRAMME DU PREMIER MODULE DE ROSO DÉDIÉ À LA RECHERCHE DE SONDES SPÉCIFIQUES.....166
FIGURE B.1.9 ORGANIGRAMME DU SECOND MODULE DE ROSO DÉDIÉ À LA RECHERCHE DE SONDES OPTIMALES. 167
FIGURE B.1.10 EXEMPLE DE L'UTILISATION DE L'ALGORITHME DE RECHERCHE DES ÉPINGLES À CHEVEUX ET DU CALCUL DE L'ÉNERGIE LIBRE ASSOCIÉE À LA CONFORMATION LA PLUS STABLE. 168
FIGURE B.1.11 PAGE D'ACCUEIL POUR L'UTILISATION EN LIGNE DE ROSO SUR LE SITE DU PBIL. 171
FIGURE B.1.12 GRAPHE PRÉSENTANT LES VALEURS D'ÉNERGIE LIBRE DE FORMATION DES ÉPINGLES À CHEVEUX OBTENUES AVEC LE LOGICIEL *MFOLD* EN FONCTION DES ÉCARTS ENTRE VALEURS CALCULÉES PAR ROSO ET PAR *MFOLD*) POUR LES SONDES DE 70 MERS. CES VALEURS ONT ÉTÉ CALCULÉES POUR TROIS TEMPÉRATURES DIFFÉRENTES D'HYBRIDATION : 25 °C, 42 °C ET 65 °C. 173
FIGURE B.1.13 LOCALISATION DES DIFFÉRENTES SONDES SUR LES GÈNES DE *BUCHNERA*. 189
FIGURE B.1.14 RÉPARTITION DES SONDES DE *BUCHNERA* EN FONCTION DES VALEURS DE TM.....191
FIGURE B.1.15 RÉPARTITION DES SONDES DE *BUCHNERA* EN FONCTION DES TAUX DE SIMILITUDE. SEULES LES SONDES QUI PRÉSENTENT DES TAUX DE SIMILITUDE SUPÉRIEURS À 80 % POSSÈDENT DES RISQUES RÉELS D'HYBRIDATION ASPÉCIFIQUE. 191
FIGURE B.1.16 EXEMPLE DE LA SONDE NORMALE ET DE LA SONDE TÉMOIN POUR LE GÈNE *AROH*.....192
FIGURE B.2.1 PLAN DE LA PREMIÈRE MINI-PUCE *BUCHNERA*. 196
FIGURE C.1.1 SCHÉMA DU PLAN EXPÉRIMENTAL ASSOCIANT LES HUIT LAMES (RÉPÉTÉ DEUX FOIS). CHAQUE FLÈCHE REPRÉSENTE UNE HYBRIDATION SUR UNE MÊME LAME ET LES CHIFFRES

CORRESPONDENT AUX MILIEUX NUTRITIONNELS SUR LESQUELS ONT ÉTÉ ÉLEVÉS LES PUCERONS.....205
FIGURE C.2.1 PLAN DE DÉPÔT DE LA PUCE *BUCHNERA*. 208
FIGURE C.2.2 DISPOSITIF EXPÉRIMENTAL DE PRÉPARATION DES MILIEUX ARTIFICIELS.....210
FIGURE C.3.1 REPRÉSENTATION DU POIDS MOYEN DES PUCERONS EN FONCTION DU TAUX D'ACIDES AMINÉS ESSENTIELS ET DE LA CONCENTRATION EN SACCHAROSE DANS LE MILIEU NUTRITIONNEL. 221
FIGURE C.3.2 EXEMPLE DE RÉSULTAT POUR UNE HYBRIDATION EFFECTUÉE SUR LA PUCE *BUCHNERA* (LAME 16) ET APRÈS SUPERPOSITION DES IMAGES OBTENUES POUR CHACUN DES DEUX FLUOROCHORMES AVEC LE LOGICIEL *GENEPIX*. 222
FIGURE C.3.3 REPRÉSENTATION MA POUR LA LAME 2 (EN HAUT) ET LA LAME 10 (EN BAS) AVANT ET APRÈS NORMALISATION AVEC $M=LOG_2(R/G)$ ET $A=LOG_2(RG)^{0,5}$. LES POINTS EN ROUGE REPRÉSENTENT LES PLOTS INVARIANTS QUI SONT UTILISÉS POUR LA NORMALISATION. 223
FIGURE C.3.4 REPRÉSENTATION MA POUR LA LAME 3 (À GAUCHE) ET LA LAME 11 (À DROITE) AVANT ET APRÈS NORMALISATION. LES POINTS EN ROSE REPRÉSENTENT LES RÉPONSES DES PLOTS INVARIANTS UTILISÉS POUR LA NORMALISATION ET LES POINTS BLEUS LES RÉPONSES DES PLOTS CONTENANT DU TAMPON. LES POINTS EN ROUGE REPRÉSENTENT LES INTENSITÉS DES CIBLES DE CONTRÔLE *PELL* INTRODUITES EN RAPPORT 1:1 (CE RAPPORT EST REPRÉSENTÉ PAR LA LIGNE GRISE M=0) ET LES POINTS EN VERT CELLES DES CIBLES *PELK* INTRODUITES EN RAPPORT 1:3 (CE RAPPORT EST REPRÉSENTÉ PAR LA LIGNE HORIZONTALE VERTE). LES STRUCTURES PARTICULIÈRES SITUÉES À GAUCHE DES GRAPHES APRÈS NORMALISATION REPRÉSENTENT LES PLOTS CORRIGÉS PAR LA MÉTHODE D'EDWARDS (2003). POUR CETTE EXPÉRIENCE, ILS SERONT TOUS ÉLIMINÉS AU COURS DE L'ÉTAPE DE MOYENNAGE QUALITÉ DÉPENDANTE.....224
FIGURE C.3.5 REPRÉSENTATION DES RAPPORTS D'INTENSITÉS POUR LA LAME 16 (EN ROUGE) ET LA LAME 13 (EN VERT) POUR LES ARN DE TRANSFERT (À GAUCHE) ET POUR L'ENSEMBLE DES GÈNES (À DROITE). 226
FIGURE C.3.6 GRAPHE « VOLCANO » REPRÉSENTANT LES GÈNES DIFFÉRENTIELLEMENT EXPRIMÉS POUR L'EFFET *SACCHAROSE*. LES POINTS EN BLEU REPRÉSENTENT LES GÈNES POUR LESQUELS AUCUN EFFET N'EST OBSERVÉ, LES POINTS EN ORANGE LES GÈNES DÉTECTÉS PAR LE TEST *FS* ET LES POINTS EN VERT SONT LES GÈNES DÉTECTÉS PAR LE TEST F_3. 227
FIGURE C.3.7 REPRÉSENTATIONS DES GÈNES EN GROUPES DE PROFILS D'EXPRESSION POUR LA LISTE DE GÈNES PRÉSENTANT UNE INTERACTION SIGNIFICATIVE. 228
FIGURE C.3.8 REPRÉSENTATION DES NIVEAU D'EXPRESSION MOYENS CALCULÉS À PARTIR DES MESURES D'INTENSITÉS RÉELLES EN FONCTION DES NIVEAUX D'EXPRESSION ESTIMÉS PAR L'ANALYSE DE VARIANCE. 230
FIGURE C.3.9 REPRÉSENTATION DES FRÉQUENCES DES NIVEAUX D'EXPRESSION (SANS TRANSFORMATION LOGARITHMIQUE) SUR LE BRIN DIRECT (EN HAUT) ET INDIRECT (EN BAS). 231

FIGURE C.3.10 REPRÉSENTATION DU TAUX DE GC DES GÈNES CHEZ *BUCHNERA* EN FONCTION DE LEUR NIVEAU MOYEN D'EXPRESSION.232
FIGURE C.3.11 REPRÉSENTATION DU NIVEAU MOYEN D'EXPRESSION DES GÈNES DE *BUCHNERA* EN FONCTION DU *CAI* DES GÈNES HOMOLOGUES CHEZ *ESCHERICHIA* COLI. 234
FIGURE C.3.12 REPRÉSENTATION DES FRÉQUENCES D'APPARITION DE 6 GRANDES CATÉGORIES FONCTIONNELLES AU SEIN DE LA CLASSIFICATION DES GÈNES OBTENUS APRÈS ANALYSE DE VARIANCE (PROBABILITÉS DU TEST *FS*) POUR L'EFFET *AA* (EN HAUT) ET L'EFFET *SACCHAROSE* (EN BAS). 237
FIGURE C.3.13 REPRÉSENTATION DES FRÉQUENCES D'APPARITION DE 17 CATÉGORIES FONCTIONNELLES AU SEIN DE LA CLASSIFICATION DES GÈNES OBTENUS APRÈS ANALYSE DE VARIANCE (PROBABILITÉS DU TEST *FS*) POUR L'EFFET *AA* (EN HAUT) ET L'EFFET *SACCHAROSE* (EN BAS).....238
FIGURE C.3.14 RÉPARTITION DES GÈNES SUR-EXPRIMÉS (EN ROUGE) ET SOUS-EXPRIMÉS (EN JAUNE) POUR LES DIFFÉRENTES CONDITIONS D'ÉTUDE (À PARTIR DES RÉSULTATS DES TESTS *FS* ET F_3). POUR L'EFFET « ACIDES AMINÉS ESSENTIELS » LES RAPPORTS UTILISÉS SONT 50 % : 25 % ET POUR L'EFFET « CONCENTRATION EN SACCHAROSE » LES RAPPORTS SONT 0,5 M : 1 M.....239
FIGURE C.3.15 REPRÉSENTATION THÉORIQUE DE LA VOIE DE BIOSYNTHÈSE DE LA PHÉNYLALANINE, DE LA TYROSINE ET DU TRYPTOPHANE (CARTE MÉTABOLIQUE DISPONIBLE DANS KEGG). LES ENZYMES EN VERT CORRESPONDENT AUX GÈNES PRÉSENTS CHEZ *BUCHNERA* DONT L'EXPRESSION NE VARIE PAS. LES ENZYMES EN BLEU CORRESPONDENT AUX GÈNES QUI SONT SUREXPRIMÉS EN CONDITION DE STRESS OSMOTIQUE, QUEL QUE SOIT LE TAUX D'ACIDES AMINÉS ESSENTIELS OU POUR 25 % D'ACIDES AMINÉS ESSENTIELS. LES ENZYMES EN ROUGE CORRESPONDENT EN REVANCHE AUX GÈNES SOUS-EXPRIMÉS. 242
FIGURE C.3.16 REPRÉSENTATION THÉORIQUE DE LA VOIE DES PENTOSES PHOSPHATES (CARTE MÉTABOLIQUE DISPONIBLE DANS KEGG). LES ENZYMES EN VERT CORRESPONDENT AUX GÈNES PRÉSENTS CHEZ *BUCHNERA* DONT L'EXPRESSION NE VARIE PAS ET LES ENZYMES EN ROUGE CORRESPONDENT AUX GÈNES QUI SONT SOUS-EXPRIMÉS EN CONDITION DE STRESS OSMOTIQUE, QUEL QUE SOIT LE TAUX D'ACIDES AMINÉS ESSENTIELS OU POUR 25 % D'ACIDES AMINÉS ESSENTIELS. AUCUN DES GÈNES NE PRÉSENTE DE SUREXPRESSION. 243
FIGURE C.3.17 REPRÉSENTATION THÉORIQUE DE LA FORMATION DU FLAGELLE (CARTE MÉTABOLIQUE DISPONIBLE DANS KEGG). LES PROTÉINES EN VERT CORRESPONDENT AUX GÈNES PRÉSENTS CHEZ *BUCHNERA* DONT L'EXPRESSION NE VARIE PAS. LES PROTÉINES EN BLEU CORRESPONDENT AUX GÈNES QUI SONT SUREXPRIMÉS EN CONDITION DE STRESS OSMOTIQUE, QUEL QUE SOIT LE TAUX D'ACIDES AMINÉS ESSENTIELS OU POUR 25 % D'ACIDES AMINÉS ESSENTIELS. AUCUN DES GÈNES NE PRÉSENTE DE SOUS-EXPRESSION. 244
FIGURE C.4.1 GRAPHES DES RÉSIDUS OBTENUS LORS DE L'AJUSTEMENT DU MODÈLE D'ANOVA AVEC INTERACTION, POUR CHACUN DES DEUX FLUOROCHROMES. 248

FIGURE C.6.1 SI VOUS POUVIEZ GROSSIR UNE CELLULE UN MILLION DE FOIS, POUR OBTENIR DES MOLÉCULES DE LA TAILLE DES OBJETS QUOTIDIENS, QUE VERRIEZ-VOUS ? 259
FIGURE D.1 REPRÉSENTATION SCHÉMATIQUE DU TRAVAIL RÉALISÉ DANS CETTE THÈSE POUR CHACUNE DES ÉTAPES D'UTILISATION D'UNE PUCE À ADN. 264
FIGURE F.1.1 EXEMPLE DE LECTURE D'UN COUPLE DE NUCLÉOTIDES DANS LA TABLE CONTENANT LES DONNÉES THERMODYNAMIQUES D'ÉNERGIE D'INTERACTION ENTRE BASES PARFAITEMENT HOMOLOGUES. SUR LA SÉQUENCE, POUR CHAQUE COUPLE DE BASES CONTIGUËS, LA BASE SITUÉE À GAUCHE (BASE G) EST LUE EN LIGNE ET LA BASE SITUÉE À DROITE (BASE D) EN COLONNE.....310
FIGURE F.1.2 EXEMPLE DE LECTURE D'UN COUPLE DE NUCLÉOTIDES DANS LA TABLE CONTENANT LES DONNÉES THERMODYNAMIQUES D'ÉNERGIE D'INTERACTION ENTRE BASES POUVANT PRÉSENTER DES MÉSAPPARIEMENTS.....312

Liste des tableaux

TABLEAU A.1.1 NOMBRE DE GÈNES CHEZ BUCHNERA PAR CATÉGORIE FONCTIONNELLE EN FONCTION DE LA CLASSIFICATION DE RILEY (1993). LA LISTE COMPLÈTE DES GÈNES EST DISPONIBLE SUR LE SITE DE L'INSTITUT RIKEN : HTTP://BUCHNERA.GSC.RIKEN.GO.JP. 65
TABLEAU B.1.1 RÉCAPITULATIF DES PRINCIPAUX PARAMÈTRES DU LOGICIEL *BLAST* AVEC LEURS VALEURS PAR DÉFAUT. LA RECHERCHE PAR DÉFAUT EST RÉALISÉE SANS ÉLIMINER LES SÉQUENCES DE FAIBLES COMPLEXITÉ (-F F), SUR LES DEUX BRINS D'ADN (-S 3), AVEC UN MOT DE 11 BASES (-W 11) ET DES VALEURS DE GAIN ET DE PÉNALITÉ AU MOMENT DE L'EXTENSION RESPECTIVEMENT DE 1 ET –3 (–R 1 -Q –3). LA TAILLE DE LA BASE EST ADAPTÉE AUTOMATIQUEMENT À LA BASE FOURNIE EN ENTRÉE (-Z 0) ET LES RÉSULTATS POUR LESQUELS LA VALEUR DE E EST INFÉRIEURE À 10 (-E 10) NE SONT PAS AFFICHÉS. 153
TABLEAU B.1.2 RÉSULTATS DES RECHERCHES DE *BLASTN* AVEC LES FICHIERS DE SÉQUENCES CONTENANT UNE SOUS-SÉQUENCE DE 35, 50 ET 100 BASES AVEC DIFFÉRENTS TAUX DE SIMILITUDE. LES COLONNES 3 ET 4 PRÉSENTENT LES RÉSULTATS OBTENUS AVEC LES PARAMÈTRES PAR DÉFAUT ET LES COLONNES 5 ET 6 CEUX QUI SONT OBTENUS AVEC LES PARAMÈTRES AJUSTÉS. ND SIGNIFIE QUE LA SOUS-SÉQUENCE N'A PAS ÉTÉ DÉTECTÉE. 156
TABLEAU B.1.3 RÉSULTATS DES RECHERCHES DE *BLASTN* AVEC LES FICHIERS CONTENANT DES SÉQUENCES DE 1000 BASES AVEC DIFFÉRENTS TAUX D'IDENTITÉ. À PARTIR DE 75 % D'IDENTITÉ, LES PARAMÈTRES PAR DÉFAUT NE PERMETTENT PLUS UNE DÉTECTION CORRECTE DE LA SÉQUENCE DE 1000 BASES (DÉTECTION UNIQUEMENT D'UNE SÉQUENCE DE 72 BASES À 95 % D'IDENTITÉ). POUR DES TAUX D'IDENTITÉ INFÉRIEUR 70 % SEULS LES PARAMÈTRES AJUSTÉS PERMETTENT DE DÉTECTER LA SÉQUENCE ET SONT DONC PRÉSENTÉS DANS LE TABLEAU. 156
TABLEAU B.1.4 RÉSULTATS DES RECHERCHES DE *BLASTN* AVEC LES FICHIERS CONTENANT DES SÉQUENCES DE 1000 BASES AVEC UNE DIVERGENCE GRADUELLE TOUTES LES 42 PAIRES DE BASES COMPRISE ENTRE 1 ET 50 %. 157
TABLEAU B.1.5 PREMIÈRES UTILISATIONS DU LOGICIEL ROSO. 174
TABLEAU B.1.6 PRÉSENTATION DES PRINCIPAUX LOGICIELS DE DÉTERMINATION DE SONDES OLIGONUCLÉOTIDIQUES. LES RÈGLES DE LOCKHART *ET AL.* (1996) QUI SONT UTILISÉES PAR CERTAINS LOGICIELS SONT UN ENSEMBLE DE RÈGLES DE DÉCISION QUI PERMETTENT D'EXCLURE LES SONDES QUI VÉRIFIENT AU MOINS UNE DES CONDITIONS SUIVANTES : (1) LE NOMBRE DE BASES SEULES (A,T, C OU G) DÉPASSE LA MOITIÉ DE LA TAILLE DE LA SONDE, (2) LA

LONGUEUR TOTALE DE SÉQUENCES DE BASES IDENTIQUES DÉPASSE UN QUART DE LA SÉQUENCE, (3) LE TAUX DE GC EST AU-DESSOUS DE 40 % OU AU-DESSUS DE 60 % ET (4) IL N'EXISTE PAS DE SÉQUENCES PALINDROMIQUES DANS LA SÉQUENCE. L'ABRÉVIATION PVV, QUI EST UTILISÉE DANS LE TABLEAU, SIGNIFIE QUE LE LOGICIEL UTILISE LE MODÈLE THERMODYNAMIQUE DU PLUS PROCHE VOISIN POUR LE CALCUL DU TM.....184

TABLEAU B.1.7 LISTE DES NEUF PSEUDOGÈNES DE *BUCHNERA APHIDICOLA* (D'APRÈS LES DONNÉES OBTENUES SUR LE SITE DE L'INSTITUT RIKEN.....187

TABLEAU B.1.8 PRÉSENTATION DES CINQ ÉTAPES DE LA DÉMARCHE DE RECHERCHE DES SONDES POUR *BUCHNERA*. LES DIFFÉRENTS CRITÈRES QUI SONT UTILISÉS SONT : (1) N4 : ABSENCE DE SÉQUENCES DE 4 NUCLÉOTIDES IDENTIQUES, (2) LOCALISATION : DEUX ZONES SONT EXCLUES 50 BASES EN 5' ET 100 BASES EN 3', (3) SEUIL DE REJET DES ÉNERGIES LIBRES DE FORMATION DES ÉPINGLES À CHEVEUX ET DES HOMODIMÈRES (EN KCAL.MOL^{-1}).....189

TABLEAU B.1.9 DÉMARCHE GÉNÉRALE D'OPTIMISATION PERMETTANT UNE UTILISATION DE ROSO EN PLUSIEURS ÉTAPES SUCCESSIVES. LES SEUILS DE REJET DES ÉNERGIES LIBRES DE FORMATION DES STRUCTURES SECONDAIRES (EN KCAL.MOL^{-1}) DÉPENDENT DU TAUX DE GC DE L'ORGANISME D'ÉTUDE CONTRAIREMENT AUX MODIFICATIONS DE LA TAILLE DES SONDES (EN MERS).....193

TABLEAU B.2.1 QUANTITÉS DE CIBLES DE CONTRÔLE AJOUTÉES POUR LES TROIS TYPES D'HYBRIDATION TESTÉES : LAME QUANTIFOIL® AVEC HYBRIDATION MANUELLE (1) OU AUTOMATIQUE (3) ET LAMES ROSA® AVEC HYBRIDATION MANUELLE (2).....199

TABLEAU C.1.1 PRÉSENTATION DES QUATRE MILIEUX NUTRITIONNELS UTILISÉS POUR L'ÉLEVAGE DES PUCERONS. 204

TABLEAU C.1.2 PRÉSENTATION DES HYBRIDATIONS RÉALISÉES SUR LES HUIT LAMES DE L'ÉTUDE AVEC LE FLUOROCHROME ASSOCIÉ À CHAQUE ÉCHANTILLON.....205

TABLEAU C.2.1 PRÉSENTATION DES VECTEURS ET DES ENZYMES DE RESTRICTION UTILISÉS POUR LA PRÉPARATION DES CIBLES DE CONTRÔLE. 211

TABLEAU C.2.2 QUANTITÉS DE CIBLES DE CONTRÔLE AJOUTÉES POUR L'ENSEMBLE DES ÉCHANTILLONS.....212

TABLEAU C.3.1 RÉCAPITULATIFS DES RÉSULTATS OBTENUS SUR TRENTE LARVES ÉLEVÉES DURANT CINQ JOURS SUR LES QUATRE MILIEUX NUTRITIONNELS DE L'EXPÉRIENCE (DEUX CONCENTRATIONS EN SACCHAROSE ET DEUX TAUX D'ACIDES AMINÉS ESSENTIELS). 219

TABLEAU C.3.2 RÉPARTITION EN POURCENTAGE DES DONNÉES EN FONCTION DES INDICES QUALITÉ (-100 : MAUVAIS, -50 :NON DÉTECTÉ, 0 : CORRECT, 100 : BON) APRÈS FILTRATION AVEC LE LOGICIEL *GENEPIX* SELON LA DÉMARCHE PRÉSENTÉE DANS LA PARTIE « MATÉRIEL ET MÉTHODES ». 225

TABLEAU C.3.3 RÉPARTITION DES DONNÉES POUR LES 617 GÈNES EN FONCTION DES INDICES QUALITÉ APRÈS MOYENNAGE INTRA ET INTER-SONDES.....225

TABLEAU C.3.4 RÉPARTITION DES EFFECTIFS POUR LA RÉALISATION DU TEST DE CHI2 COMPARANT NIVEAUX D'EXPRESSION ET DISTANCES À L'ORIGINE DE RÉPLICATION DES GÈNES CHEZ *BUCHNERA*. 231

TABLEAU C.3.5 PRÉSENTATION DES CATÉGORIES FONCTIONNELLES UTILISÉES POUR LE CLASSEMENT DES GÈNES CHEZ *BUCHNERA* (D'APRÈS RILEY, 1993).....236
TABLEAU F.1.1 TABLE DES TERMES DE SYMÉTRIE ET D'INITIATION POUR LE CALCUL DES ÉNERGIES D'HYBRIDATION D'UNE SOLUTION À PH 7 CONTENANT 1 M DE NACL (D'APRÈS SANTALUCIA, 1998). 309
TABLEAU F.1.2 TABLE DES ÉNERGIES D'INTERACTION ENTRE BASES HOMOLOGUES POUR UNE SOLUTION À PH 7 CONTENANT 1 M DE NACL. L'ENTHALPIE (ΔH) EST EXPRIMÉE EN KCAL.MOL^{-1} ET L'ENTROPIE (ΔS) EN CAL.MOL^{-1} (D'APRÈS SANTALUCIA, 1998). LA PREMIÈRE BASE D'UN COUPLE DE NUCLÉOTIDES AU SEIN D'UNE SÉQUENCE EST LUE EN LIGNE ET LA SECONDE EN COLONNE.....309
TABLEAU F.1.3 TABLE DES ENTHALPIES (ΔH EN KCAL.MOL^{-1}) ET DES ENTROPIES (ΔS EN CAL.MOL^{-1}) D'HYBRIDATION POUR UNE SOLUTION À PH 7 CONTENANT 1 M DE NACL. LES DEUX DIMENSIONS DE LA TABLE UTILISÉE DANS ROSO CORRESPONDENT AUX DEUX LIGNES DE CHAQUE CASE DE LA TABLE. LA PREMIÈRE LIGNE CONTIENT LES VALEURS DES ENTHALPIES ET LA SECONDE LES VALEURS DES ENTROPIES. LA VALEUR 0 A ÉTÉ UTILISÉE POUR L'ENTHALPIE COMME POUR L'ENTROPIE LORSQU'IL N'EXISTE PAS DE DONNÉES THERMODYNAMIQUES DISPONIBLES. LE COUPLE DE NUCLÉOTIDES DU BRIN 3' EST LU EN LIGNE ET LE COUPLE DU BRIN 5' EN COLONNE. 311
TABLEAU F.1.4 ÉNERGIES LIBRES DE FORMATION DES BOUCLES DES ÉPINGLES À CHEVEUX (EN KCAL.MOL^{-1}). L'ÉNERGIE LIBRE DE FORMATION D'UNE BOUCLE DÉPEND UNIQUEMENT DE SA TAILLE (D'APRÈS FREIER *ET AL.*, 1986 ; GROEBE ET UHLENBECK, 1988).....312
TABLEAU F.3.1 COMPOSITION DU MILIEU AP3 (POUR 100 ML DE MILIEU). 316
TABLEAU F.3.2 COMPOSITION DES MILIEUX. 318
TABLEAU F.4.1 LISTE DES GÈNES EXPRIMÉS DE FAÇON DIFFÉRENTIELLE POUR LE TERME D'INTERACTION *AA : SACCHAROSE* (OBTENUE AVEC LES TESTS F_3 ET *FS,* SEUIL DE SIGNIFICATIVITÉ DE LA PROBABILITÉ ASSOCIÉE À CHAQUE GÈNE : 0,05). M25 ET M50 REPRÉSENTENT LES RAPPORTS M ESTIMÉS POUR LES CONCENTRATIONS EN SACCHAROSE (0,5 M : 1 M), RESPECTIVEMENT POUR DES TAUX D'ACIDES AMINÉS ESSENTIELS DE 25 ET 50 %. M1 ET M0,5 REPRÉSENTENT LES RAPPORTS M ESTIMÉS POUR LES TAUX D'ACIDES AMINÉS ESSENTIELS (50 % : 25 %), RESPECTIVEMENT POUR DES CONCENTRATIONS EN SACCHAROSE DE 1 ET 0,5 M.....319
TABLEAU F.4.2 LISTE DES GÈNES EXPRIMÉS DE FAÇON DIFFÉRENTIELLE POUR LE FACTEUR *AA* (OBTENUE AVEC LES TESTS F_3 ET *FS,* SEUIL DE SIGNIFICATIVITÉ DE LA PROBABILITÉ ASSOCIÉE À CHAQUE GÈNE : 0,05). 321
TABLEAU F.4.3 LISTE DES GÈNES EXPRIMÉS DE FAÇON DIFFÉRENTIELLE POUR LE FACTEUR *SACCHAROSE* (OBTENUE AVEC LES TESTS F_3 ET *FS,* SEUIL DE SIGNIFICATIVITÉ DE LA PROBABILITÉ ASSOCIÉE : 0,05). ...322

Avant-propos

« Les choses ne changent pas. Tu changes ta façon de regarder c'est tout. »
Carlos Castaneda[1]

L'obtention en juillet 1995 de la première séquence complète du génome d'un organisme vivant, *Haemophilus influenzae* (Fleischmann *et al.*, 1995) a été accueillie de manière contrastée par la communauté scientifique (Rechenmann et Gautier, 2000). Si certains chercheurs l'ont pressentie comme un événement majeur ouvrant des voies radicalement nouvelles d'investigation du vivant, d'autres l'ont perçue au mieux comme un exploit purement technique. Le séquençage des nombreux organismes modèles qui a suivi s'est accompagné de progrès technologiques fulgurants permettant l'apparition d'outils nouveaux. Parmi ceux-ci, les puces à ADN offrent la possibilité d'obtenir des estimations simultanées des niveaux d'expression de plusieurs milliers de gènes d'un organisme au cours d'une expérience unique (Gerson, 2002). Cette prouesse technologique marque l'avènement de la transcriptomique.

Aujourd'hui, alors qu'une première version du génome humain est disponible (Venter *et al.*, 2001) et qu'une base de données comme celle de l'Institut Européen de Bioinformatique[2] (EMBL-EBI) contient plus de 70 milliards de nucléotides, les divergences d'appréciation concernant le séquençage et l'engouement pour la technologie des puces ont laissé la place à de nouveaux défis. En effet, la quantité considérable de données obtenues et leur nature particulière est une révolution qui pose le problème de la qualité, de l'analyse et du stockage de ces données (Hedge *et al.*, 2000). La biologie, peu atteinte jusqu'à présent par le formalisme mathématique, l'analyse statistique et l'informatique, a donc vu entrer ces disciplines dans le quotidien de ses chercheurs. Ces approches dites *in silico* qui complètent les études *in vivo* et *in vitro* sont rassemblées

[1] Castaneda, C. *Voir : les enseignements d'un sorcier yaqui*. Gallimard, Paris (1973).
[2] http://www3.ebi.ac.uk/Services/DBStats/.

sous le terme de « bioinformatique » dont il est difficile de donner une définition précise qui fasse l'unanimité. Ce terme n'est apparu dans la littérature scientifique qu'au début des années 1990, cependant ce domaine de recherche existait bien avant l'essor de la génomique et des dizaines de laboratoires dans le monde travaillent depuis longtemps en biomathématiques ou biométrie. Les premières étapes de la bioinformatique coïncident avec celles de la biologie moléculaire (Ouzounis et Valencia, 2003). Aujourd'hui son enjeu est double avec d'une part le développement de méthodes d'acquisition, de contrôle et d'analyse des données transcriptomiques (Brazma et Vilo, 2000), et d'autre part le passage du niveau de l'analyse des données à celui de la connaissance (Rechenmann, 2000). L'essor de la génomique offre aujourd'hui à la biologie la quantité de données nécessaires à une nouvelle approche expérimentale engagée dès 1978 par des précurseurs comme Pierre Delattre et Michel Thellier, fondateurs de l'actuelle « Société Francophone de Biologie Théorique ». Il s'agit de poursuivre cet effort de modélisation afin d'établir entre les données une cohérence nécessaire à la compréhension des organismes en tant que systèmes complexes (Legay, 1996).

Bien avant le développement des puces à ADN, l'UMR INRA/INSA « Biologie Fonctionnelle, Insectes et Interactions » (BF2I) étudiait le modèle de la bactérie *Buchnera aphidicola,* une Entérobactériacée proche d'*Escherichia coli* qui vit en symbiose avec le puceron du pois *Acyrthosiphon pisum.* Ces insectes se nourrissent exclusivement de la sève phloémienne des plantes qui est un aliment très déséquilibré, riche en sucres mais pauvre en acides aminés et tout particulièrement en acides aminés essentiels. De nombreux travaux physiologiques ont donc été réalisés pour étudier leur nutrition et notamment le métabolisme de certains acides aminés et du saccharose (Febvay *et al.*, 1995) (Febvay *et al.*, 1999). Des chromatographies d'acides aminés réalisées à partir de broyats de pucerons ont ainsi permis de montrer que des pucerons privés de leurs bactéries symbiotiques et nourris sur un aliment déséquilibré conservent un profil d'acides aminés libres déséquilibré. Au contraire, les pucerons symbiotiques sont capables, sur ce même milieu, de reconstituer un profil « normal ». Cette observation laisse supposer que *Buchnera* est capable de fournir à son hôte les acides aminés qui lui manquent.

Récemment, le séquençage de cette *Buchnera* {Shigenobu, 2000 #381} a montré que son génome extrêmement réduit ne contient que 619 gènes. Contrairement à la bactérie *Escherichia coli*, le génome de *Buchnera* a perdu l'essentiel de ses gènes de régulation. Certains gènes du métabolisme général sont également absents, mais une des caractéristiques fonctionnelles de ce génome est la conservation de la plupart des voies de biosynthèse des acides aminés essentiels. Le séquençage du génome de *Buchnera* et le développement des

puces à ADN offrent à présent la possibilité d'explorer l'ensemble des réponses métaboliques et régulatrices de cet organisme.

Dans ce contexte, les objectifs de cette thèse ont été de développer un ensemble de méthodologies, nécessaires à la conception et à l'utilisation d'une puce à ADN dédiée à *Buchnera,* pour étudier la capacité de la bactérie à adapter son métabolisme aux besoins du puceron.

Une étude bibliographique de ce double contexte méthodologique et biologique fait l'objet de la première partie de cette thèse. La deuxième partie est consacrée aux aspects méthodologiques avec le développement du logiciel ROSO (Recherche et Optimisation de Sondes oligonucléotides) et la conception d'une puce dédiée à *Buchnera*. Enfin, la troisième partie a pour objet l'utilisation de cette puce pour étudier l'impact de la bactérie *Buchnera* sur les capacités physiologiques de son hôte et sa capacité à réguler l'expression de ses gènes. Pour cela, les transcriptomes de bactéries issues de pucerons élevés sur des milieux contenant des quantités variables d'acides aminés et de saccharose ont été comparés. Compte tenu de la spécificité de chacun des aspects de ce travail, le plan, volontairement très détaillé de ce manuscrit, devrait permettre au lecteur de retrouver rapidement les parties susceptibles de l'intéresser.

L'ensemble de ce travail a naturellement imposé la mise en place d'une démarche pluridisciplinaire. Cette thèse m'a donc permis d'aborder des domaines aussi divers que l'écriture de lignes de code et les joies du développement logiciel avec ROSO, l'élevage des pucerons et les interminables comptages quotidiens de larves, le plaisir des « manips à la paillasse » avec parfois des extractions et des purifications d'ARN sans ARN, des marquages non marqués, des pannes d'appareil à hybrider juste avant les hybridations ou de scanner juste avant les lectures d'images, le côté psychédélique des grilles d'acquisition de données avec leur multitude de points rouges et verts à inspecter manuellement, le charme inestimable de la manipulation et de l'analyse statistique des données sous R et bien sûr le casse-tête des voies métaboliques. Mais cette thèse m'a également offert la possibilité de travailler avec des personnes différentes. Et si elle existe aujourd'hui, c'est aussi un peu grâce à Laurent Duret pour ROSO, Bruno Spataro pour sa mise en place sur le site du Pôle Bioinformatique Lyonnais (PBIL), Angela Douglas et Claire Allen pour leur accueil à York, Gabrielle Duport pour l'élevage des pucerons, Jacques Bernillon pour la fabrication des puces, Federica Calevro pour les expérimentations, Nicolas Morin pour les analyses statistiques et Yvan Rahbé pour les interprétations biologiques. Enfin ça n'aurait pas été pareil sans Gérard Febvay qui m'a permis de me rendre à de nombreuses formations, Hubert Charles qui a réponse

à presque toutes les questions insolubles et bien sûr Jean-Michel Fayard qui y a toujours cru.

Partie A
Étude bibliographique

1

1 Contexte biologique

« Former un couple c'est n'être qu'un ; mais lequel ? »
Oscar Wilde

1.1 La symbiose

1.1.1 Généralités

Étymologiquement le mot symbiose provient du grec « *symbiôsis* » qui signifie vie (biose) avec (sym). Il est employé pour la première fois par l'historien grec Polybe (205 avant J.-C.) pour parler de la vie en communauté des compagnons. Il réapparaît ensuite dans un manuscrit latin d'Althussius (1603) où il s'intègre dans un vocabulaire à la fois philosophique et naturaliste destiné à l'analyse de la vie politique (Perru, 2003). Mais ce n'est qu'à partir de 1877, qu'il est employé d'abord en anglais et en allemand, puis en français, pour désigner les associations d'organismes biologiques. En fait, c'est Albert Frank qui utilise pour la première fois, en 1877, le terme de « *symbiotismus* », pour caractériser les lichens et définir la réalité biologique de ce phénomène :

> *« La relation en question est encore quelque chose de plus que le simple parasitisme au sens habituel, en fait le parasite et l'hôte sont réunis dès l'origine pour constituer un nouvel organisme unifié qu'aucune des deux autres parties n'entreprend de construire pour elle-même, et où les deux compagnons se partagent le travail en vue de leur nutrition ».*

Alors que Frank vient de définir la symbiose comme « vivre ensemble » (*zusammenleben*) entre l'algue et le champignon, le lichénologue Anton de Bary travaille, lui aussi, à approfondir le concept de symbiose à partir de l'exemple des lichens, mais sans restriction et dans une perspective probablement plus ouverte. Il définit en 1879 la symbiose comme étant un état de fait qui résume *« les phénomènes de la vie en commun d'organismes différents »* (De Bary, 1879) :

> *« On peut alors appliquer le terme de symbiose à toutes les relations, telles que celles qui existent entre les insectes qui entrent dans les fleurs, et les fleurs qui reçoivent le pollen des insectes, entre les ani-*

> *maux qui cherchent leur nourriture ou un abri et les autres animaux ou les plantes qui les leur procurent. Je n'ai aucune objection à faire contre cette généralisation, je me suis efforcé de montrer que tous ces phénomènes se touchent ».*

Il distingue également pour la première fois deux catégories différentes : la symbiose antagoniste, dans laquelle il y a lutte et la symbiose mutualiste dans laquelle il y a avantage réciproque pour les symbiotes.

Actuellement, cette vision de la symbiose n'a pas changé. En revanche, elle exclut généralement la symbiose antagoniste ou parasitisme comme le montre la définition de l'*Encyclopedia Universalis* (2001). Elle rappelle que si tous les intermédiaires existent entre symbiose au sens strict et parasitisme, l'appellation « symbiose » est généralement réservée au cas d'associations coopératives dans lesquelles les relations entre les deux partenaires tendent, pour l'un comme pour l'autre, à un équilibre entre les profits et les pertes, ou sont favorables à l'un des partenaires sans nuire sensiblement à l'autre.

Les deux classes de symbiose initialement définies par de Bary, ont laissé la place à cinq niveaux d'association qui correspondent aussi probablement en partie à des degrés différents d'évolution ou de coévolution (Nardon et Grenier, 1993) :

- La symbiose antagoniste ou parasitisme dans laquelle il existe une agressivité réciproque des deux partenaires qui mettent en place des mécanismes de défense, l'hôte est globalement affecté.
- La symbiose primaire ou symbiose facultative qui affecte peu l'hôte et pour laquelle les symbiotes n'ont pas de localisation précise (infection à virus ou à bactéries chez certaines Drosophiles).
- La symbiose secondaire ou symbiose chronique dans laquelle l'interaction est plus forte. Elle concerne l'ensemble de la population (contrairement aux précédentes), se caractérise par une amélioration des performances de l'hôte et une localisation précise des symbiotes (symbioses racinaires des légumineuses).
- L'endocytobiose intégrée dans laquelle toute l'espèce est symbiotique. Le symbiote intracellulaire est hébergé dans des structures spécialisées qui sont appelées mycétocytes quand le symbiote est un champignon ou une levure et bactériocytes lorsqu'il s'agit d'une bactérie. Le symbiote est devenu obligatoire, intégré et améliore nettement les performances de l'hôte. Les symbiotes ont tendance à se comporter comme des organites cellulaires. Il est possible que des transferts de gènes du symbiote vers le noyau aient déjà eu lieu (puceron).

- La symbiose organogénétique avec un symbiote qui est devenu un organite cytoplasmique. À ce niveau, il existe un transfert important de gènes vers le noyau de la cellule hôte. Actuellement, la genèse de la cellule eucaryote est expliquée par de nombreux biologistes comme des associations successives de procaryotes avec une cellule hôte. Ces procaryotes ont ensuite évolué pour donner naissance aux plastes et aux mitochondries, ce qui est maintenant admis.

Cette gradation en cinq niveaux suggère un continuum de relations depuis le parasitisme jusqu'à la symbiose organogénétique. Néanmoins, des travaux plus récents d'analyse des génomes complets de différentes bactéries parasites ou intracellulaires obligatoires proposent des scénarios différents (Moran et Wernegreen, 2000). En effet, lorsqu'une bactérie devient le parasite exclusif d'un hôte, elle perd la plupart de ses gènes non essentiels comme les gènes de biosynthèse de nombreux nutriments. Incapable de synthétiser des molécules pour son hôte, elle peut alors très difficilement devenir un symbiote. Quoi qu'il en soit l'ensemble de ces hypothèses reste encore à l'heure actuelle du domaine de la spéculation. En revanche si la symbiose a longtemps été considérée comme une curiosité biologique, il est aujourd'hui admis que tous les êtres vivants sont symbiotiques. Ainsi, bien que tous les niveaux de symbiose soient représentés de façon variée dans le monde vivant, l'endocytobiose a certainement joué un rôle fondamental dans l'évolution (Lang *et al.*, 2000).

1.1.2 Une origine endocytobiotique pour la cellule eucaryote

L'endocytobiose concerne toutes les formes de vie à deux dans lesquelles l'un des deux partenaires vit dans le corps de l'autre (Boullard, 1990). L'endocytobiose est plus qu'une relation étroite entre deux organismes, puisqu'il s'agit d'une symbiose intracellulaire au sens strict qui a pris très tôt une grande importance d'un point de vue évolutif. En effet, dès 1883, Schimper imagine une origine endocytobiotique pour les plastes des cellules eucaryotes :

> *« S'il est définitivement établi que les plastes ne sont pas formés de novo dans la cellule œuf, alors leur situation dans la cellule où ils se trouvent rappelle celle de symbiontes. Peut-être une plante verte n'est-elle que l'union entre un organisme incolore et un microbe possédant les pigments chlorophylliens. »*

À l'époque, cette hypothèse se heurte à l'incompréhension de la majorité de ses pairs et elle est rapidement abandonnée. Elle n'est remise au goût du jour qu'en 1967, sous l'impulsion de Margulis. Elle utilise comme argument majeur la mise en évidence en 1962 d'un génome dans les plastes (Ris et Plaut, 1962),

puis dans les mitochondries l'année suivante. Cet ADN ressemble à celui des procaryotes car il n'est pas isolé au sein de l'organite et se présente sous la forme de plusieurs copies identiques d'une molécule circulaire. Il existe également des ribosomes dans les mitochondries et les plastes, qui participent à la synthèse de protéines et ont la taille et la composition en ARN des ribosomes de procaryotes. Dans les années 1990, la biologie moléculaire a permis de souligner la parenté de ces organites avec les procaryotes (Reith et Munholland, 1993).

Ainsi, l'acquisition des mitochondries se serait produite par endocytobiose il y a 2 à 3 milliards d'années, constituant une étape majeure pour les eucaryotes. Elle marque l'acquisition de la respiration cellulaire par la cellule eucaryote qui devient capable d'utiliser l'oxygène pour oxyder les sucres et produire son énergie. Compte tenu de la rareté actuelle des eucaryotes dépourvus de ces organites, cette endocytobiose semble avoir été déterminante dans le succès évolutif des eucaryotes. Les plastes auraient été acquis plus tardivement, probablement entre 1,2 et 2 milliards d'années, par certains eucaryotes qui forment les différentes lignées actuelles de végétaux.

Une fois que l'association est devenue permanente, un contrôle réciproque de la multiplication de chacun des partenaires s'instaure pour la survie de l'organisme chimère formé. La voie est alors ouverte à une évolution conjointe (ou coévolution) des partenaires, et des échanges réciproques peuvent s'établir. Pour l'endocytobiote, la vie à l'intérieur d'une cellule est extrêmement confortable car l'hôte supporte les agressions du milieu extérieur. Cet habitat protégé explique la disparition des enveloppes protectrices. Mais l'évolution régressive la plus curieuse est certainement la réduction de taille du génome des organites par rapport aux formes libres dont ils sont issus. Ainsi, certains gènes nécessaires pour se repérer, se déplacer, se protéger dans le milieu extérieur ont pu disparaître, alors que d'autres en revanche existent toujours mais ont été transférés dans le noyau de la cellule hôte. Ces transferts scellent l'association entre les partenaires, et expliquent l'impossibilité de cultiver isolément les organites qui sont devenus complètement dépendants de leur hôte. La symbiose est donc devenue obligatoire et les partenaires forment un nouvel organisme chimérique. Des régulations couplent l'expression des deux génomes, qui concourent ensemble au bon fonctionnement de la nouvelle entité physiologique. L'endocytobiose a donc profondément influé sur l'évolution des eucaryotes en leur permettant d'acquérir des potentialités métaboliques de procaryotes. Plus qu'une curiosité biologique, la symbiose est certainement l'un des moteurs les plus puissants de l'évolution du monde vivant. Elle crée très rapidement des organismes chimériques qui peuvent engendrer

des lignées nouvelles. Elle rapproche des partenaires et favorise des transferts de gènes massifs qui créent des génomes eux aussi chimériques car le génome nucléaire contient des gènes eucaryotes, mais aussi des gènes d'origine bactérienne, issus des mitochondries ou des plastes.

1.1.3 Endocytobiose et unité de sélection

D'un point de vue évolutif, la sélection peut intervenir à trois niveaux. Tout d'abord, au niveau de l'entité chimérique elle-même appelée aussi symbiocosme (Nardon, 1995) qui s'adapte au milieu environnant. Puis au niveau de l'hôte pour qui la symbiose peut être un facteur d'adaptation permettant une diversification des formes de vie par l'acquisition par exemple de nouvelles capacités métaboliques. Enfin, au niveau du symbiote, qui dans les cas extrêmes peut devenir un organite cytoplasmique entièrement contrôlé par l'hôte (cf. 1.1.2).

1.1.31 Impact sur l'hôte

Le rôle de la symbiose intracellulaire dans le pouvoir adaptatif des hôtes à des environnements très différents a longtemps été ignoré ou sous-estimé. Pourtant, l'endocytobiose a permis à des hôtes divers d'exploiter et de coloniser des milieux pauvres ou inhospitaliers. L'adaptation des pucerons au phloème, un milieu nutritionnel particulièrement déséquilibré semble ainsi être liée à l'acquisition des symbiotes. Par ailleurs, les nouvelles capacités métaboliques et écologiques héritées des symbiotes, isolent écologiquement les groupes qui les possèdent de ceux qui en sont dépourvus. En ce sens, la symbiose intracellulaire favoriserait la divergence des populations et pourrait représenter une barrière écologique favorable à la spéciation sympatrique (ou spéciation non géographique). De même, des symbiotes comme les *Wolbachia* peuvent intervenir directement sur la reproduction de leur hôte, en induisant par exemple des incompatibilités qui isolent la population contenant les symbiotes (Moran et Baumann, 1994).

1.1.32 Impact sur le symbiote

La vie intracellulaire et la transmission maternelle des symbiotes favorisent une évolution originale des génomes. Tant pour les endocytobiotes que pour les parasites intracellulaires, il a été observé un manque de recombinaisons génétiques, une fixation importante de mutations délétères acquises par dérive génétique, un taux de mutations important et un biais mutationnel élevé (Wernegreen, 2002).

Le manque de recombinaisons génétiques dans les populations d'endocytobiotes résulte d'une part de la séquestration des symbiotes dans les

cellules hôtes et d'autre part de la perte possible de certains éléments (comme les séquences répétées ou les transposons) qui sont impliqués indirectement dans la recombinaison. Par ailleurs, le fait que seul un petit nombre de symbiotes soit transmis à la descendance de l'hôte à chaque génération a pour conséquence une diminution drastique de l'effectif efficace de la population. Ce phénomène appelé « *Muller's ratchet* » ou « cliquet de Muller » (processus nommé ainsi après avoir été décrit pour la première fois par Muller en 1964) ou « goulots d'étranglement » se traduit en effet par une accélération de la dérive génétique qui favorise la fixation de mutations délétères (Moran, 1996).

Ce manque de recombinaisons et cette accélération de la dérive génétique aboutissent à une augmentation de la vitesse d'évolution chez les symbiotes intracellulaires. Parallèlement, les génomes symbiotiques présentent une accumulation très importante de bases A et T, aussi bien dans les régions codantes que non codantes. Ce phénomène, qui a également été observé dans les mitochondries (Crozier et Crozier, 1993), semble être une caractéristique des génomes des organismes vivant dans un milieu intracellulaire. Son origine et les pressions sélectives qui le maintiennent ne sont pas encore complètement élucidées bien qu'une nette corrélation entre taux de GC et taille des génomes bactériens ait été observée. Certains auteurs l'expliquent par une déficience ou une absence des enzymes capables de corriger les erreurs de réplication (Wernegreen, 2002). D'autres se basent sur une « faiblesse » de la polymérase dans les régions de l'ADN déjà particulièrement riches en bases A qui aurait pour conséquence l'ajout de bases A et T par l'enzyme dans ces régions (Tamas, 2002).

Enfin la taille des génomes d'endocytobiotes est souvent très petite en comparaison des groupes bactériens dont ils sont issus (Andersson *et al.*, 2002 ; Andersson et Kurland, 1998). Deux explications ont été proposées à cette particularité. D'une part, des mécanismes de délétions massifs et de réarrangements chromosomiques permettraient de limiter les effets de l'accumulation de mutations délétères (Andersson et Andersson, 1999 ; Ochman et Moran, 2001) et d'autre part, la vie dans un milieu intracellulaire assurant une fourniture de nutriments et éventuellement d'enzymes expliquerait la perte de gènes devenus inutiles aux endocytobiotes (Moran, 1996).

1.1.33 Impact sur le symbiocosme

La dépendance mutuelle des deux partenaires montre que l'association entre l'hôte et ses symbiotes est une véritable entité coadaptée. En effet, toutes les tentatives de cultures *in vitro* de bactéries endocytobiotiques ont jusqu'à présent échoué, confirmant la perte d'autonomie des symbiotes en l'absence de l'hôte (Hinde, 1971b). Parallèlement, dans le cas de symbioses anciennes, l'hôte ne survit pas ou est généralement très diminué en l'absence de ses sym-

biotes. Cet état de coadaptation est souvent le résultat d'une longue coévolution entre les deux génomes des partenaires.

Mais plus encore que cet aspect de coévolution, il est possible de voir le symbiote comme un ensemble de gènes prêts à s'exprimer, acquis simultanément et transmis à toute la descendance par voie maternelle. Ce processus d'acquisition génique est donc bien plus efficace que la mutagenèse ou les remaniements chromosomiques. En ce sens, la symbiose apparaît comme un mécanisme sophistiqué de prédation génique et comme la source d'innovation la plus efficace. Elle ne peut pas s'interpréter bien sûr en termes de gradualisme dans la mesure où le symbiocosme est en fait un groupe de génomes qui ne dérivent pas les uns des autres, mais se sont assemblés selon un processus particulier (Nardon, 1995).

La définition de la symbiose, la théorie de l'origine endocytobiotique des cellules eucaryotes et tout particulièrement les aspects évolutifs de l'endocytobiose mettent en avant la difficulté à définir un concept *a priori* évident, celui de l'individu.

1.1.4 Le concept d'individu

La symbiose, loin d'être une curiosité biologique, est un phénomène révélateur d'une tendance à l'association qui existe de façon universelle dans le monde vivant et qui remet en question le concept d'individualité biologique (Perru, 1997). Ainsi, ce que nous appelons « un homme » contient 10^{13} cellules eucaryotes, 10^{13} bactéries sur la peau et les muqueuses et 10^{14} bactéries dans l'appareil digestif. De même, dans les cas de symbioses plus évoluées qui associent des microorganismes avec des métazoaires, le nombre de symbiotes est du même ordre de grandeur que le nombre de cellules de l'hôte (Nardon et Grenier, 1993). Alors, qu'est-ce qu'un individu ? D'un point de vue étymologique, le terme « individu » provient du latin « *individuum* » qui signifie « corps indivisible ». D'un point de vue biologique, un individu est donc un corps organisé dont l'ensemble de ses constituants possède une certaine homogénéité génétique. Ce corps vit une existence propre et ne peut être divisé sans être détruit. Mais en appliquant strictement cette définition, comme le souligne Margulis (1993), seules les bactéries méritent le terme d'individu, les autres formes étant des assemblages plus ou moins sophistiqués d'organismes différents.

L'autonomie d'un individu est donc limitée, que se soit au niveau de l'organisme avec la digestion, ou au niveau de l'écosystème avec la nutrition ou la reproduction. Poussant ce raisonnement à l'extrême, Giglio-Tos définit en 1903 le monde vivant comme une immense symbiose où chaque espèce ne peut

vivre qu'aux dépens des substances nourrissantes fournies par les autres. Le concept de symbiose signifie alors une interdépendance au plan de la nutrition, chaque espèce assimilant des substances provenant d'autres espèces auxquelles elle fournit à son tour de la nourriture. Pour lui, l'individu n'existe pas car il est toujours possible de le considérer comme symbiose d'individualités élémentaires.

Cette difficulté à définir l'individu montre la relativité qui existe en biologie où le réel peut être vu comme un système continu dans lequel il est nécessaire de définir les limites des systèmes étudiés en fonction des questions posées (Kupiec et Sonigo, 2000). En ce sens, l'individu constitue plus une approximation de travail « utile » qu'une réalité, et pour l'étude d'un système dynamique ce sont les interactions qui deviennent les éléments déterminants. En ce qui concerne l'étude de la symbiose, l'association des deux partenaires crée une nouvelle entité biologique, le « symbiocosme », qui peut s'interpréter comme un micro écosystème, avec des interactions entre les partenaires et avec le milieu. Il s'agit finalement d'une communauté interdépendante, avec perte d'autonomie du symbiote, et évolution vers une nouvelle forme d'individualité (Nardon et Grenier, 1993).

Cette présentation de la symbiose montre qu'aujourd'hui la question n'est plus vraiment, contrairement aux querelles initiales, de savoir si la symbiose est plus ou moins mutualiste ou parasitaire. Il s'agit plutôt de décrire ce que représente le complexe vivant qui émerge de ces relations, et qui peut être étudié comme un système dans lequel l'association des deux organismes donne naissance à des propriétés émergentes qui ne sont pas celles des deux entités séparées.

1.2 Les deux partenaires de l'association puceron - bactérie

La symbiose puceron - bactérie est un modèle étudié depuis de nombreuses années au laboratoire BF2I en raison des caractéristiques inhabituelles de sa physiologie nutritionnelle et de son impact agronomique. En effet, le puceron se nourrit exclusivement du phloème des plantes qui est un liquide de haute pression osmotique excessivement déséquilibré. Pour survivre sur ce milieu pauvre et fortement osmotique, le puceron a développé d'une part de remarquables adaptations anatomiques, physiologiques et biochimiques et d'autre part une intime symbiose avec des micro-organismes, et principalement avec une bactérie du genre *Buchnera,* qui l'approvisionne en nutriments difficiles à obtenir directement du phloème.

Cette association est indispensable pour le puceron comme pour *Buchnera*. En effet, il est possible d'éliminer les bactéries des pucerons par un traitement antibiotique, mais ces pucerons dépourvus artificiellement de bactéries (appelés pucerons aposymbiotiques) présentent une croissance fortement ralentie et leur descendance n'est pas viable (Douglas, 1992 ; Mittler, 1971). De même la bactérie *Buchnera* peut être isolée par dissection mais ne survit que quelques heures en dehors du puceron (Whitehead et Douglas, 1993b).

De nombreuses études attribuent le début de cette association à l'infection d'un ancêtre commun à l'ensemble des espèces de pucerons par une bactérie, il y a environ 180 à 250 millions d'années (Moran *et al.*, 1993). Cet événement a été suivi d'une coévolution stricte des deux partenaires sans transmission horizontale d'un autre symbiote (Clark *et al.*, 2000).

La compréhension des mécanismes d'adaptation de la symbiose puceron - bactérie au phloème est un premier pas pour appréhender le fonctionnement global de ce système. Au-delà de cet intérêt académique, cette compréhension permettra éventuellement de lutter plus efficacement contre le puceron du pois, véritable ravageur des cultures légumières, tant par ses dommages directs que par la transmission de virus aux plantes dont il est responsable.

1.2.1 Du côté du puceron *Acyrthosiphon pisum*

1.2.11 Caractéristiques biologiques

Le puceron du pois *Acyrthosiphon pisum* (cf. **Figure A.1.1**) appartient à la famille des Aphididae et à l'ordre des Hémiptères. Les adultes aptères ou ailés ont un cycle de reproduction annuel hétérogonique impliquant une alternance de plusieurs générations parthénogénétiques (les femelles produisent leur descendance sans accouplement) et une seule génération sexuée. Les femelles parthénogéniques sont toujours vivipares, c'est-à-dire qu'elles donnent naissance à de jeunes larves capables de s'alimenter et de se déplacer immédiatement. Ce cycle complexe lui permet d'exploiter de façon optimale les ressources de sa plante hôte (Baumann et Baumann, 1994).

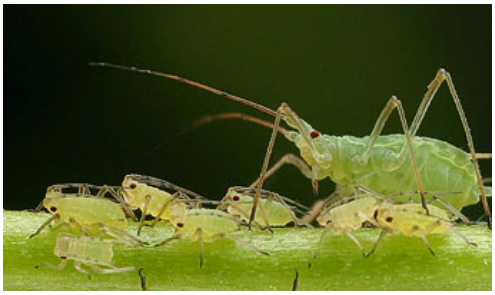


Figure A.1.1 Larves de pucerons sur leur plante hôte (à gauche) et puceron adulte avec ses larves en vue rapprochée (à droite) (Rahbé, Y., communication personnelle).

1.2.12 Nutrition

1.2.121 Composition du phloème

Les pucerons se nourrissent exclusivement de la sève phloémienne des plantes qui est composée essentiellement de sucres, et tout particulièrement de saccharose qui associe une grande stabilité chimique et une faible viscosité en solution concentrée. Les principaux composants azotés sont des acides aminés libres dont la concentration est comprise entre 50 et 800 mM. Les acides aminés les plus représentés sont généralement l'aspartate, le glutamate, l'asparagine et la glutamine. Les acides aminés essentiels, c'est-à-dire les acides aminés que les animaux ne peuvent synthétiser *de novo,* sont pour les pucerons aposymbiotiques, l'histidine, l'isoleucine, la leucine, la lysine, la méthionine, la phénylalanine, la thréonine, le tryptophane et la valine (Mittler, 1971). Ils sont présents en faible concentration (Sandstrom et Moran, 2001 ; Sandstrom *et al.*, 2000) et représentent seulement 10 à 30 % de l'ensemble des acides aminés (Abisgold *et al.*, 1994). Enfin, le phloème contient des substances inorganiques. Il s'agit principalement de potassium et de phosphate à une concentration de l'ordre de 1 à 5 $g.l^{-1}$ et de micro-nutriments (fer, manganèse, zinc, cuivre et molybdène). En revanche, les lipides tout comme les stérols ne sont présents qu'en infimes quantités dans la sève.

1.2.122 Prise alimentaire

Le saccharose et les acides aminés, en tant que principales sources respectivement de carbone et d'azote dans la sève phloémienne, ont fait l'objet de nombreuses études nutritionnelles chez le puceron. Il a ainsi été observé que la prise alimentaire des pucerons varie en fonction de la teneur nutritive de ces deux sources alimentaires. Des études ont montré que le saccharose à une concentration supérieure à 0,1 M et la méthionine peuvent être considérés comme des phagostimulants. De plus, les pucerons adoptent des réponses nutritionnelles compensatoires en augmentant leur prise alimentaire en présence

d'un aliment contenant de faibles concentrations en saccharose (Abisgold *et al.*, 1994). De la même façon, les pucerons montrent une réponse alimentaire compensatoire pour des concentrations en acides aminés comprises entre 75 et 250 mM, mais pas pour des concentrations plus élevées (Abisgold *et al.*, 1994). En revanche, ce mécanisme est altéré chez les pucerons aposymbiotiques (Prosser *et al.*, 1992).

1.2.123 Utilisation du saccharose

Le saccharose présent en très forte concentration dans la sève phloémienne est responsable d'une pression osmotique deux à quatre fois plus élevée que celle des fluides de l'insecte (Douglas *et al.*, 2001). Le puceron a donc développé des stratégies lui permettant de maintenir l'équilibre hydrique et d'acquérir les autres nutriments de la sève en quantité suffisante (Rhodes *et al.*, 1997 ; Wilkinson *et al.*, 1997). Il est capable de diminuer cette pression par la production d'oligosaccharides à partir des sucres du milieu et cela, indépendamment de la présence de *Buchnera* (Wilkinson *et al.*, 1997).

Les pucerons utilisent essentiellement le saccharose comme source d'énergie (Febvay *et al.*, 1995 ; Rhodes *et al.*, 1996). Cependant, bien que le saccharose soit composé d'un glucose et d'un fructose, les pucerons utilisent préférentiellement le fructose, et pour de fortes concentrations en saccharose, éliminent le glucose sous forme d'oligosaccharides retrouvés dans le miellat (Ashford *et al.*, 2000).

1.2.124 Utilisation des acides aminés

Si le puceron utilise le saccharose comme source d'énergie, les acides aminés sont par contre réservés à la production de composés structuraux (Rhodes *et al.*, 1996). Les acides aminés aromatiques comme la tyrosine et son précurseur, la phénylalanine, sont ainsi des acides aminés indispensables au développement larvaire des insectes (Dadd, 1973). Ils interviennent tout particulièrement dans la formation de l'exosquelette, fondamental pour le développement des insectes qui synthétisent une nouvelle cuticule après chaque mue larvaire.

Les pucerons sont capables de survivre sur différentes plantes hôtes dont les phloèmes contiennent des proportions extrêmement variables en acides aminés (Bernays et Klein, 2002 ; Sandstrom et Pettersson, 1994). Il est à noter que même sur une plante hôte unique, la composition du phloème varie parfois dans des proportions importantes suivant la localisation de la plante, la saison, en réponse à la prise alimentaire des pucerons (Telang *et al.*, 1999) ou encore sous l'effet de champignons pathogènes (Johnson *et al.*, 2003). Des études ont donc été menées pour étudier l'impact de la concentration en acides aminés essentiels dans le milieu nutritionnel des pucerons. Il a ainsi été montré qu'une

variation de la concentration en acides aminés essentiels comprise entre 20 % (taux de la sève phloémienne) et 50 % (taux des milieux artificiels) de la totalité des acides aminés du milieu n'a pas de répercussion sur leur croissance (Prosser et Douglas, 1992). En revanche pour des concentrations plus faibles, le poids des pucerons diminue.

1.2.13 Localisation des symbiotes

Parmi les micro-organismes présents chez le puceron, *Buchnera,* en tant que symbiote primaire ou principal, peut représenter plus de 90 % des bactéries. Par opposition, les bactéries restantes sont appelées symbiotes secondaires ou parfois bactéries accessoires. Ces symbiotes secondaires ne sont pas présents de façon universelle dans les populations naturelles de pucerons. Chez les pucerons qui en possèdent, elles peuvent se trouver dans l'hémolymphe des insectes comme dans certaines cellules de l'appareil digestif (Darby *et al.*, 2001 ; Fukatsu *et al.*, 2000 ; Sandstrom *et al.*, 2001).

Les bactéries *Buchnera* sont présentes à une densité d'environ 10^7 cellules.mg^{-1} de pucerons frais, ce qui correspond à approximativement 10 % du volume total de l'insecte (Baumann *et al.*, 1996 ; Humphreys et Douglas, 1997 ; Wilkinson *et al.*, 2001). La population de bactéries augmente durant les stades larvaires, pour atteindre son maximum durant la période de reproduction des jeunes adultes et finalement décliner chez les pucerons âgés (Koga *et al.*, 2003). La dynamique de la population reflète donc l'activité biologique des pucerons, suggérant un éventuel contrôle de leur prolifération par le puceron (Moran *et al.*, 1993).

Contrairement aux bactéries accessoires, *Buchnera* est localisée dans des structures particulières (cf. **Figure A.1.2**). Il s'agit de cellules polyploïdes appelées bactériocytes (Griffiths et Beck, 1973) dont la seule fonction est apparemment d'assurer la survie des bactéries qui occupent 60 à 70 % de leur cytoplasme (Whitehead et Douglas, 1993a). Un puceron adulte possède entre 60 et 80 bactériocytes, contenant au total plus de 5 millions de bactéries (Baumann et Baumann, 1994 ; Douglas et Dixon, 1987). Ces bactériocytes sont regroupés dans une structure bilobée appelée bactériome qui est située dans l'hémolymphe de l'insecte à proximité des ovarioles (Baumann *et al.*, 1995). Cette structure continue et dorsale par rapport à l'appareil digestif forme un V dont la base est dirigée vers la partie postérieure de l'insecte.

À l'intérieur du cytoplasme des bactériocytes, les bactéries sont entourées individuellement ou par petits groupes, d'une membrane d'origine eucaryote appelée membrane symbiosomale. Bien que les propriétés de cette mem-

brane n'ont pas encore été élucidés, elle joue probablement un rôle clé dans l'association symbiotique en contrôlant les entrées et les sorties de nutriments.

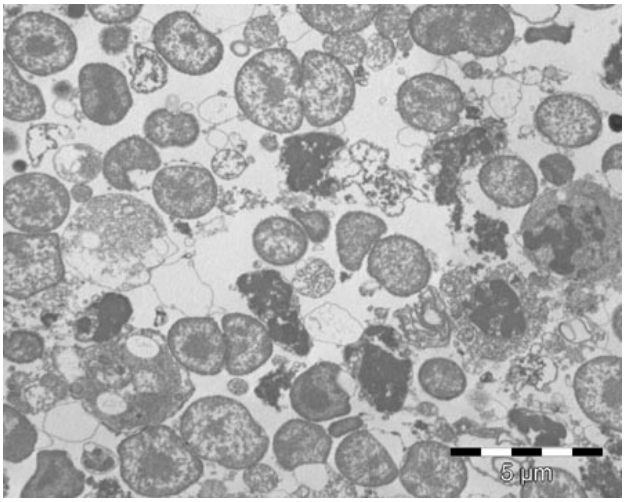


Figure A.1.2 Des bactériocytes de pucerons du pois adultes (observation par microscopie électronique à transmission) (Rahbé, Y., communication personnelle).

Les bactériocytes se différencient très tôt dans l'embryon. En conséquence, les pucerons parthénogénétiques qui portent les embryons dans leurs ovaires possèdent à la fois des bactériocytes dans l'hémolymphe maternelle et des bactériocytes dans les embryons. De la naissance au milieu de la période reproductive des pucerons, le nombre de bactéries d'origine embryonnaire est même plus important que celui d'origine maternelle et peut représenter jusqu'à 75 % de la population totale (Humphreys et Douglas, 1997 ; Whitehead et Douglas, 1993a).

Les bactéries sont transmises par voie verticale aux embryons, via les ovaires maternels (Baumann et Baumann, 1994). Chez les pucerons ovipares, un mécanisme d'exocytose assure le transfert des bactéries à l'œuf non fécondé tandis que chez les pucerons vivipares les bactéries sont transmises aux embryons au cours du stade blastoderme (Brough et Dixon, 1990 ; Hinde, 1971a). Il existe deux populations distinctes de bactériocytes chez les embryons. La première population est présente avant la transmission des bactéries maternelles aux embryons et la seconde, qui apparaît plus tard au cours du développement, rejoint ensuite les premières cellules. La mise en place de ces deux populations cellulaires ne dépend pas de *Buchnera* car même en l'absence de symbiotes, la formation des bactériocytes est initiée (Braendle *et al.*, 2003).

1.2.2 Du côté de la bactérie *Buchnera aphidicola*

Buchnera aphidicola est une bactérie gram-négative sphérique ou légèrement ovale de 2 à 5 µm de diamètre (Houk et Griffiths, 1980) (cf. **Figure A.1.3**). Elle appartient au sous-groupe γ-3 des protéobactéries (Munson *et al.*, 1991) et fait partie, tout comme *Escherichia coli*, des Entérobactériacées.

La bactérie *Buchnera* n'est pas cultivable en dehors de son hôte, ce qui rend son étude directe difficile. En revanche, son génome est complètement séquencé pour trois pucerons : *Acyrthosiphon pisum* (Ap) {Shigenobu, 2000 #381}, *Schizaphis graminum* (Sg) (Tamas *et al.*, 2002) et *Baizongia pistaciae* (Bp) (Van Ham *et al.*, 2003). Ainsi, bien qu'il soit difficile de déterminer son taux de croissance, l'étude de son génome montre que seule une copie de l'opéron des ARN ribosomiaux est présente, ce qui est spécifique d'une bactérie à croissance lente (Baumann *et al.*, 1995). Cet opéron est divisé en deux unités de transcription *rrs* et *rrl-rrf* (Unterman *et al.*, 1989). Il semble de plus que la division des bactéries soit dépendante du stade de développement de son hôte. Elle est par exemple réduite de moitié au cours du quatrième et dernier stade larvaire du puceron (Whitehead et Douglas, 1993a). De la même façon, l'étude de ses capacités métaboliques reste délicate.

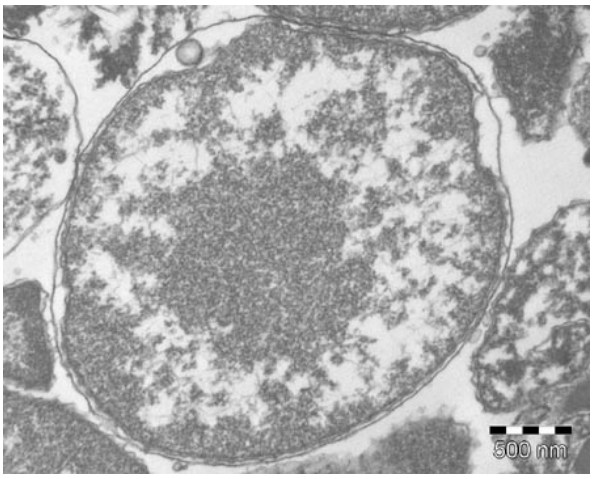


Figure A.1.3 Une bactérie *Buchnera* entourée de ses deux membranes (comme toutes les bactéries gram négative) et d'une troisième membrane d'origine eucaryote : la membrane symbiosomale (observation en microscopie électronique à transmission) (Rahbé, Y., communication personnelle).

1.2.21 Un génome avec une organisation originale

Compte tenu de la disposition et de l'annotation des gènes chez *Buchnera*, son génome est souvent considéré comme un sous-ensemble de gènes d'*Escherichia coli* (Charles et Ishikawa, 1999 ; Moran et Wernegreen, 2000).

La majorité des gènes de *Buchnera* possèdent en effet des gènes orthologues déjà séquencés et disponibles dans les bases de données. Seuls quatre gènes (*yba1*, *yba2*, *yba3* et *yba4*) ont été annotés comme orphelins. Deux explications ont été avancées concernant leur origine. La première est que ces gènes ont été perdus récemment chez *Escherichia coli* et la seconde que leur dégradation chez *Buchnera* ne permet plus leur identification par homologie (Tamas, 2002). Quoi qu'il en soit, les connaissances disponibles pour *Escherichia coli* rendent cette ressemblance globale particulièrement intéressante pour l'étude de *Buchnera*, même si son génome montre une organisation originale à tous les niveaux.

1.2.211 Au niveau global

Une taille réduite

Le génome de *Buchnera* Ap est composé d'un chromosome circulaire de 680 681 paires de bases et deux plasmides portant un total de 619 gènes dont 9 pseudogènes {Shigenobu, 2000 #381}. Cette taille est extrêmement réduite en comparaison des 4 à 5 Mpb des génomes des Entérobactéries libres (Charles et Ishikawa, 1999).

Il semble que cette réduction de taille ait eu lieu avant la divergence des lignées de *Buchnera* appartenant aux différents hôtes. Ainsi, la taille du génome de la bactérie libre correspondant au dernier ancêtre commun à *Buchnera* et *Escherichia coli* est estimée entre 1818 (Silva *et al.*, 2001) et 2425 gènes (Moran et Mira, 2001). Quant à la taille du génome du dernier ancêtre symbiotique commun aux espèces actuellement séquencées de *Buchnera,* elle est estimée approximativement à 640 gènes (Silva *et al.*, 2003). La redondance de gènes entre l'hôte et le symbiote et la stabilité du milieu intracellulaire ont vraisemblablement diminué de façon drastique la pression de sélection sur un grand nombre de gènes avec pour conséquence des délétions rapides et massives dans le génome de *Buchnera*. Cette hypothèse semble beaucoup plus probable qu'une sélection directe visant à une compaction du génome (Mira *et al.*, 2001), d'autant plus que le chromosome de *Buchnera* est soumis à une forte polyploïdie (cf. paragraphe suivant). Par ailleurs, le faible nombre de symbiotes transmis au cours des générations de pucerons a certainement favorisé la fixation de mutations délétères et donc l'accélération de la dérive génétique (Moran, 1996) (cf. 1.1.32, phénomène dit de « *Muller's rachet* »). Naturellement, le faible nombre de gènes restants a alors été soumis à une pression de sélection beaucoup plus forte et a donc peu évolué depuis le dernier ancêtre commun. Cette période de stabilité génomique, appelée stase génomique dure au moins depuis les 50 à 70 millions d'années séparant la divergence des puce-

rons *Acyrthosiphon pisum* et *Schizaphis graminum* (Tamas *et al.*, 2002). Elle est confirmée par la présence d'un très faible nombre de pseudogènes (neuf) dans le génome de *Buchnera* Ap, qui ont pour la plupart été supprimés dès le début de la réduction massive du nombre de gènes.

Par ailleurs, une étude portant sur la taille du génome de *Buchnera* issues de pucerons plus éloignés phylogénétiquement a montré que les pucerons du genre *Cinara* (Lachninae) possèdent des bactéries avec un génome contenant seulement 450 kbp (Gil *et al.*, 2002). La réduction du génome s'est donc poursuivie pour certaines lignées. Cette taille est même inférieure aux 580 kpb codant les 480 gènes de *Mycoplasma genitalium*, le plus petit génome libre et autonome connu actuellement (Fraser *et al.*, 1995). Cette étude suggère que le génome de *Buchnera* est toujours en cours de réduction contrairement à l'hypothèse précédente de stase génomique (Silva *et al.*, 2001).

Une forte polyploïdie

Chaque *Buchnera* contient entre 60 et 80 copies du chromosome (Komaki et Ishikawa, 1999). Komaki et Ishikawa (2000) ont montré de plus que cette forte polyploïdie varie avec l'âge et le stade de développement du puceron. Ainsi, le nombre de copies varie de 50 chez les embryons à plus de 100 chez les adultes avant de redescendre à 80 chez les pucerons âgés. Ces auteurs expliquent cette polyploïdie comme un moyen de conserver pour chaque gène des copies fonctionnelles non touchées par les nombreuses mutations délétères. Cependant, s'il est vrai que certaines études en évolution moléculaire suggèrent effectivement que la transmission verticale des symbiotes pose le problème de l'accumulation de mutations délétères, d'autres semblent suggérer le contraire. Ainsi, la comparaison entre les génomes de *Buchnera* issues des pucerons *Uroleucon ambrosia* et *Pemphigus obesinymphae* montre une similarité remarquable. La majorité des polymorphismes concernent de rares allèles et semblent résulter d'un polymorphisme ancestral, probablement lié à des fluctuations démographiques des populations de pucerons (Abbot et Moran, 2002). Le faible nombre de copies chez les embryons pourrait résulter d'une élimination des copies mutées au moment de la transmission des bactéries à la génération suivante, cependant le mécanisme de transmission n'a pas encore été décrit. Quant à la diminution du nombre de copies chez les insectes âgés, elle serait simplement liée à une dégradation de l'ADN génomique ou à une perte de la capacité de réplication de l'ADN (Komaki et Ishikawa, 2000). Il est également possible d'imaginer que la polyploïdie joue un rôle dans les interactions entre le puceron et les bactéries et notamment dans la régulation de l'expression des gènes. Néanmoins, à l'heure actuelle aucune étude n'a permis de démontrer une telle possibilité.

La présence de deux plasmides

Chez *Acyrthosiphon pisum,* comme chez la plupart des autres pucerons, les bactéries *Buchnera* possèdent deux plasmides, présents en de nombreux exemplaires. Le plasmide Leucine (pLeu) est constitué de 786 pb qui codent les 7 gènes de l'opéron de biosynthèse de la leucine (Silva *et al.*, 1998), tandis que le plasmide Tryptophane (pTrp) porte de cinq à dix répétitions en tandem d'une partie de l'opéron tryptophane (Panina *et al.*, 2001). Le nombre de copies de ces deux plasmides dépend du puceron hôte. Il n'a pas été déterminé chez *Acyrthosiphon pisum*, mais chez *Schizaphis graminum* les nombres de copies des gènes situés sur le plasmide tryptophane et le plasmide leucine par rapport aux gènes localisés sur le chromosome sont, respectivement, de 14,5 et 23,5, révélant une amplification importante (Plague *et al.*, 2003).

La leucine et le tryptophane sont indispensables à la croissance et à la survie de l'hôte et jouent un rôle essentiel dans la relation symbiotique. Certains auteurs expliquent donc le transfert des voies de biosynthèse de la leucine et du tryptophane sur deux plasmides afin de permettre leur surproduction pour le puceron (Baumann *et al.*, 1995 ; Komaki et Ishikawa, 2000), et l'adaptation de cette production aux variations environnementales indépendamment des gènes situés sur le chromosome (Plague *et al.*, 2003). Néanmoins, une étude expérimentale a récemment démontré que la production de tryptophane n'était pas corrélée avec le rapport entre le nombre de copies d'un gène codant une enzyme clé de la voie de biosynthèse du tryptophane sur le plasmide et le nombre de copies du gène sur le chromosome (Birkle *et al.*, 2002). Cette étude ne prend cependant pas en compte les copies inactives du gène et ne permet pas d'exclure un contrôle par rétroaction du nombre de copies du plasmide, en fonction de la concentration en tryptophane.

L'organisation du génome de *Buchnera* à un niveau global montre une réduction drastique de taille accompagnée par une extraordinaire quantité de copies aussi bien du chromosome que des plasmides. Cela rappelle les caractéristiques des plastes et des mitochondries, et ce d'autant plus que la bactérie est en partie transmise par voie maternelle entre les générations de pucerons.

1.2.212 Au niveau des gènes

Le génome de *Buchnera* Ap contient 610 gènes (dont 10 gènes localisés sur les plasmides) qui représentent 88 % de la séquence et 9 pseudogènes.

Synténie

Il existe un phénomène de synténie lorsqu'une conservation de l'ordre des gènes homologues entre les génomes de deux espèces est observable sur la

séquence du chromosome. En comparaison avec *Escherichia coli*, le génome de *Buchnera* comporte de nombreuses pertes dont notamment les gènes répétés et une partie des gènes indispensables pour mener une existence indépendante. Ainsi, certains gènes du métabolisme intermédiaire, et certains gènes impliqués dans les phénomènes de transport, de signalisation et de réparation de l'ADN sont absents {Shigenobu, 2000 #381;Silva, 2001 #680;Tamas, 2002 #220}. En revanche, une étude comparative des génomes des *Buchnera* Ap et Sg montre que le chromosome ne comporte aucune inversion, translocation, duplication ou même acquisition de gènes étrangers par transfert horizontal (Tamas, 2002). Le séquençage récent du génome de *Buchnera* Bp a permis de conforter cette observation. En effet, la comparaison des trois génomes a mis en évidence seulement quatre réarrangements entre *Buchnera* Bp et les génomes de *Buchnera* Ap et Sg (Van Ham *et al.*, 2003).

Cette remarquable stabilité génomique à l'origine d'une forte synténie entre les trois génomes s'explique par la perte de nombreux gènes impliqués dans les processus de recombinaison et d'insertion de l'ADN qui a abouti, d'une part, à la perte des capacités de réarrangement du génome, et d'autre part, à la difficulté d'acquisition de nouveaux gènes par transferts horizontaux (Silva *et al.*, 2003 ; Wernegreen *et al.*, 2000). De plus, la disparition des séquences répétées qui sont des sites potentiels de recombinaison (Tamas *et al.*, 2002), et la vie dans un milieu intracellulaire limitent considérablement la probabilité de transfert horizontaux.

Répartition des gènes sur les brins précoce et tardif du chromosome

Chez la plupart des bactéries, les gènes essentiels (c'est-à-dire tous les gènes dont la délétion par mutation ponctuelle est létale chez *Escherichia coli*) sont principalement localisés sur le brin précoce du chromosome (Rocha et Danchin, 2003b). Ainsi, des études menées chez *Bacillus subtilis* et *Escherichia coli* ont montré une fréquence élevée de gènes essentiels sur le brin précoce même pour des gènes à faible expression (Rocha et Danchin, 2003a). D'autres études ont également été réalisées pour une dizaine de γ-protéobactéries, dont *Buchnera*. Cependant des biais ont été observés dans les génomes des bactéries intracellulaires obligatoires comme *Buchnera*. Ils sont sans doute liés au problème posé par la définition de l'essentialité d'un gène. En effet, compte tenu du mode de vie particulier de *Buchnera*, il est loin d'être évident que les gènes essentiels chez *Escherichia coli* le soient chez *Buchnera*. Une définition adéquate de l'essentialité des gènes chez *Buchnera* couplée à une étude de leur expression permettrait d'étudier de façon pertinente le biais de positionnement des gènes et éventuellement de retrouver les résultats observés chez les γ-protéobactéries libres. Une première étude indique pour le moment que les gè-

nes supposés les plus fortement exprimés sont préférentiellement localisés sur le brin précoce du chromosome (Rispe *et al.*, 2004). Il existe de plus une dissymétrie de composition en bases très marquée entre les deux brins avec un excès de T et de G sur le brin direct (Rispe *et al.*, 2004).

1.2.213 Au niveau des séquences

Le génome de *Buchnera* avec un taux de GC de seulement 26,3 % est l'un des génomes présentant le biais le plus important en bases A et T (Clark *et al.*, 1999). Ce biais compositionnel est expliqué par un fort biais mutationnel vers les bases A et T durant la réduction massive du génome (Tamas *et al.*, 2002). Chez toutes les bactéries, il semble qu'une richesse en bases A et T soit corrélée à une taille relativement faible des gènes. Ceci peut s'expliquer de façon neutraliste en observant simplement que les différents codons « stop » riches en bases A et T, possèdent une probabilité d'apparition plus élevée chez les organismes à faible taux de GC. *Buchnera* n'échappe pas à cette règle avec une taille moyenne de 988 pb pour l'ensemble de ses gènes, ce qui est relativement faible en comparaison d'*Escherichia coli* (Charles *et al.*, 1999).

Les régions intergéniques sont encore plus riches en bases A et T que la moyenne du génome avec, pour certaines, un taux avoisinant les 90 % {Shigenobu, 2000 #381}. En revanche les gènes codant pour l'ARN 16S montrent une différence moins significative de leur taux en AT par rapport aux bactéries non symbiotiques, ce qui indique probablement une forte pression de sélection sur ces gènes (Woolfit et Bromham, 2003).

Des études basées sur l'utilisation des différents acides aminés pour la synthèse protéique chez *Buchnera,* par rapport à *Escherichia coli,* montrent l'existence d'une corrélation entre l'usage de certains acides aminés pour la synthèse protéique et l'expression des gènes correspondant chez *Buchnera* (Palacios et Wernegreen, 2002 ; Rispe *et al.*, 2004). En effet, les acides aminés des protéines codées par les gènes les plus exprimés chez *Buchnera* ne sont généralement pas des acides aminés aromatiques. Cette absence peut s'expliquer par le fait que ces acides aminés sont particulièrement coûteux à produire. De plus, les gènes les plus exprimés présentent également des taux en G et en C plus élevés que le reste du génome. Cette constatation s'explique sans doute par l'existence d'une pression de sélection plus importante sur les gènes fortement exprimés qui limiterait le biais mutationnel vers les bases A et T. Il est également possible d'imaginer que *Buchnera* évite d'utiliser les acides aminés aromatiques afin de les fournir préférentiellement au puceron.

1.2.22 Un métabolisme essentiellement dédié à la biosynthèse des acides aminés

Buchnera fournit à son hôte les acides aminés essentiels qui sont absents de son milieu nutritionnel (Douglas, 1998 ; Sasaki et Ishikawa, 1995) et le séquençage du génome de la bactérie a conforté cette évidence. En effet, bien que *Buchnera* ne possède plus que 35 % des gènes d'*Escherichia coli*, son génome contient encore 45 % des gènes codant pour la biosynthèse des acides aminés (Douglas *et al.*, 2001) (cf. **Tableau A.1.1**).

1.2.221 Les voies de biosynthèse des acides aminés

L'analyse du génome de *Buchnera* montre que les voies de biosynthèse des acides aminés non essentiels sont pour la plupart complètement absentes (à part celle de la cystéine), contrairement à celles de la biosynthèse des neuf acides aminés essentiels pour le puceron qui sont parfaitement conservées (cf. **Figure A.1.4**). Il semble donc que les processus évolutifs favorisant la perte des gènes chez *Buchnera* (Rispe et Moran, 2000) soient compensés par une sélection retenant les gènes de biosynthèse des acides aminés essentiels. Le génome de la bactérie contient ainsi 10 % de gènes ayant des fonctions dans le métabolisme des acides aminés essentiels, alors qu'*Escherichia coli* n'en compte que 2 %. L'étude des voies de biosynthèse de *Buchnera* laisse donc supposer que la bactérie fournit à son hôte les acides aminés qu'il ne peut synthétiser et réciproquement que l'hôte apporte à la bactérie les acides aminés qui lui font défaut. De plus, comme les précurseurs de plusieurs acides aminés essentiels sont des acides aminés non essentiels (glutamate et aspartate par exemple, cf. **Figure A.1.4**). les voies de biosynthèse de l'hôte et du symbiote sont non seulement complémentaires mais également dépendantes les unes des autres {Shigenobu, 2000 #381}.

Tableau A.1.1 Nombre de gènes chez Buchnera par catégorie fonctionnelle en fonction de la classification de Riley (1993). La liste complète des gènes est disponible sur le site de l'institut RIKEN : http://Buchnera.gsc.riken.go.jp.

Métabolisme des petites molécules	
Dégradation des petites molécules	3
Métabolisme énergétique	51
Glycolyse	9
Déshydrogénation du pyruvate	3
Cycle de l'acide tricarboxylique	2
Voie des pentoses phosphate	6
Respiration	23
Fermentation	0
Force motrice ATP-proton	8
Métabolisme central intermédiaire	13
Biosynthèse des acides aminés	55
Voie du glutamate	8
Voie de l'aspartate	11
Voie de la sérine	4
Voie des acides aminés aromatiques	16
Histidine	8
Voie du pyruvate	0
Voie des chaînes branchées	8
Biosynthèse de polyamine	2
Purines, pyrimidines, nucléosides et nucléotides	34
Biosynthèse des cofacteurs, groupements prosthétiques et transporteurs	26
Biosynthèse des acides gras	6
Fonctions de régulation	
	7
Métabolisme des macromolécules	
Synthèse et modification des macromolécules	187
ARN ribosomiaux et ARN stable	4
Modification des protéines ribosomales, synthèse et modification de protéines	54
Maturation et modification des ribosomes	0
ARN de transfert	32
Synthèse des ARNt aminoacylés et modification	34
Nucléoprotéines	2
Réplication de l'ADN, restriction, modification et recombinaison	32
Traduction et modification des protéines	18
Synthèse des ARN, modification des ARN et transcription	9
Polysaccharides (cytoplasmique)	0
Phospholipides	2
Dégradation des macromolécules	22
Enveloppe cellulaire	46
Membranes, lipoprotéines et porines	4
Polysaccharides de surface, lipopolysaccharides, et antigènes	2
Structures de surface	25
Muréine et peptidoglycane	15
Processus cellulaires	
Transport et protéines de liaison	18
Chaperonnes	11
Division cellulaire	12
Chimiotactisme et mobilité	0
Sécrétion de protéines et de peptide	7
Adaptation osmotique	0
Détoxification	3
Autres	
Fonctions en relation avec la colicine	1
Sensibilité aux médicaments et analogues	2
Adaptations et conditions atypiques	2
Autres non classés	111
Autres catégories	28
Hypothétiques conservés	79
Non connus (unique chez *Buchnera*)	4

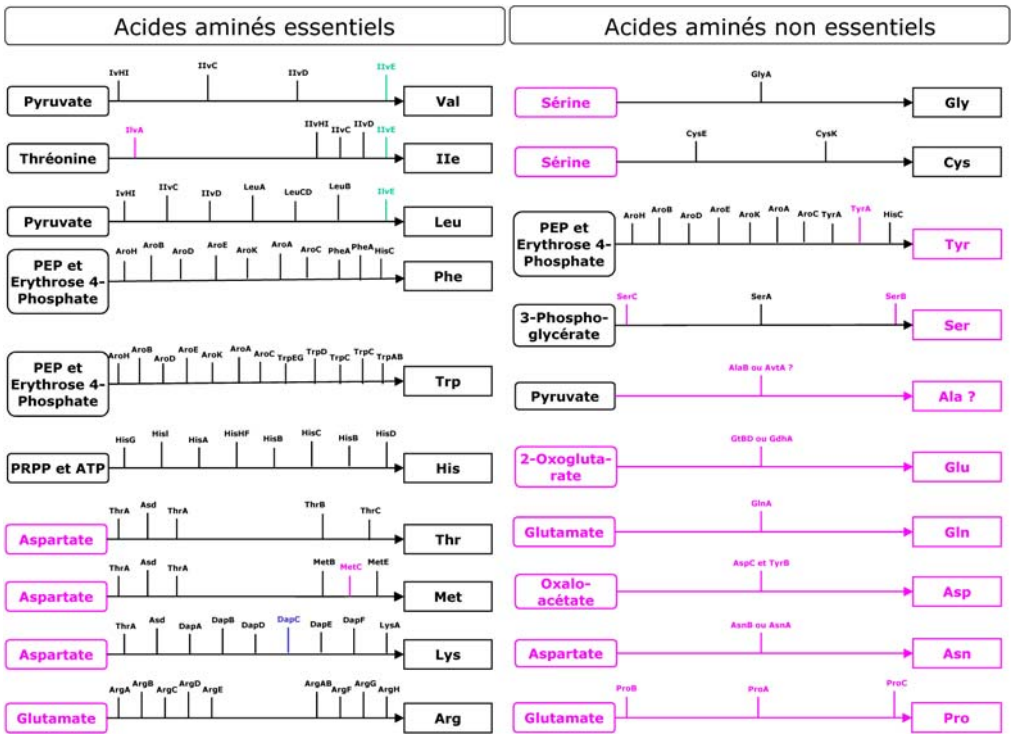


Figure A.1.4 Représentation des voies de biosynthèse des acides aminés essentiels (à gauche) et non essentiels (à droite) déduites de l'annotation du génome de *Buchnera* {Shigenobu, 2000 #381}. Les étapes pour lesquelles aucun gène n'a été annoté sont indiquées en rose.

Voies de biosynthèse des acides aminés essentiels : L'enzyme *ilvE* (en vert) est absente des voies de biosynthèse de la valine, l'isoleucine et la leucine, mais une autre aminotransférase assure sans doute ces étapes. Dans la voie de biosynthèse de la lysine, le gène *dapC* (en bleu) n'a été identifié chez aucun organisme. Enfin, la dernière étape de biosynthèse de la phénylalanine est catalysée par *TyrB* chez *Escherichia coli* mais peut être remplacée par *HisC* chez *Buchnera*.

Voies de biosynthèse des acides aminés non essentiels : De même que chez *Escherichia coli*, la voie de biosynthèse de l'alanine n'a pas été élucidée.

L'observation attentive des voies de biosynthèse des acides aminés essentiels montre néanmoins que deux gènes sont absents. Le premier est le gène *ilvA*, qui code pour la thréonine désaminase, première enzyme de la voie de biosynthèse de l'isoleucine, et le second est *metC*, qui code pour la cystathione β-lyase, enzyme de la voie de biosynthèse de la méthionine. Pourtant les données concernant la nutrition des pucerons suggèrent que *Buchnera* est capable de synthétiser l'isoleucine et la méthionine. Plusieurs hypothèses ont été avancées pour expliquer cette incohérence entre données génomiques et physiologiques. Il est en effet possible que d'autres enzymes catalysent ces réactions en

raison d'une faible spécificité de substrat, ou encore que *ilvA* et *metC* aient été transférés vers le génome du puceron, que ces deux gènes soient transcrits et traduits dans les bactériocytes puis que les enzymes soient fournies à *Buchnera* (Douglas, 2003).

Contrairement à *Escherichia coli* qui possède 86 ARN de transfert, *Buchnera* ne contient plus que 32 ARN de transfert, ce qui implique que la plupart des acides aminés n'ont qu'un seul ARN de transfert (Moran et Mira, 2001). Ce nombre limité d'ARN de transfert pourrait éventuellement permettre une vitesse plus élevée de traduction et donc de biosynthèse des acides aminés pour son hôte.

Enfin, il est à noter que, contrairement aux parasites obligatoires, *Buchnera* possède non seulement les voies de biosynthèse des acides aminés essentiels mais que la plupart des voies de biosynthèse des nucléotides sont également complètes. Aucune étude n'a été menée à ce jour pour déterminer si la bactérie produit des nucléotides pour son propre usage ou pour son hôte {Shigenobu, 2000 #381}.

1.2.222 Le métabolisme énergétique

Les données génomiques montrent que *Buchnera* dispose d'un métabolisme aérobie, ce qui est compatible avec sa localisation dans le bactériocyte, qui reçoit certainement l'oxygène circulant dans les trachées de l'insecte et dont le cytoplasme contient de nombreuses mitochondries. En revanche *Buchnera* a perdu les gènes impliqués dans les phénomènes de fermentation et de respiration anaérobie.

L'opéron de la déshydrogénase NADH (*nuo*) et l'opéron du cytochrome C (*cyo*) sont parfaitement conservés, en revanche la voie de biosynthèse de l'ubiquinone est absente. La bactérie possède également l'opéron d'une ATP synthase (type F_0F_1) indiquant sa capacité à produire de l'ATP à partir du gradient électrochimique de protons généré par le système de transport d'électrons.

Buchnera possède l'ensemble des gènes du cycle du pentose phosphate et ceux de la glycolyse, excepté le gène codant pour l'hexokinase qui permet la phosphorylation du glucose. Cependant, il est possible que la bactérie importe et phosphoryle le glucose en même temps, d'autant plus que la bactérie possède deux gènes (*crr* et *ptsG*) spécifiques de l'importation du glucose (Tamas *et al.*, 2001). En revanche seuls deux gènes (*sucA* et *sucB*) du cycle de l'acide tricarboxylique (TCA) sont présents dans le génome. Ce cycle de la respiration aérobie permet normalement l'oxydation du pyruvate en dioxyde de carbone. Cependant, les gènes codant pour le complexe pyruvate déshydrogénase sont

présents, suggérant que le pyruvate produit par la glycolyse peut être converti en acétyl-coA (Tamas *et al.*, 2001).

1.2.223 Que reste-t-il des autres voies ?

Buchnera semble posséder une capacité limitée de réparation de l'ADN et présente un répertoire plutôt inhabituel de gènes appartenant à cette catégorie. En effet, *Buchnera* ne semble pas posséder de méthylase (Akman et Aksoy, 2001) et le gène *recA* est absent alors qu'il s'agit d'un élément crucial pour les recombinaisons homologues (Hughes, 2000). *Buchnera* est actuellement le seul organisme connu à posséder *recBCD* sans *recA* et ce, bien que de nombreuses espèces aient perdu *recA* (Marais *et al.*, 1996). Il est possible que le produit du gène *recBCD* soit une exonucléase qui assure la fonction de réparation à la place de la réparation par recombinaison de *recA* (Van Ham *et al.*, 2003). De même dans le système de réparation *uvr*, *Buchnera* a perdu le gène *uvrAB*C mais conservé *mfd*. Or, excepté chez *Buchnera*, ces deux gènes sont conservés dans tous les génomes séquencés d'eubactéries (Eisen et Hanawalt, 1999). Cet inventaire unique de gènes montre que les mécanismes de réparation et de recombinaison de l'ADN sont sévèrement détériorés chez *Buchnera*. Quant à l'absence de nombreux gènes impliqués dans le système SOS (*recA*, l*exA*, *umuCD* et *uvrABC*), elle indique que le génome de *Buchnera* est potentiellement vulnérable aux dommages de l'ADN.

En ce qui concerne les aspects structuraux, *Buchnera* ne semble pas capable de fabriquer des lipopolysaccharides. Les gènes impliqués dans leur biosynthèse sont en effet tous absents, exceptés *lpcA* et *kdtB*. Par ailleurs, seul un petit nombre de gènes codants pour des lipoprotéines ou des protéines membranaires sont présents. La membrane de *Buchnera* est donc particulièrement vulnérable, ce qui s'explique certainement par son mode de vie intracellulaire qui la protège des attaques extérieures. En revanche, l'absence des gènes responsables de la biosynthèse des phospholipides (excepté *cls* codant pour la cardiolipine synthétase) est plus surprenante car les phospholipides sont des composants indispensables de la membrane. Il est donc possible que *Buchnera* importe les phospholipides de son hôte ou qu'elle les synthétise avec des enzymes dont les gènes ont été transférés dans le noyau de son hôte. Ces constatations peuvent aussi indiquer que *Buchnera* est dans un processus d'abandon des voies de biosynthèse de son enveloppe cellulaire, enveloppe qui pourrait être remplacée par une membrane d'origine eucaryote.

1.2.224 Le transport des métabolites

L'état de symbiose obligatoire qui existe entre *Buchnera* et le puceron nécessite certainement de nombreux échanges entre le symbiote et le cytoplasme de la

cellule hôte. Pourtant, seuls 18 gènes, soit 3 % du génome de la bactérie sont annotés comme transporteurs (cf. **Tableau A.1.1**). À titre de comparaison, le génome d'*Escherichia coli* compte 661 gènes potentiellement impliqués dans le transport[3], soit plus de 15 % de son génome.

Le génome de *Buchnera* ne contient qu'un faible nombre de gènes impliqués dans le système de transport spécifique ABC qui assure le transfert de nutriments et autres substances à travers la membrane, souvent contre un gradient de concentration ou par hydrolyse d'ATP. Ce système est pourtant présent de façon ubiquitaire chez les bactéries (Tomii et Kanehisa, 1998). Le seul système de transport spécifique annoté chez *Buchnera* est le complexe phosphoenolpyruvate-carbohydrate phosphotransférase (PTS) qui permet l'importation du glucose et du mannitol. Pour ce qui est du transport non spécifique, les protéines hypothétiques *YnfM* et *YajR* sont probablement des transporteurs de faible affinité. De même *GlpF* et les porines *ompA* et *ompF* sont probablement impliqués dans des phénomènes de diffusion passive et les gènes du système de sécrétion protéique *sec* sont conservés {Shigenobu, 2000 #381}.

En ce qui concerne le transport des acides aminés, il n'existe actuellement aucune donnée permettant de valider l'existence d'un transport actif d'acides aminés vers l'extérieur ou vers l'intérieur de la bactérie et aucun gène n'a pu lui être attribué. En revanche, le génome de *Buchnera* contient un nombre étonnamment élevé de gènes codant pour des protéines du flagelle. Les gènes du flagelle chez *Buchnera* sont situés dans deux opérons *flg* et *fli* contenant 12 gènes chacun. Il existe également un petit opéron supplémentaire composé de 2 gènes (*flhA* et *flhB*). Ces 26 gènes représentent une grande partie du génome de *Buchnera*, suggérant une fonction importante (Tamas *et al.*, 2001). Plusieurs arguments ont été avancés pour un rôle possible de ces gènes dans le transport. Il est en effet possible que la bactérie utilise son flagelle pour effectuer des transports actifs, comme cela a déjà été observé chez *Salmonella typhimurium* (Blattner *et al.*, 1997 ; Kubori *et al.*, 1998) et *Yersinia enterocolitica*. Pour cette dernière, le flagelle assure à la fois des fonctions d'import et d'export des protéines (Young *et al.*, 1999). Chez *Buchnera*, la séquence des gènes du flagelle montre un niveau élevé de divergence par rapport à *Escherichia coli,* suggérant également une modification de leur fonction (Tamas *et al.*, 2002). Par ailleurs, le flagelle est généralement composé de trois éléments : un corps basal, un crochet et un filament. Or *Buchnera* ne possède pas le gène *fliC* qui est normalement impliqué dans la synthèse du filament responsable de la

[3] http://ecocyc.org.

mobilité cellulaire. Les cellules de *Buchnera* ne semblent d'ailleurs pas posséder de flagelle et aucun mouvement n'a jamais été observé, d'autant que la bactérie a également perdu les gènes impliqués dans les phénomènes de chimiotactisme.

Une seconde alternative est que le transport est assuré par l'hôte, probablement au niveau de la membrane symbiosomale entourant les bactéries. Actuellement, il n'existe aucune donnée expérimentale permettant de confirmer cette hypothèse et les propriétés de cette membrane sont complètement inconnues. De même d'autres protéines de la bactérie, et notamment celles impliquées dans le transport, pourraient être en voie de remplacement par des protéines de l'hôte, comme cela a déjà été observé dans les mitochondries (Kurland et Andersson, 2000). Si tel est le cas, il devient difficile de définir avec précision le passage de l'état d'organisme à celui d'organite cellulaire (Andersson, 2000).

1.2.23 Une régulation de l'expression des gènes plutôt énigmatique

1.2.231 De rares éléments de régulation

En comparaison des bactéries libres, le génome des bactéries parasitaires et symbiotiques ne contient que très peu d'éléments de régulation de la transcription (Wilcox *et al.*, 2003). *Buchnera* n'échappe pas à ce constat et l'analyse de son génome a mis en évidence l'absence de nombreux systèmes de régulation présents chez *Escherichia coli*.

Les gènes codant pour les deux systèmes impliqués normalement dans le contrôle de l'expression des gènes en fonction des conditions environnementales sont complètement absents. Excepté *dnaA*, les autres types de régulateurs dont les systèmes d'atténuation de la transcription sont également absents {Panina, 2001 #178;Shigenobu, 2000 #381}. De plus, le génome de *Buchnera* ne contient ni les gènes codant pour l'adénylate cyclase (*cyaA*), ni ceux codant pour les récepteurs protéiques de l'AMP cyclique (*crp*). La répression catabolique par l'AMP cyclique n'est donc pas possible chez *Buchnera*, malgré la présence dans le génome de protéines généralement impliquées dans cette réponse régulatrice. La bactérie ne possède donc plus les éléments de régulation de la transcription permettant de répondre à un changement de source de carbone dans l'environnement, ce qui est cohérent avec son mode de vie intracellulaire. En revanche, la bactérie possède un régulateur normalement impliqué dans le stockage du carbone, codé par le gène *csrA*, qui est peut-être impliqué dans une régulation globale et post-transcriptionnelle du métabolisme des sucres (Nogueira et Springer, 2000). Enfin, contrairement à *Bacillus subtilis* et *Escherichia coli* qui possèdent respectivement 18 et 7 gènes codant pour des facteurs

de transcription sigma, le génome de *Buchnera* ne possède que le gène *rpoH* qui code pour le facteur σ32, et le gène *rpoD* qui code pour le facteur σ70 {Shigenobu, 2000 #381}.

La perte des promoteurs semble associée de façon fréquente à l'apparition d'un nouveau groupe de gènes contigus qui forment une nouvelle unité de transcription. D'autant plus que dans la plupart des cas, le regroupement des gènes au sein d'un polycistron a également été suivi de la perte des gènes présents sur le brin opposé de l'ADN (Moran et Mira, 2001). Ce regroupement des gènes au sein d'une même unité de transcription peut être interprété comme un mécanisme favorisant l'efficacité de la transcription ou de la traduction (Shcherbakov et Garber, 2000).

Il est important de noter que l'hypothèse de la perte des régions de régulation est basée sur les homologies de séquences entre le génome de *Buchnera* et le génome actuel d'*Escherichia coli*. Or aucune étude expérimentale n'a permis d'écarter la possibilité que les promoteurs apparemment dégradés de *Buchnera* soient encore fonctionnels, ou que le symbiote puisse utiliser des promoteurs différents de ceux qui ont été identifiés chez *Escherichia coli*. De plus, la recherche de motifs de régulation dans les régions intergéniques de *Buchnera* est rendue particulièrement difficile par la richesse importante en AT (Baumann *et al.*, 1995). Enfin, la position des codons « start » des gènes n'a pas été vérifiée expérimentalement et peut, en cas d'erreur, biaiser la recherche des séquences régulatrices (Moran et Mira, 2001). De même, de nombreux gènes régulés au niveau transcriptionnel chez *Escherichia coli* n'ont pas de sites de liaisons potentiels dans les génomes des autres γ-protéobactéries. Il est possible que cela soit simplement un artefact lié aux techniques de recherche utilisées qui ne permettent de retrouver que des régions hautement conservées par rapport à *Escherichia coli.* Or si ces régions ont pour la plupart été déterminées expérimentalement chez *Escherichia coli,* les sites correspondant chez les autres bactéries (et en particulier chez *Buchnera* en raison du fort biais AT) n'ont pas été découverts de façon expérimentale mais par analyse bioinformatique.

Les conséquences de la perte des gènes de régulation sont également à modérer en fonction du rôle de l'hôte dans la relation endocytobiotique. D'une part, il est possible que le puceron fournisse au symbiote les protéines ou molécules régulatrices dont ce dernier a besoin. D'autre part, il existe peut-être des transferts de gènes de la bactérie vers le noyau de l'hôte. Cette dernière hypothèse pourra être vérifiée lorsque le génome complet du puceron sera disponible.

L'étude du génome de bactéries parasitaires telles que *Mycoplasma genitalium* (Fraser *et al.*, 1995) ou *Rickettsia prowazekii* a mis en évidence des

pertes similaires des systèmes de régulation. Cependant, la perte des régulateurs s'accompagne en général de la disparition simultanée des gènes qui sont sous leur contrôle. Le génome de *Buchnera* est pour cela un spécimen unique, dans le sens où il possède encore beaucoup de gènes cibles des systèmes de régulation absents. Sur ce constat, {Shigenobu, 2000 #381} imaginent *Buchnera* comme un organite spécialisé dans la fourniture d'acides aminés essentiels, de la même façon que les mitochondries fournissent l'énergie aux cellules eucaryotes. Il existe cependant une différence fondamentale entre la fourniture d'énergie et celle de nutriments, car la fourniture d'acides aminés nécessite une adaptation constante avec l'environnement. Quelques expériences ont donc été orientées vers l'étude de la capacité de *Buchnera* à répondre à des variations environnementales, elles sont présentées dans les paragraphes suivants.

1.2.232 Régulation du choc thermique

Le mécanisme de régulation du choc thermique a été le plus étudié chez de nombreuses bactéries modèles car il compte parmi les plus conservés et les plus fortement induits au niveau transcriptionnel (Neidhart *et al.*, 1996). Il s'agit de plus d'une des expériences les plus simples à mettre en œuvre. Ce choix est certainement discutable pour une bactérie comme *Buchnera* qui, en plus de vivre dans un milieu intracellulaire particulièrement protégé des variations environnementales, ne semble pas avoir conservé l'ensemble des gènes impliqués dans la protection contre le choc thermique. Elle n'a cependant pas échappé à cette première expérience portant sur la régulation de l'expression des gènes (Wilcox *et al.*, 2003).

Les pucerons, comme l'ensemble des insectes ne sont pas capables de réguler la température. Des études préliminaires ont montré que les bactéries *Buchnera* meurent lorsque les pucerons sont soumis à des températures supérieures à 37 °C (Ohtaka et Ishikawa, 1991) et que la densité des symbiotes est fortement réduite durant la saison estivale (Montllor *et al.*, 2002).

Chez *Escherichia coli*, le choc thermique induit une augmentation de l'expression du gène codant pour le facteur de transcription σE. σE est lui-même responsable de l'expression du gène *rpoH* codant pour le facteur σ32 et de la déstabilisation d'une structure secondaire du transcrit *rpoH*, diminuant son affinité pour les gènes *Dnak* et *Dnaj*. Cette affinité est particulièrement élevée en condition normale. Cela permet donc la synthèse et la libération du facteur σ32 qui active les gènes codant pour les protéines impliquées dans la réponse au choc thermique dont la protéine *GroEL*. *Buchnera* possède vingt des trente-cinq gènes impliqués dans le choc thermique chez *Escherichia coli* dont

les gènes *rpoH* codant le facteur σ32 et le gène *mopA* codant pour la protéine *GroEL* (Wilcox *et al.*, 2003).

Une étude globale de l'expression des gènes du symbiote *Buchnera* de *Schizaphis graminum* a été réalisée pour étudier sa réponse au choc thermique (Wilcox *et al.*, 2003). Pour cela, les pucerons ont été soumis à des températures élevées puis l'expression des gènes des bactéries symbiotiques a été étudiée au moyen d'une puce à ADN avec pour sondes des produits de PCR représentant l'ensemble du génome. Les données obtenues ont été comparées à des données d'expression d'*Escherichia coli* et de *Bacillus subtilis* soumis également à un stress thermique. Cette méthodologie est discutable dans la mesure où *Buchnera* bénéficie de la protection du milieu intracellulaire de l'hôte. D'autant plus que les bactéries ne sont pas réellement dans les mêmes conditions. Alors que *Buchnera* est exposée durant 120 minutes à une température qui atteint progressivement 36 °C, *Escherichia coli* est exposée 7 minutes à 50 °C (Richmond *et al.*, 1999) et *Bacillus subtilis* (Helmann *et al.*, 2001) 2 minutes à 48 °C.

Les résultats de cette expérience montrent néanmoins que dix gènes de *Buchnera* présentent une surexpression. Parmi ceux-ci, cinq appartiennent au régulon de choc thermique d'*Escherichia coli* (*mopA*, *mopB*, *dnaK*, *grpE*, *ibpA*) et un sixième, *dnaJ* fait également partie des gènes de stress thermique chez *Escherichia coli*. Sur les dix-huit gènes de *Buchnera* Sg qui sont orthologues des gènes de choc thermique surexprimés chez *Escherichia coli*, six présentent donc une induction significative, soit 1/3. Par ailleurs, deux autres gènes, *Fpr* (exprimé au cours du stress oxydant *chez E coli*) et *yieA* sont surexprimés, probablement en même temps que *ibpA*, avec lequel ils forment un polycistron en raison de la réduction du génome. Pour les deux derniers gènes (*ychF* et *rne*), les auteurs n'avancent aucune raison à leur surexpression (Wilcox *et al.*, 2003). Ils ne commentent pas d'avantage l'absence de surexpression du gène *rpoH,* qui joue pourtant théoriquement un rôle clé dans l'activation de la réponse au choc thermique. Et malgré les nombreux aspects discutables de cette expérience, les auteurs concluent cette étude sur l'incapacité de *Buchnera* à réguler l'expression de ses gènes en réponse au choc thermique.

Il est possible cependant d'avoir une autre interprétation de cette expérience. En conditions normales, la synthèse de la protéine chaperonne *GroEL* est élevée chez la bactérie, suggérant que le promoteur est constitutivement actif. Cependant, le niveau de σ32 dans les cellules de *Buchnera* n'est pas détectable (Wilcox *et al.*, 2003). L'étude des séquences montre que le gène *rpoH*, contrairement aux autres bactéries, n'est pas précédé d'une séquence promotrice spécifique du facteur σE, suggérant que le gène *rpoH* ne présente pas de séquence régulatrice (Yuzawa *et al.*, 1993). De plus, le transcrit *rpoH* ne possède pas la structure secondaire qui permet normalement la régulation de sa tra-

duction (Nakahigashi *et al.*, 1995). Enfin, en dépit d'une structure relativement bien conservée, σ32 est incapable de complémenter la fonction chez des mutants d'*Escherichia coli* (Sato et Ishikawa, 1997). Ces études semblent suggérer que la perte de la fonction de σ32 chez *Buchnera* n'est pas due à des modifications importantes, mais plutôt à de faibles changements qui affectent l'ensemble de la molécule. Par ailleurs, le taux de AT dans la région promotrice du choc thermique est extrêmement élevé. Ce biais AT pourrait favoriser l'expression de l'opéron sans nécessiter une activation par le σ32, comme cela a été montré pour les opérons des ARN ribosomiaux chez *Escherichia coli* (Petho *et al.*, 1986). Ces études montrent donc une expression constitutive des protéines de l'opéron choc thermique.

En ce qui concerne la surproduction de la protéine *GroEL* en conditions normales chez *Buchnera*, cette caractéristique est également retrouvée chez les mitochondries et les plastes (Tamas, 2002). Mais bien que *GroEL* représente 10 % des protéines retrouvées chez *Buchnera* (Sato et Ishikawa, 1997), sa fonction n'est pas encore complètement élucidée. Elle pourrait jouer un rôle dans le transport membranaire ou le repliement des protéines dégradées. Une étude bioinformatique du repliement des protéines de *Buchnera* et d'autres bactéries intracellulaires, a d'ailleurs montré une efficacité de repliement plus faible que celles des protéines des bactéries libres (Van Ham *et al.*, 2003). L'effet des mutations délétères qui affectent le génome de *Buchnera* pourrait ainsi être compensé par la surexpression constitutive de la chaperonne *GroEL* (Moran, 1996). De même, le repliement très bien conservé de la chaperonne *DnaK* suggère que la protéine soumise à une relativement forte pression de sélection, assure une fonction similaire (Fares *et al.*, 2002a). Des expériences menées chez *Escherichia coli* ont confirmé que les chaperonnes pouvaient compenser la présence de mutations délétères en assurant tout de même un repliement correct des protéines (Fares *et al.*, 2002b). L'étude de *GroEL* et des mécanismes de régulation associés suggère donc que *Buchnera* utilise les gènes normalement dédiés à la réponse au choc thermique pour d'autres fonctions que celles qui sont assurées chez *Escherichia coli*, probablement en raison de son mode de vie intracellulaire. Dans ce cas, la question n'est plus de savoir si *Buchnera* possède encore la capacité à réguler l'expression de ces gènes en réponse au choc thermique, puisque certains des gènes impliqués normalement dans cette réponse sont peut-être dédiés à une autre fonction essentielle.

Il semble donc que l'étude de la régulation du choc thermique ne soit pas la plus appropriée pour analyser les mécanismes de régulation d'une bactérie intracellulaire telle que *Buchnera,* et ce d'autant plus que le rôle principal de la bactérie est certainement la biosynthèse d'acides aminés essentiels pour son

hôte. Le travail expérimental réalisé dans cette thèse a donc été orienté vers l'étude de la régulation de la biosynthèse des acides aminés chez *Buchnera*.

1.2.233 Régulation de la biosynthèse des acides aminés

La synthèse des acides aminés chez les bactéries est généralement régulée à un niveau post-traductionnel par un système de rétroaction négative de l'acide aminé produit sur la dernière enzyme de sa voie de biosynthèse (régulation allostérique). Un acide aminé est donc synthétisé en abondance uniquement s'il est présent en faible quantité. En plus de cette première possibilité, de nombreux systèmes de régulation existent au niveau transcriptionnel. Cependant, chez *Buchnera,* l'annotation du génome indique que les régulateurs, incluant les systèmes d'atténuation, sont apparemment absents pour tous les gènes contribuant à la synthèse des acides aminés (Tamas *et al.*, 2002).

Des études théoriques ont donc été menées pour étudier de façon plus spécifique la régulation de la synthèse des acides aminés. Elles ont d'abord été conduites pour quelques acides aminés en particulier.

Ainsi, une analyse de la régulation de la synthèse de la phénylalanine par l'étude de la séquence génomique semble indiquer une surproduction constitutive de cet acide aminé. La première raison est l'absence de région d'atténuation, pourtant souvent observée chez d'autres bactéries, et la seconde est la dégradation du site de liaison allostérique de la phénylalanine sur l'enzyme intervenant dans sa synthèse. Cette dégradation empêche vraisemblablement le rétrocontrôle négatif de la phénylalanine sur sa propre synthèse (Jimenez *et al.*, 2000). La perte apparente de ces éléments de régulation contraste avec les études nutritionnelles qui suggèrent que *Buchnera* est capable de fournir au puceron la phénylalanine absente de son milieu nutritionnel (Douglas *et al.*, 2001).

D'autres études ont porté sur l'analyse de la régulation du tryptophane. Une analyse théorique du génome de la bactérie semble indiquer que la production de tryptophane est libre de tout contrôle (Panina *et al.*, 2001). Cette étude met néanmoins en évidence une grande variété de systèmes de régulation chez différentes bactéries, suggérant que le mode de régulation des bactéries est sous l'influence d'une forte pression de sélection qui dépend fortement du mode de vie des bactéries. Cela peut suggérer la possibilité d'un contrôle différent pour une bactérie intracellulaire comme *Buchnera*. Une autre étude a été menée avec des clones différents de pucerons pour analyser le niveau d'amplification des gènes de la biosynthèse du tryptophane en comparaison avec le taux de synthèse du tryptophane. Pour cela, les rapports entre le nombre de copies du gène de l'anthranilate synthase sur le plasmide tryptophane (*trpG*) et le nombre de copies du gène porté par le chromosome (*trpE*) ont été calculés pour tous les

clones. L'amplification des gènes portés par les plasmides varie de deux à seize fois suivant les clones, et bien que la production de tryptophane soit très variable, elle n'est pas corrélée avec le niveau d'amplification (Birkle *et al.*, 2002). Cependant cette étude se limite à l'analyse d'une seule enzyme. Or, le contrôle de la synthèse d'un métabolite ne dépend certainement pas du changement d'activité d'une seule enzyme, mais de nombreuses enzymes. Elle ne peut donc pas à elle seule remettre en cause la possibilité d'un contrôle de la biosynthèse du tryptophane par l'amplification des gènes sur le plasmide.

En ce qui concerne la leucine, quelques éléments de régulation ont tout de même été découverts. Ainsi, une analyse des régions intergéniques du plasmide leucine a révélé la présence de séquences bien conservées en amont du gène *repA2* qui pourraient correspondre à des sites de fixation de la sous-unité σ70 de l'ARN polymérase. Ces régions sont également présentes dans l'opéron des gènes des ARN 16S et 23S (Munson *et al.*, 1993). Par ailleurs, une autre étude a mis en évidence la présence d'une longue séquence palindromique proche de la région intergénique séparant *repA2* et *leuA*. Cette séquence est relativement bien conservée dans les différentes espèces de *Buchnera*, ce qui suggère une contrainte fonctionnelle forte (Silva *et al.*, 1998). Elle pourrait éventuellement être impliquée dans le contrôle de l'expression de l'opéron leucine par la formation d'une épingle à cheveux qui permettrait la continuité de la transcription (Rutberg, 1997) ou la stabilisation des transcrits (Emory *et al.*, 1992 ; Wong et Chang, 1986), comme cela a été observé chez d'autres bactéries.

Une étude très générale de l'expression des gènes par PCR soustractive pour des *Buchnera* issues de pucerons jeunes et âgés montre une surexpression de deux gènes du métabolisme des acides aminés (*argA* et *thrB*) impliqués respectivement dans les voies de synthèse du glutamate à partir de l'arginine et de la thréonine à partir de l'aspartate (Nakabachi et Ishikawa, 1997). Cette expérience suggère que les bactéries sont capables d'adapter leur production d'acides aminés en fonction de l'âge des pucerons.

Une expérience menée pour étudier plus spécifiquement la régulation de la biosynthèse des acides aminés de façon globale a également été réalisée. Pour cela, la puce utilisée pour l'étude du choc thermique chez la bactérie symbiotique de *Schizaphis graminum* a été de nouveau employée (Moran, 2003). Elle révèle que *Buchnera* produit un nombre important de transcrits codant pour la biosynthèse de l'arginine et de la lysine qui sont deux acides aminés absents du milieu nutritionnel des pucerons des céréales. L'étude porte sur des bactéries issues de pucerons élevés sur des plantes présentant des taux en acides aminés favorables et défavorables (d'après la définition de Telang *et al.*, 1999). L'étude semble montrer une faible différence d'expression entre les deux types de plantes sur lesquelles les pucerons ont été élevés, mais aucune valeur n'est

explicitée. L'auteur avance plusieurs hypothèses pour expliquer ces résultats. Il envisage une absence de régulation ou un biais éventuel lié à une forte hétérogénéité dans la population bactérienne. En effet, les pucerons en plus de leur propre population, dite maternelle, de bactéries symbiotiques, contiennent une forte densité de bactéries dans les embryons (cf. 1.2.13). Enfin il n'exclut pas la possibilité d'une régulation assurée par un transport actif d'acides aminés par l'hôte lui-même. Il convient certainement d'attendre la publication de cette étude (les résultats n'ont pour l'instant fait l'objet que d'une présentation lors d'un colloque). De plus les auteurs ne semblent pas maîtriser et connaître parfaitement la composition de la sève des plantes utilisées.

Les processus régulant la synthèse des acides aminés essentiels sont-ils forcément visibles dans le génome de *Buchnera* ? Si la synthèse des acides aminés essentiels est limitée par le substrat, c'est-à-dire si elle est régulée par l'approvisionnement en précurseurs, alors la réponse sera négative. Cependant, une régulation de la biosynthèse des acides aminés essentiels chez *Buchnera* par la production semble peu cohérente avec son mode de vie symbiotique. Il serait sans doute plus « intéressant » pour le symbiocosme que la production en acides aminés essentiels soit contrôlée par la fourniture en précurseurs. Par ailleurs, les gènes de la biosynthèse des acides aminés sont fortement exprimés chez *Buchnera* (Baumann *et al.*, 1999). Or une modélisation de la dynamique des voies métaboliques chez *Escherichia coli* a montré que la régulation est positive pour des niveaux élevés d'expression des gènes et négative pour des niveaux faibles (Neidhart *et al.*, 1996). Cet argument laisse envisager un contrôle non pas négatif, mais positif de la synthèse d'acides aminés chez *Buchnera*.

Ainsi, le contrôle de l'expression des gènes de la biosynthèse des acides aminés reste encore une énigme chez *Buchnera*. Il est vrai que les caractéristiques du génome rendent la recherche des éléments de régulation difficile et que seules quelques expériences ont été réalisées pour étudier spécifiquement cette régulation chez *Buchnera*. Mais, même si les éléments traditionnels de régulation sont réellement absents, le mode de vie particulier de *Buchnera* laisse envisager que les mécanismes mis en œuvre pour la régulation de la biosynthèse des acides aminés puissent être différents de ceux qui existent chez les autres bactéries.

1.2.234 Et si la réponse était ailleurs ?

Certains auteurs ont émis l'hypothèse que le génome « dégénéré » de *Buchnera* n'est plus capable d'assurer la régulation de l'expression de ses gènes. C'est sans doute sous-estimer la diversité des systèmes de régulation mis en place au cours de l'évolution chez les organismes vivants. Sans remettre en cause les mécanismes de régulation transcriptionnelle mis en évidence chez

tous les organismes et tout particulièrement chez les bactéries, ce constat suggère l'existence possible de systèmes de régulation chez *Buchnera* différents des systèmes classiques qui sont connus chez les autres bactéries. Il existe déjà quelques exemples de mécanismes alternatifs moins complexes à mettre en œuvre et sans doute plus rapides que les systèmes indirects nécessitant la mise en œuvre de molécules régulatrices. Il ne s'agit pas de dresser ici une liste exhaustive de tous les mécanismes possibles de régulation mais simplement de montrer qu'il existe une grande variété de systèmes de régulation pouvant être utilisés en fonction des circonstances par les bactéries.

Le mécanisme de régulation le plus élémentaire concerne les variations du surenroulement négatif de la chromatine sur les nucléosomes. En effet, le chromosome bactérien d'*Escherichia coli* est mille fois plus long que la cellule, impliquant la formation de boucles dont les ouvertures permettant la transcription peuvent être indépendantes. Plusieurs mécanismes sont possibles et ont été observés chez *Escherichia coli*. Le surenroulement négatif de l'ADN est maintenu par les gyrases (Westerhoff *et al.*, 1988) dont l'activité dépend fortement du rapport de concentration entre ATP et ADP, rapport qui est directement impliqué dans les phénomènes métaboliques. Par ailleurs, l'acétylation des histones en relaxant le surenroulement permet l'activation de l'expression des gènes (Sheridan *et al.*, 1998). Enfin les régions riches en AT déstabilisent le surenroulement, ce qui faciliterait l'expression des gènes correspondants. Ce surenroulement joue un rôle important dans les modifications de l'expression des gènes chez *Escherichia coli* au cours de la croissance (Kusano *et al.*, 1996) ou du stress osmotique (Higgins *et al.*, 1988). Des mutations touchant les gènes codant pour les topoisomérases aboutissent d'ailleurs à des modifications du surenroulement de l'ADN qui affectent les niveaux de l'expression d'un grand nombre de protéines (Steck *et al.*, 1993).

Il est possible que ce type de mécanisme joue un rôle chez *Buchnera*. En effet, Nakabachi et Ishikawa (2000) ont montré que les cellules de *Buchnera* contiennent un quantité importante de polyamine sous forme de spermidine. Les polyamines sont des cations organiques qui interviennent dans une grande variété de processus biologiques et sont probablement impliquées dans la régulation des différents niveaux de condensation de la chromatine chez les eucaryotes (Matthews, 1993). Les polyamines semblent également intervenir dans les phénomènes de liaisons spécifiques à l'ADN et jouent sans doute un rôle important dans les phénomènes de régulation de la structure tertiaire des génomes tant chez les eucaryotes que chez les procaryotes (Feuerstein *et al.*, 1991). Or la bactérie n'a pas la capacité de synthétiser les polyamines. Elle possède seulement les gènes (*speD* et *speE*) qui permettent de convertir la putrescine en spermidine. Il semble donc que l'hôte fournisse à *Buchnera* des polyamines

sous forme de putrescine que la bactérie convertit en spermidine. Une étude a en effet mis en évidence une abondante production de S-adénosylméthionine décarboxylase (SAMDC) dans le cytoplasme des bactériocytes des pucerons jeunes (Nakabachi et Ishikawa, 2001). Cette enzyme catalyse une des étapes limitantes de la biosynthèse des polyamines chez les eucaryotes. Par ailleurs, même en l'absence d'un rôle de la spermidine, le mécanisme d'enroulement et d'ouverture de la chromatine permettant la transcription des gènes pourrait être assuré par les chaperonnes. Or la chaperonne *GroEL* est produite de façon constitutive et à un niveau élevé chez *Buchnera* (Sato et Ishikawa, 1997), (cf 1.2.232).

Si tel est le cas, la polyploïdie offrirait alors une régulation extrêmement fine à ce mécanisme de régulation. En effet, si la régulation la plus évidente semble en général l'adaptation du nombre de copies d'ARN messager dans la cellule, rien n'empêche d'imaginer que cette variation soit réalisée directement au niveau du génome avec un phénomène de polyploïdie, tout comme elle est possible au niveau du protéome par des phénomènes de régulation post-traductionnelle ou par la régulation des activités enzymatiques (Schaechter, 2001).

Par ailleurs, toujours en ce qui concerne la structure du génome, il est possible que la distance entre gènes impliqués dans une même voie métabolique le long du chromosome permette leur régulation commune en fonction du repliement du chromosome. Cette régulation ne nécessiterait donc que très peu d'opérons. Cette organisation périodique de la transcription a été mise en évidence chez la levure (Kepes, 2003) mais également chez *Escherichia coli* (Kepes, 2004).

Évidemment, ces mécanismes alternatifs de régulation, complètement spéculatifs, ne remettent pas en cause la possibilité d'une régulation à un niveau transcriptionnel. Il est tout à fait possible que des éléments de régulation n'aient pas encore été découverts dans le génome de *Buchnera* ou que les gènes de *Buchnera* ne possèdent pas les mêmes fonctions que ceux de l'actuelle *Escherichia coli*. En effet, de nombreux gènes codent pour des protéines hypothétiques dans le génome de *Buchnera*. Par ailleurs, l'exemple du gène *ftsZ* conforte l'hypothèse que certains des gènes de *Buchnera* jouent peut-être un rôle différent de celui qui a été attribué aux gènes homologues d'*Escherichia coli*. Ce gène, impliqué dans la division cellulaire est responsable de la séparation de la cellule mère en deux cellule filles chez *Escherichia coli.* Bien que ce gène présente une forte analogie chez *Buchnera*, son introduction chez des mutants d'*Escherichia coli* aboutit à la formation de cellules allongées mais pas à leur séparation (Baumann et Baumann, 1998). De plus il n'existe actuellement au-

cun doute sur le fait que de nombreuses protéines sont impliquées dans différentes activités physiologiques. Ainsi, la protéine *RecA* assure par exemple le rôle de recombinase ou de coprotéase selon les circonstances (Schaechter, 2001). Dans les mitochondries de levure, la chaperonne *Hsp60*, analogue de la protéine *GroEL* de *Buchnera,* est par ailleurs capable de se fixer sur l'ADN mitochondrial pour activer la réplication (Kaufman *et al.*, 2000). Hall *et al.* (2004) ont également montré qu'une enzyme mitochondriale impliquée dans la biosynthèse de l'arginine était capable de réguler directement l'expression des gènes dans le compartiment mitochondrial mais aussi dans le noyau de la cellule. Chez *Buchnera*, il est également possible que les pseudogènes aient un rôle dans la régulation de l'expression des gènes qui leur sont associés, comme cela a déjà été montré chez la Souris (Hirotsune *et al.*, 2003). Et même en l'absence d'éléments de régulation, il est toujours possible d'imaginer une régulation basée sur une variation non pas de la synthèse mais de la dégradation des ARN messagers.

Enfin les mécanismes de régulation pourraient être situés à l'extérieur de *Buchnera*, c'est-à-dire assurés par le puceron, comme c'est le cas pour les mitochondries ou les plastes et les cellules qui les abritent. Aucune donnée n'est actuellement disponible concernant la présence de gènes de *Buchnera* dans le génome du puceron, bien que le transfert des gènes *ilvA* et *metC* vers le noyau du puceron soit une explication possible de la capacité de *Buchnera* à synthétiser respectivement l'isoleucine et la méthionine (cf 1.2.221). De même des éléments de régulation pourraient être localisés dans le noyau du puceron. Enfin, en ce qui concerne la régulation de la synthèse des acides aminés essentiels chez *Buchnera,* aucune étude ne peut écarter la possibilité d'une régulation par le puceron de la fourniture des précurseurs via la membrane symbiosomale.

En conclusion, *Buchnera* présente toutes les caractéristiques de l'endocytobiose, à tel point qu'il est parfois difficile de savoir s'il est encore possible de parler d'individu ou s'il n'est pas préférable de parler d'organite. D'un point de vue génomique, la bactérie a en effet perdu les gènes devenus inutiles dans le milieu intracellulaire mais conservé ceux qui sont essentiels pour la survie du puceron. D'un point de vue métabolique, il existe une forte complémentarité avec son hôte. Quant à la régulation, elle reste encore une énigme à décrypter.

1.2.3 Le couple du point de vue de la problématique du métabolisme des acides aminés

Malgré l'existence d'une grande diversité de types de symbiose, Buchner (1965) est le premier a avoir proposé un rôle fonctionnel unique pour l'ensemble des symbiotes. Leur « raison d'être », comme il l'écrit lui-même, est l'approvisionnement de l'hôte en nutriments absents ou présents en quantité très limitée dans le milieu nutritionnel, et que l'hôte est incapable de synthétiser lui-même. Pour avancer cela, il s'appuie sur l'existence d'une étroite corrélation entre le régime alimentaire des insectes et la présence des symbiotes. En effet, les symbiotes sont généralement absents chez les insectes vivant sur des milieux nutritionnels complets, mais existent chez ceux dont l'alimentation est carencée. Ainsi les insectes carnivores sont dépourvus de symbiotes, alors que les insectes qui se nourrissent de sang ou de la sève des plantes en possèdent.

Le couple formé par *Buchnera* et le puceron n'échappe pas à cette règle et de nombreuses études ont révélé que leur association était essentiellement nutritionnelle. Et même si un rôle de recyclage de l'azote à partir de l'acide urique a également été avancé pour *Buchnera* (Douglas, 1998 ; Whitehead *et al.*, 1992), la fonction principale du symbiote reste la synthèse des neuf acides aminés essentiels pour le puceron. Cette fourniture endogène d'acides aminés essentiels est cruciale pour le puceron, car dans le phloème les ratios acides aminés essentiels : acides aminés non essentiels sont compris entre 1 : 3 et 1 : 10 alors que ce ratio est de 1 : 1 dans ses protéines. La nutrition des pucerons a donc fait l'objet d'études utilisant des milieux nutritifs carencés et d'analyses métaboliques avec incorporation de traceurs radioactifs. L'ensemble de ces études est basé sur la comparaison de pucerons symbiotiques et aposymbiotiques. Quelques travaux ont également été réalisés sur des préparations de bactéries isolées, mais ils ne présentent que peu d'intérêt pour l'étude du couple symbiotique. En effet, bien que ces bactéries isolées soient viables quelques heures, leur métabolisme est différent de celui des bactéries en symbiose, comme le montre l'observation des différentes classes de protéines produites (Ishikawa, 1984).

1.2.31 Des études nutritionnelles

De nombreuses études nutritionnelles ont été réalisées par l'élimination d'un acide aminé du milieu nutritif des pucerons. Ces expériences sont réalisées sur différents milieux dépourvus successivement de chacun des vingt acides aminés pour des pucerons symbiotiques et pour des pucerons aposymbiotiques.

La première étude réalisée chez le puceron *Myzus persicae* (Mittler, 1971) montre que les pucerons symbiotiques supportent toutes les carences, hormis l'absence de méthionine en raison de ses propriétés phagostimulantes. En revanche, les pucerons aposymbiotiques ont besoin de tous les acides aminés essentiels, montrant le rôle des bactéries symbiotiques dans la fourniture des acides aminés essentiels.

La croissance de larves de pucerons a ensuite été étudiée chez le puceron *Acyrthosiphon pisum* sur milieu équilibré puis déséquilibré en acides aminés essentiels. Les résultats montrent que les larves aposymbiotiques ont une croissance deux fois plus rapide sur milieu équilibré (c'est-à-dire sur un milieu dont la composition permet d'assurer une croissance normale des pucerons) alors que les larves symbiotiques ont une croissance identique sur les deux milieux (Prosser et Douglas, 1992). Cependant même sur un milieu parfaitement équilibré, les performances des pucerons aposymbiotiques et symbiotiques ne sont jamais comparables, ce qui suggère que la compensation du milieu nutritionnel ne peut pas remplacer *Buchnera* (Prosser *et al.*, 1992).

1.2.32 Des études métaboliques

Les études métaboliques sont toutes basées sur des analyses biochimiques utilisant des marqueurs radioactifs.

Une première expérience basée sur l'utilisation d'acides aminés marqués au ^{14}C a permis de comparer le devenir des acides aminés chez des pucerons symbiotiques et aposymbiotiques élevés sur milieu équilibré et déséquilibré en acides aminés. Les résultats montrent que les pucerons aposymbiotiques conservent un profil d'acides aminés libres déséquilibré, contrairement aux pucerons symbiotiques qui sont capables de reconstituer un profil « normal » (Liadouze *et al.*, 1995). Une expérience complémentaire menée sur des milieux contenant des quantités variables d'acides aminés montre que les pucerons n'adaptent pas leur assimilation en fonction du milieu (Wilkinson et Ishikawa, 1999). Ces expériences suggèrent une éventuelle régulation de l'apport en acides aminés par une productivité différentielle de *Buchnera*. Par ailleurs, la synthèse de trois acides aminés essentiels (thréonine, isoleucine et lysine) a été mise en évidence chez les pucerons symbiotiques (Febvay *et al.*, 1995) à partir du carbone des acides aminés libres non essentiels (Liadouze *et al.*, 1996).

L'étude du devenir du glutamate marqué au ^{14}C révèle que ce dernier joue un rôle primordial dans la synthèse des acides aminés. Cette constatation est d'autant plus intéressante que d'autres expériences montrent que le glutamate semble être le principal composé nutritionnel apporté à *Buchnera* (Whitehead et Douglas, 1993b ; Febvay *et al.*, 1995 ; Sasaki et Ishikawa, 1995). Des expériences ont montré qu'il pénètre dans les bactériocytes où il représente

30 % des acides aminés totaux (Whitehead *et al.*, 1992). Par ailleurs, la glutamine entre également dans les bactériocytes à l'aide d'une enzyme, la glutamyltranspeptidase (Sasaki et Ishikawa, 1993) et la réaction de désamination de la glutamine en glutamate existe dans les bactériocytes (Sasaki et Ishikawa, 1995). Le puceron fournit divers autres acides aminés non essentiels à la bactérie dont la valine, la proline et surtout l'aspartate (Whitehead et Douglas, 1993b) qui s'accumulent chez le puceron aposymbiotique (cf. **Figure A.1.5**). L'aspartate aminotransférase, une enzyme responsable de la désamination de l'asparagine en aspartate, est en effet surproduite chez les pucerons (Nakabachi et Ishikawa, 1997) et l'aspartate est converti en glutamate dans les bactériocytes (Febvay *et al.*, 1995).

Enfin, le devenir du saccharose, composé majoritaire du phloème, a également été étudié. Une expérience réalisée sur un milieu contenant du saccharose marqué au ^{14}C, montre ainsi que chez les pucerons symbiotiques la radioactivité est incorporée dans tous les acides aminés composant les protéines. La même étude réalisée sur des pucerons aposymbiotiques montre en revanche que les acides aminés essentiels ne contiennent pas de ^{14}C radioactif. Les résultats complets de cette étude indiquent que, chez les pucerons symbiotiques, le saccharose participe à la synthèse de pratiquement tous les acides aminés (sauf l'histidine et l'arginine) par l'intermédiaire des voies de la glycolyse (Febvay *et al.*, 1999).

L'ensemble de ces études tant nutritionnelles que métaboliques a permis de définir un premier modèle des voies de biosynthèse des acides aminés pour le couple formé par *Buchnera* et son hôte (cf. **Figure A.1.5**). La bactérie est capable de synthétiser les acides aminés qui sont essentiels pour son hôte mais elle a besoin de plusieurs acides aminés non essentiels et principalement de glutamate et d'aspartate qui sont probablement importés du cytoplasme de l'hôte (Zientz *et al.*, 2001).

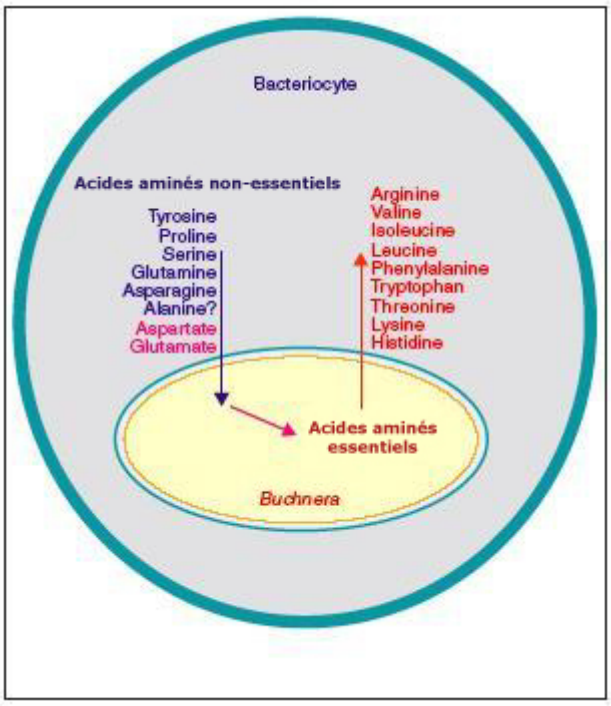


Figure A.1.5 Modèle des interactions existant entre *Buchnera* et son hôte pour le métabolisme des acides aminés. *Buchnera* est capable de synthétiser les acides aminés qui sont essentiels pour son hôte (en rouge) et semble avoir besoin de plusieurs acides aminés non essentiels (en bleu). Parmi ces acides aminés, les deux principaux sont l'aspartate et le glutamate (en rose) (d'après Zientz *et al.*, 2001).

1.2.4 Un couple ou un mariage à trois ?

Les populations naturelles de puceron du pois hébergent *Buchnera* qui en tant que symbiote primaire et obligatoire contribue directement à leur survie en leur fournissant les acides aminés essentiels absents de leur milieu nutritif. Mais les pucerons hébergent aussi fréquemment de nombreuses bactéries secondaires (Fukatsu *et al.*, 2000 ; Tsuchida *et al.*, 2002), (cf. **Figure A.1.6**). Ces γ-protéobactéries sont de cinq types : le type R (PASS), le type T (PABS), le type U (PAUS), *Rickettsia* (PAR) et *Spiroplasma* (Haynes *et al.*, 2003). Contrairement à *Buchnera*, elles peuvent être transmises de façon horizontale (Sandstrom *et al.*, 2001). Elles semblent impliquées dans des rôles spécifiques, comme celui de permettre une meilleure tolérance des pucerons aux températures élevées (Montllor *et al.*, 2002) et une résistance à certains parasitoïdes d'Hyménoptères. Elles semblent également permettre aux pucerons d'exploiter une plus large gamme de plantes hôtes (Tsuchida *et al.*, 2004).

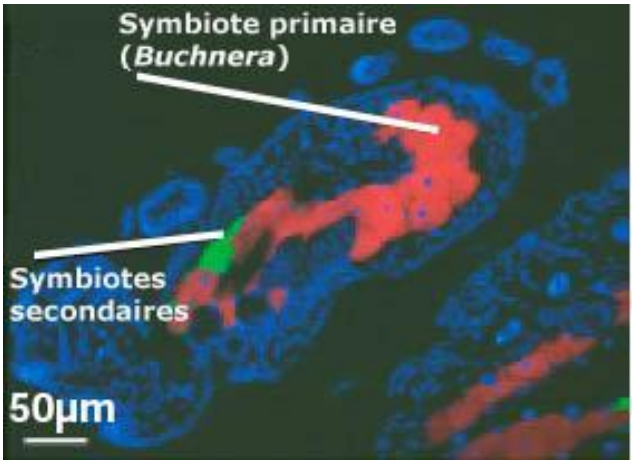


Figure A.1.6 Distribution des symbiotes primaires PASS (en rouge) et secondaires (en vert) chez un embryon de puceron du pois (image obtenue par FISH et observation au microscope à transmission), (Koga, R., communication personnelle).

Buchnera est vulnérable aux températures élevées. Lorsque les pucerons sont soumis à des températures supérieures à 30 °C, ils ne sont plus capables de se reproduire et les bactéries *Buchnera* meurent. Une infection avec des PASS améliore la reproduction des pucerons en conditions de stress thermique (Montllor *et al.*, 2002), ce qui laisse envisager que les bactéries secondaires sont capables de compenser la déficience de *Buchnera* lorsque la température est élevée. La fonction des bactéries secondaires serait alors de favoriser la survie des pucerons durant la saison d'été (Koga *et al.*, 2003).

Ces bactéries secondaires sont peut-être en compétition avec *Buchnera* tant pour les ressources que pour l'espace disponible ou, au contraire, coopèrent pour la fourniture de nutriments aux pucerons. Des interactions symbiotes-symbiotes et symbiotes-hôte existent donc au sein d'un super système symbiotique. Cependant les bases physiologiques de ces interactions n'ont été que peu étudiées jusqu'à présent. Des études ont montré que des pucerons rendus aposymbiotiques présentent une extraordinaire prolifération de micro-organismes (Nakabachi *et al.*, 2003) et qu'une infection par des symbiotes secondaires permet la survie et la reproduction des pucerons aposymbiotiques sur plusieurs générations (Koga *et al.*, 2003). Il semble en effet qu'en l'absence de *Buchnera*, les bactéries symbiotiques secondaires colonisent les bactériocytes et soient capables d'assumer le rôle des symbiotes primaires, établissant ainsi un nouveau système symbiotique. Pour d'autres auteurs, ces symbiotes secondaires pallient les fonctions déficientes de *Buchnera*, notamment dans les lignées où le génome de *Buchnera* est extrêmement réduit. Ils avancent pour preuve la grande quantité de symbiotes secondaires dans ces lignées de pucerons (Chen *et al.*, 2000).

Le puceron et *Buchnera* forment bien une entité biologique, cependant les expériences précédentes montrent que d'autres bactéries peuvent potentiellement remplacer *Buchnera*. Nul n'est irremplaçable, et pas même un symbiote obligatoire ! C'est ce qu'illustre l'exemple des pucerons appartenant au groupe des *Cerataphidini* chez lesquels les bactéries *Buchnera* ont été complètement perdues et remplacées par un symbiote ascomycète (Fukatsu et Ishikawa, 1996).

Même si certains auteurs ont suggéré que la bactérie *Buchnera* pourrait à terme être remplacée par un nouveau symbiote (Moran et Baumann, 1994 ; Von Dohlen *et al.*, 2001), l'acquisition de symbiotes secondaires ayant des rôles biologiques variés semble surtout destinée à compléter ceux que le symbiote primaire ne peut pas assurer. Il est tentant de comparer ce phénomène à celui de la duplication génique qui offre au gène dupliqué la capacité d'évoluer vers de nouvelles fonctions car les contraintes fonctionnelles restent assumées par le gène original. De la même façon, dans le super système symbiotique, les symbiotes secondaires offrent une ouverture vers de nouvelles fonctions biologiques alors que le symbiote primaire est maintenu. L'endocytobiose est une source d'innovation importante pour l'évolution (cf.1.1.33) et les symbiotes facultatifs le prouvent en apportant un ensemble de nouveaux gènes au couple initial (Margulis et Fenster, 1991).

1.3 Les objectifs de la thèse du point de vue biologique

Presque cent ans après les premières observations de Buchner, il ne fait plus aucun doute que *Buchnera* fournit à son hôte les acides aminés essentiels absents de son milieu nutritionnel. En revanche, la régulation de l'expression des gènes chez *Buchnera*, et tout particulièrement de l'expression des gènes responsables de la biosynthèse des acides aminés reste encore aujourd'hui une énigme. Certains auteurs considèrent la stabilité de l'environnement de *Buchnera* comme une raison suffisante pour expliquer l'absence de systèmes de régulation (Goebel et Gross, 2001). Mais même s'il est possible de souscrire à la vision de *Buchnera* comme organite (Andersson, 2000), il n'est pas nécessaire d'en conclure que le génome de *Buchnera* est complètement dégénéré. Les chloroplastes et les mitochondries sont en effet encore capables de réguler l'expression de leurs gènes en fonction des besoins de leur cellule hôte. Enfin, si l'association existe avec succès depuis plus de 200 millions d'années (Tamas, 2002), c'est probablement que le génome de *Buchnera* est loin d'avoir révélé toutes ses possibilités.

Le développement de la génomique a conduit à l'émergence de nouvelles technologies, comme les puces à ADN, qui permettent d'appréhender simultanément l'expression de l'ensemble des gènes d'un organisme par l'étude

multanément l'expression de l'ensemble des gènes d'un organisme par l'étude de son transcriptome (Covacci *et al.*, 1997). Les puces à ADN ont acquis leur notoriété en permettant l'identification des gènes d'organismes modèles répondant de façon extrême à des traitements extérieurs, comme l'ont montré les recherches sur le cancer. Cependant les progrès du séquençage associés à la baisse des coûts des expériences (Gibson, 2002) ont permis aux puces à ADN de faire récemment leur apparition dans tous les domaines de la biologie et notamment dans ceux de la physiologie et l'écologie (Ye *et al.*, 2001). Des puces ont ainsi été développées ces dernières années pour étudier les interactions complexes entre hôte et bactéries (Cummings et Relman, 2000 ; Kato-Maeda *et al.*, 2001).

Dans ce contexte récent, la problématique biologique de cette thèse est de lever une partie de l'énigme de la régulation des gènes chez *Buchnera*. Compte tenu du rôle de la bactérie dans l'association symbiotique avec le puceron, l'étude du métabolisme global des acides aminés par la technologie des puces à ADN semble offrir un cadre idéal pour poser cette question. Pour cela, une étude de l'expression des gènes de *Buchnera* issues de pucerons élevés sur des milieux nutritionnels contenant des quantités variables d'acides aminés a donc été réalisée. Ce travail est en quelque sorte une façon nouvelle d'aborder la symbiose. La combinaison des précédentes études nutritionnelles et de cette étude est en effet un premier pas vers la compréhension globale des interactions du système symbiotique puceron - bactérie.

2

2 Contexte méthodologique Les puces à ADN de A à Z (ou presque)

2.1 Introduction

Loin de correspondre à une technique unique, les puces à ADN regroupent un ensemble très divers de méthodes et de technologies. L'objectif de cette partie n'est donc pas de présenter une liste exhaustive des méthodologies qui sont en constante évolution, mais plutôt de mettre en évidence les étapes essentielles de la conception et de l'utilisation des puces à ADN. Leur description concerne principalement l'analyse du transcriptome, qui est au centre de notre problématique. De plus, compte tenu de l'immense quantité d'informations disponibles, certains aspects ont été volontairement limités, en raison de notre problématique biologique, aux particularités concernant les modèles bactériens. Les nombreuses étapes de la technologie des puces à ADN sont à l'origine d'une variabilité expérimentale qui peut affecter la qualité des résultats (Nadon et Shoemaker, 2002) et qu'il est essentiel de prendre en compte (Leung et Cavalieri, 2003). Ces sources de variabilité sont donc également décrites et discutées au sein des différentes paragraphes.

2.1.1 Historique

Les premières puces à ADN sont apparues en 1993, mais leur concept date de 1987 (cité dans Bellis et Casellas, 1997). Suite logique aux anciennes méthodes de *northern blotting* (Alwine *et al.*, 1977) et d'expression différentielle (Liang et Pardee, 1992), la technologie des puces à ADN est basée sur le principe d'hybridation développé par Southern (1974). Ce principe stipule que deux fragments d'acides nucléiques complémentaires peuvent s'associer et se dissocier de façon réversible sous l'action de la chaleur et de la concentration saline du milieu. Il s'agit simplement d'une miniaturisation du système classique de reverse dot blot (Lennon et Lehrach, 1991) qui a vu le jour grâce à une technologie pluridisciplinaire intégrant l'électronique (techniques de dépôt), la chimie (préparation des lames et greffes des sondes oligonucléotidiques ou syn-

thèse *in situ*), l'analyse d'images (acquisition des données) et l'informatique (interprétation des données). Depuis leur apparition, les puces à ADN suscitent un intérêt inversement proportionnel à leur taille, avec pour preuve l'explosion du nombre de publications qui leur sont dédiées depuis 2001 (cf. **Figure A.2.1**).

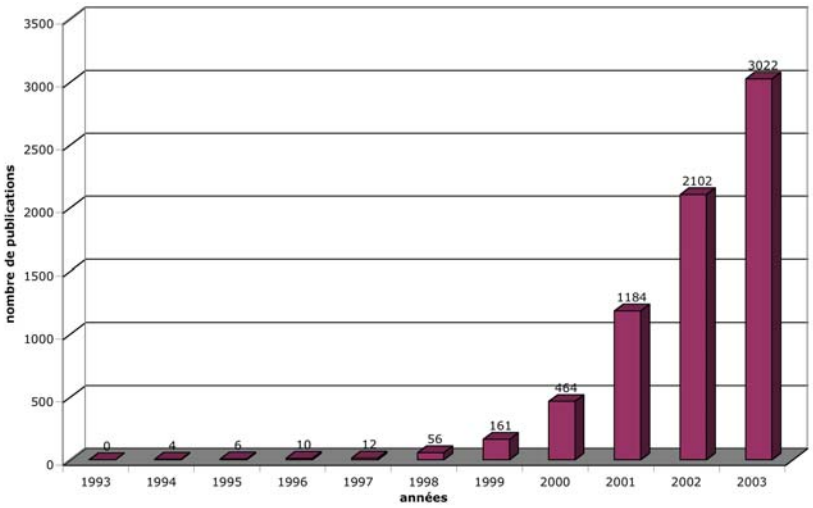


Figure A.2.1 Nombre de publications concernant les puces à ADN de 1994 à 2004 (d'après la base Entrez-PubMed[4], mots-clés : microarray* ou DNA chip*).

Historiquement les *macroarrays*, les *microarrays* et les « véritables » puces à ADN correspondent à trois méthodes différentes d'analyse (Lagoda et Regad, 2000). Les *macroarrays* utilisaient des clones d'ADN complémentaire (ADNc) disposés sur des membranes de nylon (avec un espacement de l'ordre du millimètre) en association avec des cibles radioactives. Les *microarrays*, plus miniaturisés, comportaient quelques milliers de gènes représentés par des produits PCR déposés tous les 200 à 400 microns sur une lame de verre et des cibles marquées par fluorescence. Enfin, les « véritables » puces à ADN associaient à chacun des gènes d'un organisme un ensemble d'oligonucléotides synthétisés in situ. La première de ces puces à ADN s'appelait la « *Gene ChipTM HIV PRT* ». Commercialisée en 1998 par *Affymetrix* (Santa Clara, CA, USA), elle avait été conçue pour l'analyse des mutations de la transcriptase inverse et de la protéase du virus HIV (cité dans Hinfray, 1997). La même année a vu le développement de la première puce à oligonucléotides dédiée à une bactérie, contenant un sous-ensemble de cent gènes de *Streptococcus pneumoniae* (De Saizieu *et al.*, 1998).

[4] http://www4.ncbi.nlm.nih.gov/pubmed/.

Aujourd'hui ces trois distinctions n'ont plus vraiment lieu d'être, d'autant plus que ces techniques sont utilisées de façon croisée, comme le montre l'exemple de puces à ADN utilisant des produits PCR et des cibles radioactives. Les terminologies « puce à ADN » et « *microarray* » sont donc employées de façon indifférente. Les termes de « biopuce » ou « microréseau » sont également employés dans la littérature française.

2.1.2 Principe des puces à ADN et analyse du transcriptome

L'idée conceptuelle de la puce à ADN est très simple. Il s'agit de greffer sur une surface de quelques centimètres carrés des fragments synthétiques d'ADN (les sondes) espacés de quelques micromètres et représentatifs de chacun des gènes étudiés (Ramsay, 1998 ; Rockett et Dix, 2000). Ce microdispositif est ensuite mis au contact des acides nucléiques à analyser, au cours de l'étape d'hybridation. Ces acides nucléiques, appelés cibles, correspondent aux ARNm ou aux ADNc qui ont été préalablement couplés à un marqueur fluorescent ou radioactif. Ce contact entre cibles et sondes conduit à la formation d'hybrides qualifiés par leurs coordonnées, et quantifiés grâce à la lecture des signaux radioactifs ou fluorescents.

En ce qui concerne la terminologie associée à la technique des puces à ADN, il est important de rappeler que les puces ne sont qu'un *northern blotting* inversé, où la sonde est fixe alors que la cible marquée est en solution. Cependant cette différence est à l'origine d'une confusion entre les termes « cible » (*probe*) et « sonde » (*target*). Une nomenclature a donc été recommandée (Phimister, 1999) et semble aujourd'hui bien respectée. Les sondes correspondent aux acides nucléiques fixés sur la puce, alors que les cibles représentent l'ensemble des acides nucléiques libres étudiés.

L'analyse du transcriptome nécessite de mesurer les niveaux d'expression des gènes. Cette mesure peut être réalisée par une évaluation absolue de l'expression des gènes. Ce type de mesure est réservé aux puces commercialisées par la société *Affymetrix* qui sont conçues de façon un peu particulière. Pour chaque gène, une série de dix à vingt sondes, réparties sur toute la séquence du gène, est représentée sur la lame. À chacune de ces sondes PM (*Perfect Match*) est associée une sonde MM (*MisMatch*) dont la séquence est identique à la séquence des sondes PM, excepté une mutation ponctuelle située en position centrale. Cette sonde MM permet de quantifier la part du signal aspécifique (bruit de fond) associé à la sonde PM. Le calcul du niveau d'expression d'un gène est relativement complexe, mais peut être considéré, en première approximation, comme une moyenne pondérée des différences (PM-MM) de chaque paire de sondes associées à ce gène (Chudin *et al.*, 2002).

Cependant, pour l'ensemble des puces qui sont classiquement utilisées (et qui font l'objet de cette partie), l'étude du niveau d'expression des gènes est basée sur la détermination des variations de niveau d'expression d'un organisme dans deux conditions différentes. Il s'agit donc d'une mesure différentielle de l'expression des gènes. En pratique, pour réaliser cette mesure, les cibles constituant deux échantillons d'étude sont marquées au moyen de deux fluorochromes différents. Elles sont ensuite mélangées et hybridées sur la puce (cf. **Figure A.2.2**). Après hybridation, la mesure des intensités des signaux de fluorescence relatifs à chacun des fluorochromes permet finalement de calculer pour chaque gène le rapport des intensités, qui évalue son expression différentielle.

Ces deux types de mesures ne sont cependant pas complètement disjoints. En effet, contrairement à ce que laisse penser la description de ce principe, les cibles marquées avec deux fluorochromes différents ne sont pas en compétition pour se lier sur une même sonde car les sondes sont toujours en large excès par rapport à la quantité de cibles qui est déposée. Une étude expérimentale montre même que l'intensité du signal obtenu pour un fluorochrome n'est pas affectée par la présence d'une cible marquée avec l'autre fluorochrome ('T Hoen *et al.*, 2004). Cette étude valide donc la possibilité de travailler sur les signaux aussi bien que sur les rapports des intensités et autorise la comparaison d'échantillons qui ne sont pas hybridés sur une même lame.

Basées sur ce principe, de nombreuses puces ont été développées avec des jeux de sondes spécifiques pour étudier le transcriptome d'organismes divers. Actuellement, des chercheurs essayent de concevoir une puce universelle contenant des sondes qui représentent l'ensemble des combinaisons possibles de séquences d'ADN. Le problème majeur est de définir la taille optimale des sondes. Un modèle a été réalisé pour concevoir une puce dédiée à la fois à la Souris et à la levure (Van Dam et Quake, 2002). Une première puce universelle appelée *UMAS* (*Universal Micro-Array System*) a même vu le jour. Elle contient pour sondes toutes les combinaisons possibles d'hexamères et son utilisation combinée avec une étape enzymatique de fractionnement des cibles permettrait de générer des profils d'expression pour n'importe quel organisme (Roth *et al.*, 2004).

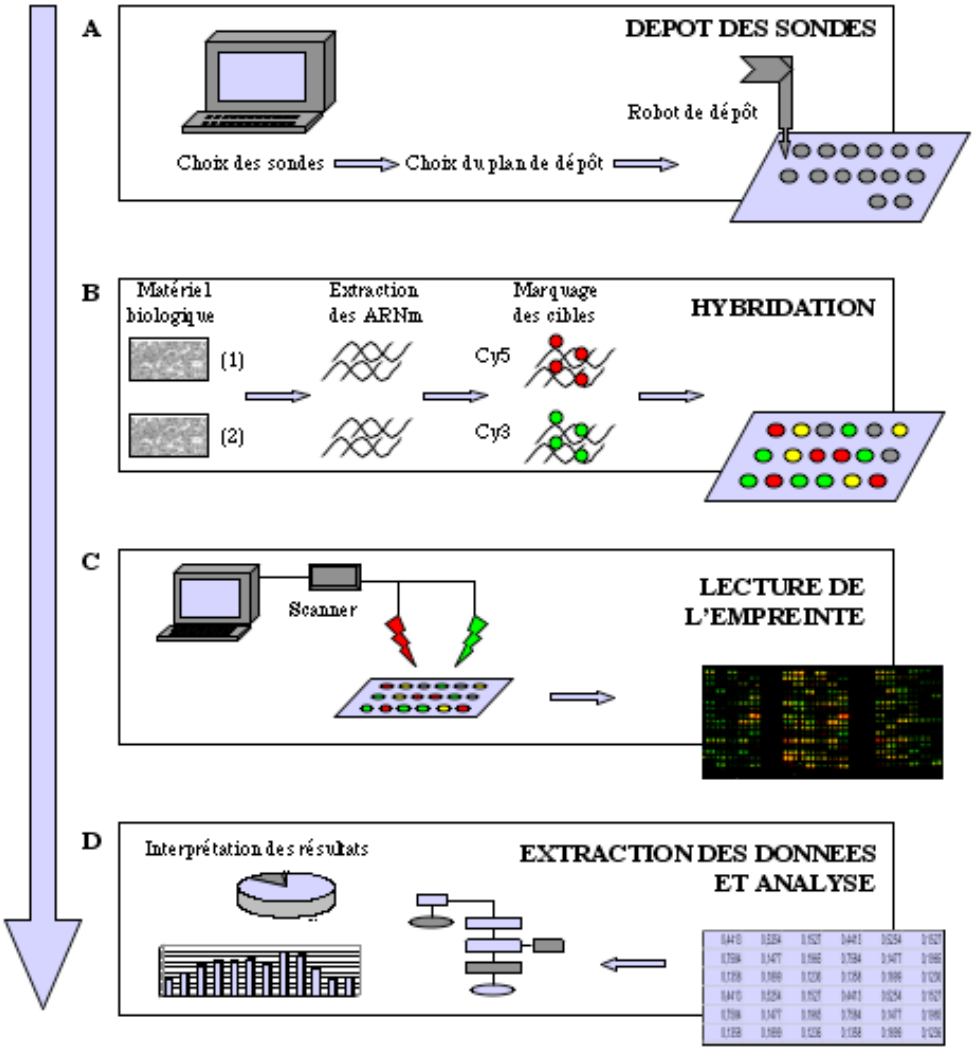


Figure A.2.2 Principe de la technologie des à puces à ADN. (A) Les séquences des sondes sont déterminées de façon à optimiser leur spécificité et leur sensibilité. Les sondes synthétisées sont déposées par un robot sur la surface de la lame selon un plan défini. (B) Les ARNm sont extraits des échantillons biologiques à comparer, marqués avec deux fluorochromes différents puis mélangés avant hybridation. (C) La lecture des lames est réalisée avec un scanner (microscope à fluorescence) couplé à un photomultiplicateur (PMT). (D) L'image est alors analysée de façon à quantifier le signal. Les données sont ensuite normalisées, analysées et interprétées.

2.1.3 Des applications variées pour les puces

2.1.31 Des puces à ADN

Parallèlement à l'analyse des profils d'expression, les puces à ADN offrent la possibilité de réaliser des études très diverses.

2.1.311 Criblage de mutations

Le criblage de mutations, appelé aussi *SNP* pour *Single Nucleotide Polymorphism*, repose sur la conception de puces à ADN capables d'analyser chaque base d'une séquence connue. Pour cela les sondes sont organisées en ensembles au sein desquels une séquence est strictement homologue à celle du type sauvage alors que les autres sont caractérisées par une substitution de base toujours localisée au milieu de la séquence (Ahrendt *et al.*, 1999). Par exemple, pour étudier une substitution de base, les sondes sont organisées en tétrades dans lesquelles une des quatre sondes possède en position centrale la base homologue à celle de la séquence sauvage, alors que trois autres sondes contiennent les trois autres bases possibles.

2.1.312 Séquençage par hybridation

Une autre application des puces à ADN a été développée pour le séquençage par hybridation. Son utilisation permet d'accélérer considérablement le procédé. Le gène étudié est en effet considéré comme un ensemble de plusieurs séquences chevauchantes dont la détermination simultanée puis l'assemblage permettent de reconstituer la séquence (Hacia, 1999). La méthode est également utilisée pour la vérification de génomes déjà séquencés par la technique de Sanger.

2.1.313 Étude phylogénétique

Les puces peuvent également être utilisées dans les études phylogénétiques et par exemple pour déterminer les distances génétiques qui existent entre bactéries (Cho et Tiedje, 2001). La comparaison de plusieurs expériences de puces montre qu'il est également possible d'identifier les niveaux d'expression et les régulations de gènes orthologues chez différents organismes (Jimenez *et al.*, 2002).

Une extension logique de ce type d'étude est la recherche de certaines fonctions chez des organismes dont le génome n'est pas connu. Une puce dédiée à *Escherichia coli* a par exemple été utilisée pour obtenir des informations sur les gènes (et donc les fonctions associées) présents chez le symbiote de la mouche tsé-tsé, très proche phylogénétiquement d'*Escherichia coli* (Akman et Aksoy, 2001).

2.1.314 Applications industrielles

Les puces trouvent également de nombreuses applications dans l'industrie. Des puces conçues pour des cibles d'ARN ribosomiaux sont ainsi utilisées pour identifier les souches pathogènes dans les hôpitaux. De même, elles sont utilisées dans le secteur agro-alimentaire pour le contrôle qualité (contrôle des micro-organismes utilisés dans certaines fabrications, détection des séquences provenant d'organismes génétiquement modifiés) et pour l'environnement (détection des agents infectieux dans l'alimentation, l'air ou l'eau), (Guschin *et al.*, 1997).

2.1.315 Caractérisation physique du génome

Deux techniques principales ont été développées pour étudier la cartographie et les modifications du génome. La première est la méthode *ChIP on chip* basée sur une technique d'immuno-précipitation de la chromatine (*ChIP*) sur une puce contenant des régions intergéniques (*on chip*). Ce type de puce permet de déterminer les sites de fixation d'une protéine sur l'ADN (Iyer *et al.*, 2001). L'ADN sur lequel sont fixées des protéines (essentiellement des facteurs de transcription) est tout d'abord extrait, puis les protéines sont précipitées avec des anticorps spécifiques. L'ADN est ensuite récupéré, marqué et hybridé sur la puce (Lo *et al.*, 2001 ; Robyr *et al.*, 2002 ; Shannon et Rao, 2002).

La seconde méthode développée est appelée *CGH* (*comparative genomic hybridization*). Ce type de puce est destiné à la comparaison des profils d'hybridation génomique dans différentes conditions et permet de repérer d'éventuelles duplications ou suppressions de séquences (Kallioniemi *et al.*, 1992 ; Pollack *et al.*, 1999).

D'autres types de puces sont également conçues pour étudier les modifications post-transcriptionnelles de l'ADN comme l'acétylation, la méthylation, la phosphorylation, et l'ubiquitination (Pollack et Iyer, 2002).

2.1.32 Et même des puces sans ADN !

Bien que les puces soient essentiellement utilisées avec des sondes d'ADN, leur succès a ouvert la voie à de nouvelles applications. Aujourd'hui la technologie des puces est utilisée pour obtenir une vision du protéome (puce à protéines) et même accéder aux phénotypes cellulaires (puces à cellules).

Étude du protéome

L'analyse du protéome n'a pas échappé au développement des puces (De Wildt *et al.*, 2000 ; Tyers et Mann, 2003). Pour observer directement et à grande échelle la présence des protéines et leur activité, des puces à protéines ont donc été conçues (Macbeath, 2002). Ces puces permettent de déterminer directement les concentrations en protéines et leur conformation spatiale, qui

dans certains cas correspond à leur degré d'activité. Pour cela, des anticorps spécifiques sont fixés dans des puits sur les puces. Ils permettent de piéger les protéines présentes dans une certaine conformation. Ainsi, il est théoriquement possible de comparer à grande échelle les concentrations de différentes protéines, et de séparer les formes actives et inactives de ces protéines. La principale difficulté de cette technologie réside dans la production d'anticorps spécifiques de chacune des protéines cellulaires (Zhu *et al.*, 2000).

Étude des phénotypes

Récemment la famille des puces s'est agrandie avec le développement des puces à cellules appelées aussi puces à transfection. Les cellules sont piégées sur une puce à l'aide d'anticorps spécifiques. Un ensemble de microcapillaires permet ensuite de véhiculer, via un liquide, les gènes utilisés pour la transfection simultanée des lignées cellulaires présentes sur la puce. Une fois le gène intégré dans le génome, la production de la protéine correspondante est responsable d'un phénotype cellulaire directement observable. Pour cela, un gène rapporteur est inséré dans le génome des cellules étudiées. Une première expérience a été réalisée avec des ADNc (Ziauddin et Sabatini, 2001), et plus récemment une autre étude a été menée en utilisant des ARN interférents (*ARN-si* pour *small interfering*), (Baghdoyan *et al.*, 2004).

2.2 Conception des puces à ADN

En dépit d'un principe relativement simple, il existe de nombreuses possibilités de conception pour les puces à ADN (Freeman *et al.*, 2000). L'objectif de cette partie n'est pas de dresser un inventaire technique de l'ensemble des possibilités existantes mais plutôt de présenter certains des aspects parmi les plus essentiels.

La fabrication de la lame sur laquelle sont déposées les sondes combine des problèmes de chimie (traitement de la surface de la lame et fixation des sondes) et de mécanique (dépôt des sondes sur la lame par un robot). Des disfonctionnements sur ces différentes étapes peuvent conduire à des irrégularités spatiales sur la lame qu'il est nécessaire de rechercher afin d'éliminer les données correspondantes lors de l'analyse.

2.2.1 Le support d'hybridation

Certaines méthodes d'analyse utilisaient des sondes fixées sur un support bien avant l'apparition des puces à ADN, notamment dans le domaine des tests de diagnostic. Cependant, ces supports ne permettent l'immobilisation que d'un faible nombre de sondes, 96 pour les plaques de microtitration en plastique

et une cinquantaine pour les bandelettes de nitrocellulose. Pour les puces, il était donc nécessaire de disposer de supports de plus grande capacité. Par ailleurs, le choix d'un support est conditionné par ses propriétés physico-chimiques, ou plus précisément, par l'adéquation de celles-ci avec les conditions dans lesquelles la puce est placée lors de la fixation des sondes et lors de son utilisation. Il est donc nécessaire, avant d'envisager l'emploi d'un support particulier, de considérer des caractéristiques comme la stabilité au pH, la résistance physique ou la stabilité chimique ainsi que la capacité à fixer de façon non spécifique les acides nucléiques (Matson *et al.*, 1994). Une fois le support choisi, sa surface est rendue réactive vis-à-vis de molécules « espaçantes » au cours d'une première étape dite de fonctionnalisation. Ces molécules de taille variable sont ensuite utilisées comme intermédiaires entre la surface et la sonde. Elles permettent de s'affranchir des propriétés de surface des supports, qui s'avèrent souvent gênantes pour la fixation des sondes comme pour l'hybridation.

2.2.11 Les supports dits neutres

Le rôle des supports neutres est uniquement d'assurer la fixation optimale des sondes. Le verre est le support classiquement utilisé, principalement pour son faible coût, son inertie et sa transparence qui facilite la détection par fluorescence. Son utilisation nécessite une activation préalable de la surface par fixation de dérivés du silane possédant des groupements actifs (comme des aldéhydes) capables de générer des liaisons covalentes avec des nucléosides modifiés (par ajout d'amine par exemple) situés à l'extrémité des sondes. D'autres types de traitement de la surface peuvent être réalisés. Par exemple, la technique dite d'impression d'ADN développée par Derisi *et al.* (1996) est basée sur l'utilisation de lame de verre traitée à la poly-L-lysine pour former une surface adhésive. Cependant, le verre présente deux inconvénients. D'une part, la densité des plots reste faible (Hoheisel et Vingron, 2000) et d'autre part, la fixation des sondes fait généralement intervenir des interactions de charges non covalentes qui ne permettent pas de créer une liaison irréversible entre la surface et la sonde. Les sondes sont donc partiellement éliminées de la surface lors de l'étape de lavage, ce qui peut aboutir à une perte de sensibilité (Rogers *et al.*, 1999 ; Schena *et al.*, 1996). Pour pallier ce problème, des équipes et des entreprises de biotechnologies ont développé des lames de verre dont la surface est recouverte de couches de polymères permettant la fixation covalente des sondes. Le laboratoire IFOS de l'Ecole Centrale de Lyon a par exemple mis au point des puces très résistantes aux lavages en utilisant une monocouche d'un polymère d'aminopropyl triéthoxysilane (ATPS) (Dugas *et al.*, 2004).

D'autres supports comme les gels de polyacrylamide peuvent également être employés. Ce type de support est utilisé dans les puces appelées *MAGIChip* (*MicroArrays of Gel Immobilised Compounds on a Chip*) (Guschin *et al.*, 1997 ; Yershov *et al.*, 1996). Le polyacrylamide est un support stable qui ne génère qu'un faible bruit de fond pour la détection par fluorescence. De plus, la méthode permet de multiplier par cent la densité de sondes. Sa seule limite reste l'accessibilité des sondes situées à l'intérieur du gel (Proudnikov *et al.*, 1998). Des membranes de nylon (Bertucci *et al.*, 1999 ; Cao *et al.*, 2002) ou encore le polypropylène sont également utilisés comme support.

2.2.12 Les supports dits actifs

Comme les supports neutres, les supports actifs permettent la fixation des sondes et l'hybridation des séquences cibles, mais ils interviennent également dans l'acquisition du signal. Ils ne facilitent pas seulement la transduction par leurs propriétés physiques, comme le verre avec sa transparence, mais jouent un véritable rôle de vecteur du signal. Les supports les plus souvent utilisés sont le silicium, les électrodes de carbone, les fibres optiques (Lee et Walt, 2000) et l'or qui, en tant que métal conducteur et inoxydable, peut intervenir dans un mode de transduction électrochimique ou optique (Beattie *et al.*, 1993).

2.2.2 Les sondes

2.2.21 Différents type de sondes

Les sondes greffées ou synthétisées peuvent être de différentes tailles en fonction de la problématique biologique, des contraintes expérimentales et des moyens disponibles. La détermination de leur séquence nécessite une analyse bioinformatique de façon à optimiser à la fois les paramètres de spécificité (unicité des sondes dans le génome) et les paramètres thermodynamiques (stabilité de l'hybride formé) dont l'ensemble des aspects est discuté dans le chapitre suivant.

Le criblage de mutation ou génotypage nécessite l'utilisation de sondes courtes (15-20 pb) (Lindblad-Toh *et al.*, 2000a ; Lindblad-Toh *et al.*, 2000b). Elles se caractérisent par une spécificité très forte (une seule base de différence entre la cible et la sonde suffit, en théorie, à interdire l'hybridation) et une sensibilité faible. Un des verrous technologiques de ces puces reste la possibilité d'effectuer des mesures quantitatives. La technologie *Affymetrix* utilise également des sondes courtes pour l'analyse de l'expression de gènes mais en multipliant le nombre de sondes par gène (cf. 2.1.2).

L'analyse du transcriptome est basée sur l'utilisation de deux types de sondes. Les premières sont les sondes oligonucléotidiques moyennes (30-70 pb) qui allient à la fois des qualités de spécificité et de sensibilité. Les secondes sont des fragments d'ADNc entiers issus d'une amplification par PCR (100 à 500 pb). Ces sondes offrent une sensibilité maximale, mais possèdent, en théorie, une spécificité plus faible que les précédentes.

Il existe cependant peu de travaux expérimentaux comparant les sondes d'ADNc et les oligonucléotides (Holloway *et al.*, 2002). D'un point de vue purement technique, les solutions de sondes d'ADNc présentent une viscosité élevée qui peut compliquer le procédé de fabrication des puces à ADN (Tomiuk et Hofman, 2001). Par ailleurs, il a été démontré sur les puces contenant des sondes d'ADNc que seules 10 à 20 % des sondes sont disponibles pour l'hybridation contre 20 à 50 % pour les puces à oligonucléotides ('T Hoen *et al.*, 2004). Des phénomènes de compétition entre cibles peuvent donc exister au sein des plots pour lesquels les transcrits sont en concentration élevée (Stillman et Tonkinson, 2001), et ceci d'autant plus que les plots de sondes d'ADNc présentent des variations de concentration plus importantes que les plots de sondes oligonucléotidiques (Hessner *et al.*, 2003).

D'un point de vue qualitatif, aucune différence de sensibilité n'a été observée lors de la comparaison de l'utilisation de sondes oligonucléotidiques (50 mers) et de sondes PCR (Kane *et al.*, 2000). Cependant, en ce qui concerne la spécificité du signal, l'utilisation de sondes oligonucléotidiques présente plusieurs avantages. Le premier concerne la possibilité de choisir une région spécifique pour chacun des transcrits et dépourvue de séquences répétées favorisant les hybridations aspécifiques (Gerhold *et al.*, 1999 ; Lipshutz *et al.*, 1999). De plus, chez les eucaryotes supérieurs beaucoup de gènes possèdent des variants en fonction de l'épissage alternatif, et la possibilité de détecter spécifiquement ces transcrits avec des sondes oligonucléotidiques est d'une grande utilité (Kane *et al.*, 2000). Un second avantage concerne la possibilité de choisir des sondes oligonucléotidiques présentant des caractéristiques similaires et optimales. Cela permet d'obtenir un comportement d'hybridation comparable pour l'ensemble des cibles étudiées, contrairement aux sondes d'ADNc, qui sont fréquemment de taille variable, ce qui peut impliquer des différences dans les cinétiques d'hybridation (Tomiuk et Hofman, 2001).

2.2.22 Les sondes de contrôle

Compte tenu de la complexité des protocoles expérimentaux mis en œuvre pour les puces à ADN, il est indispensable d'ajouter des sondes dites de contrôle aux sondes représentant les gènes d'intérêt. Différents types de témoins peuvent être déposés sur les puces à ADN. Les premiers sont les témoins

négatifs qui peuvent correspondre à des plots contenant du tampon ou à des sondes qui ne sont pas capables de s'hybrider avec les cibles de l'échantillon. Ils permettent d'estimer le bruit de fond sur la puce (Selinger *et al.*, 2000). Des témoins d'hybridation aspécifique peuvent être ajoutés. Ils correspondent à des sondes dont les séquences diffèrent de celles des gènes étudiés pour seulement une ou plusieurs bases situées en position centrale. Les hybridations observées sur l'un de ces témoins témoignent de phénomènes d'hybridation non spécifique dont il est nécessaire de tenir compte pour l'analyse des données (Wodicka *et al.*, 1997). Enfin, il est possible d'employer des témoins positifs qui pourront être utilisés au cours de l'étape de normalisation. Leurs séquences correspondent à des séquences provenant d'un autre organisme et qui sont absentes dans le génome étudié (Schena *et al.*, 1996). Les transcrits, correspondant à ces témoins positifs, sont synthétisés *in vitro* puis ajoutés à des concentrations différentes et parfaitement connues aux transcrits de l'organisme étudié juste avant la réaction de marquage.

2.2.3 Le plan de dépôt

La définition du plan de dépôt des sondes constitue un des aspects critiques de la conception des puces à ADN. La répartition des plots sur la lame peut en effet avoir un impact important sur l'étape d'analyse des résultats (Yang *et al.*, 2002b). Il est donc essentiel d'accorder une grande importance au choix du plan de dépôt des sondes. Dans l'idéal, chaque lame d'un plan expérimental devrait posséder un plan unique et aléatoire, compte tenu des biais potentiels qui peuvent être induits par les arrangements réguliers (Fisher, 1951). Cependant, les contraintes liées à l'utilisation de robots et la logistique de reconnaissance des plots rendent impossible ce type de distribution aléatoire. Bien qu'il soit nécessaire de ne pas oublier cette réalité, il est tout de même possible d'optimiser le plan de dépôt des sondes. Pour cela, le dépôt d'une même sonde sur la puce doit être répété, la meilleure solution étant évidemment de disperser ces plots répétés sur la lame. Cela permet de limiter les problèmes causés par les lavages ou les rayures sur la lame et offre la possibilité de moyenner les données pour un gène (Lee *et al.*, 2000).

2.2.4 L'adressage des sondes

L'adressage, pour reprendre une terminologie utilisée par les informaticiens, est défini comme la capacité à fixer de façon précise une molécule sur un site donné. Deux approches différentes existent pour la technologie des puces à ADN. Les sondes sont soit fixées après formation, soit synthétisées directement sur le support (synthèse *in situ*). Le choix de la méthode utilisée dépend

des contraintes d'utilisation de la puce notamment en termes de diversité, de densité et de pureté des sondes. Dans les deux cas cependant, il est essentiel de déposer les sondes en excès par rapport à la concentration en cibles utilisée. Il a en effet été montré que le signal d'hybridation n'est corrélé de façon linéaire avec la quantité d'ARNm que pour une concentration en sondes au moins dix fois plus importante que celle en cibles (Lemieux *et al.*, 1998).

2.2.41 La fixation de sondes synthétisées

Cette méthode d'adressage est utilisée pour les oligonucléotides de synthèse ou des fragments d'ADNc. Elle peut s'effectuer à l'aide de différentes techniques. La méthode la plus classiquement utilisée est une méthode mécanique. Les sondes sont fixées sur le support par un robot de dépôt dont la tête est formée de plusieurs micropipettes appelées aiguilles de dépôt (cf. **Figure A.2.3**). Cependant d'autres techniques, comme l'électrochimie sont également employées. Dans ce cas, la puce est composée d'un support en silicium où chaque plot est recouvert d'une mini-électrode en or. Chaque sonde rejoint alors un plot particulier lorsque les mini-électrodes sont mises successivement sous tension (Livache *et al.*, 1994).

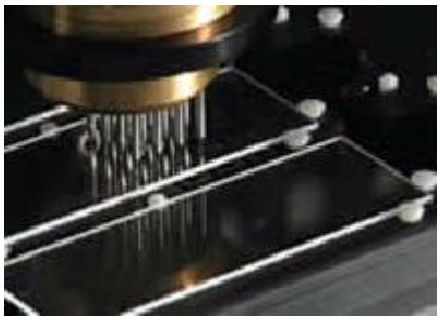


Figure A.2.3 Vue rapprochée des aiguilles d'un robot de dépôt.

2.2.42 La synthèse in situ

La synthèse *in situ* est réalisée par deux types de technologies. La première est la photolithographie développée par la société *Affymetrix*. Elle procède par dépôts de couches successives des quatre nucléotides sur un support en verre. Un masque, dont la configuration varie pour chaque couche déposée, assure une succession correcte des bases. Ce procédé extrêmement flexible ne permet cependant pas d'analyser et de corriger les éventuelles erreurs de synthèse (Mcgall *et al.*, 1996). La surface de verre est recouverte d'un polymère qui possède des groupements aminés protégés par un groupement photolabile. La photodéprotection (lyse par la lumière) de ces groupements est réalisée à travers un masque photolithographique qui permet d'exposer sélectivement les groupes fonctionnels. Durant la synthèse des oligonucléotides, des déoxynu-

cléosides possèdant un groupe protecteur photolabile à l'extrémité 5' sont ajoutés. Ils réagissent uniquement au niveau des sites qui ont été exposés à la lumière lors de l'étape précédente (Heller, 2002). Ce processus répété avec différents masques permet de contrôler l'endroit et l'ordre d'addition des nucléotides. Cette stratégie combinatoire permet de former un grand nombre de composés en un nombre réduit d'étapes (Jacobs et Fodor, 1994). La seconde technique mise en œuvre est l'impression qui est une adaptation du procédé utilisé par les imprimantes à jet d'encre. Elle repose sur la propulsion de très petites sphères de fluide dont le volume est inférieur au nanolitre (Relogio *et al.*, 2002).

2.3 Hybridation des puces à ADN

2.3.1 Préparation du matériel biologique

La préparation des échantillons est une étape critique qui est souvent sous-estimée. Pourtant, les choix effectués au cours de cette étape et la mise en œuvre de protocoles complexes sont responsables d'une variabilité à la fois biologique et expérimentale qui est rarement discutée dans les publications.

2.3.11 Choix et préparation des échantillons

Le matériel biologique est bien souvent un matériel hétérogène. Une tumeur par exemple est constituée de nombreuses cellules très différentes et le prélèvement des cellules malades s'accompagne souvent de celui de cellules saines. De même, une population bactérienne se compose de cellules dans des états physiologiques très variables. Malheureusement, travailler sur une cellule unique, en plus de la miniaturisation nécessaire, ne constitue pas une meilleure solution, car les réponses individuelles des différentes cellules sont également très variables (Blake *et al.*, 2003). L'idéal serait donc de travailler sur une population de cellules homogènes et synchronisées, ce qui est en général loin d'être le cas.

Par ailleurs, une quantité de transcrits minimum est nécessaire pour optimiser les hybridations sur puces à ADN. Cependant il est parfois difficile d'obtenir une quantité suffisante de matériel biologique au départ (Hautefort et Hinton, 2000). Ainsi, travailler sur une population de cellules homogènes n'est parfois possible qu'au prix d'une réduction drastique de la taille des échantillons. L'étude de certains modèles spécifiques (bactéries non cultivables ou prélèvements par microchirurgie) impose également des échantillons de taille réduite. Dans ce cas, il est alors nécessaire de réaliser une étape d'amplification

des ARNm (Wang *et al.*, 2000). Cette étape peut induire des biais très importants liés à des phénomènes d'amplifications sélectives. Récemment, de nouvelles techniques d'amplification ont été développées afin de maintenir au mieux les rapports entre cibles amplifiées et non amplifiées (Iscove *et al.*, 2002) pour conserver la possibilité d'effectuer des analyses quantitatives (Petalidis *et al.*, 2003).

Enfin, les conditions d'extraction des échantillons se révèlent particulièrement importantes. En effet, l'ARN messager est une molécule relativement instable et les demi-vies des différents transcrits d'une cellule eucaryote se révèlent extrêmement variables. Quant aux procaryotes, l'exemple des bactéries à croissance rapide comme *Escherichia coli* montre que le taux de renouvellement des transcrits est très élevé (Gingeras et Rosenow, 2000). Dans tous les cas, les profils d'expression obtenus ne représentent donc qu'une moyenne des niveaux d'expression et l'étape d'extraction se révèle critique. En effet, en plus de l'instabilité des transcrits, le stress de l'extraction conduit à des perturbations comme la synthèse de protéines de choc thermique ou l'activation de lipopolysaccharides, qui imposent une modification rapide des profils d'expression. Il est donc essentiel de prélever puis de congeler l'échantillon le plus rapidement possible (Watson *et al.*, 1998), ce qui nécessite la mise au point d'approches innovantes pour la préparation des échantillons. Cependant, même avec un protocole d'extraction standardisée et reproductible, il existe toujours une part de variabilité expérimentale qui doit être prise en compte.

2.3.12 Choix des cibles de référence et de contrôle

La technique des puces à ADN avec marquage par fluorescence est basée sur une mesure relative de l'expression des gènes entre deux conditions. Cette absence de mesure absolue de l'expression des gènes impose de définir une référence destinée à détecter et confirmer les gènes exprimés de façon différentielle. Il existe pour cela deux possibilités.

La première possibilité est de connaître un ensemble de gènes invariants à l'intérieur même de l'échantillon étudié. Pour cela, les « gènes de ménage » dont l'expression est constitutive dans les conditions d'étude, ont été fréquemment utilisés. Malheureusement, pour de nouvelles conditions expérimentales ou sur des modèles encore peu étudiés, leur expression constitutive est difficile à supposer *a priori* (Yang *et al.*, 2002b). De plus des études décrivent des variations parfois importantes de l'expression de ces gènes dits pourtant « de ménage ». Une étude menée au cours de la phase stationnaire chez la levure et en réponse au stress thermique chez l'Homme montre que le niveau d'expression de la plupart des gènes change au cours de l'expérience. De plus, les résultats indiquent que même une faible variation globale a un effet signifi-

catif sur le nombre de gènes détectés comme différentiellement exprimés (Van De Peppel *et al.*, 2003). Pour éviter ce type de problème Eickhoff *et al.* (1999) préconisent plutôt l'utilisation de témoins externes. Il peut s'agir de transcrits produits *in vitro* (cf. 2.2.22) et ajoutés en quantités connues aux cibles d'intérêt (Selinger *et al.*, 2000) ou de cibles de contrôle commerciales qui restent en grande majorité adaptées à l'utilisation d'échantillons d'origine eucaryote.

Une autre possibilité est de remplacer l'un des deux échantillons étudiés sur la puce par un échantillon de référence, l'idée étant de trouver une référence universelle. La première référence qui a été utilisée est l'ADN génomique qui est par définition complet et universellement disponible. D'autres références ont été mises au point pour quelques modèles. Il s'agit de mélanges commerciaux d'ARN total qui sont représentatifs de l'échantillon étudié. Ces deux méthodes testées chez l'Homme semblent donner des résultats équivalents (Weil *et al.*, 2002). Une autre étude plus récente nuance cette observation et montre que pour les gènes faiblement exprimés les meilleurs résultats sont obtenus avec une référence d'ADN génomique (Williams *et al.*, 2004). Actuellement, le choix de l'une ou l'autre de ces références dépend plus des moyens disponibles et du modèle étudié que de leur performance. En effet, chez l'Homme, de nombreux laboratoires utilisent actuellement les mélanges commerciaux d'ARN, et ce malgré leur coût. En revanche pour les autres organismes, l'ADN génomique reste une référence de choix universelle et complète. Elle a par exemple été utilisée avec succès chez la bactérie *Mycobacterium tuberculosis* (Talaat *et al.*, 2002).

2.3.2 Marquage

Le marquage des cibles met en jeu des protocoles divers en fonction des caractéristiques de l'organisme étudié. Les protocoles courants de marquage des échantillons issus de cellules eucaryotes utilisent la séquence poly-A présente à l'extrémité 3' des ARNm pour ancrer une amorce poly-T permettant la synthèse des ADNc marqués par une transcriptase inverse. Chez les procaryotes en revanche, bien que cette séquence soit présente sur de nombreux ARNm, elle intervient comme un signal de dégradation de la molécule et possède donc une courte durée de vie (Gingeras et Rosenow, 2000). De nombreuses autres méthodes de marquage spécifique ont été mises au point. Seules trois d'entre elles, représentatives de cette diversité, sont décrites dans cette partie pour montrer qu'il est essentiel de prendre en compte leurs caractéristiques au cours de l'analyse des résultats. La première qui est également la plus utilisée pour l'analyse du transcriptome met en jeu une synthèse d'ADNc à partir d'un ensemble d'amorces aléatoires ou spécifiques. La seconde correspond à un mar-

quage de l'ARN total. Ces deux approches ne permettent cependant pas un marquage spécifique des ARNm. Dans le premier cas, la synthèse des ADNc peut varier en fonction des gènes et dans le second cas, le marquage concerne également des ARN ribosomiaux et des ARN de transfert qui représentent de 97 à 99 % des ARN chez les bactéries. Pour ces deux raisons, une troisième technique a été mise au point par *Affymetrix*. Elle permet le marquage direct et spécifique des ARNm.

2.3.21 Marquage par synthèse d'ADNc

La synthèse d'ADNc est réalisée au cours d'une étape de transcription inverse qui peut être couplée à une amplification par PCR (RT-PCR) lorsque la quantité d'échantillon est insuffisante (cf. 2.3.11). L'enzyme utilisée est très sensible à de nombreux paramètres comme la composition en bases de la molécule d'ARNm, la température ou encore la concentration saline du milieu. Cette sensibilité peut générer un marquage différentiel de chacun des gènes qui est généralement limité par l'optimisation des protocoles ('T Hoen *et al.*, 2003). Ainsi, une étude menée chez *Mycobacterium tuberculosis* montre que l'utilisation d'amorces spécifiques à la place d'amorces aléatoires permet une meilleure sensibilité et une plus grande spécificité de synthèse (Talaat *et al.*, 2000). Compte tenu du nombre d'ADNc à synthétiser simultanément, des logiciels ont été développés pour minimiser le nombre d'amorces spécifiques nécessaires (Fernandes et Skiena, 2002 ; Talaat *et al.*, 2000).

L'étape de marquage à proprement parler peut avoir lieu au cours de la synthèse d'ADNc ou au cours d'une seconde étape. Elle est effectuée par incorporation d'une molécule radioactive ou fluorescente. Les cyanines fluorescentes Cy3 (vert) et Cy5 (rouge), mais aussi la fluoresceïne et la rhodamine, sont le plus souvent utilisées car elles possèdent un haut niveau d'émission photonique sous forme d'un pic étroit qui assure une meilleure sensibilité et résistent à la décoloration. Cependant, le Cy5 est parfois à l'origine d'un bruit de fond élevé sur les surfaces en verre et semble plus sensible à la décoloration que le Cy3 (Van Hal *et al.*, 2000). Lorsque le marquage a lieu au cours de la synthèse d'ADNc, il s'agit d'un protocole de marquage direct qui met en jeu un nucléotide portant la molécule fluorescente. Par opposition, le marquage indirect se déroule en deux étapes. Un nucléotide modifié est incorporé dans les ADNc au cours de la synthèse puis les ADNc sont mis en contact avec les molécules fluorescentes qui se fixent de façon spécifique aux nucléotides modifiés (cf. **Figure A.2.4**).

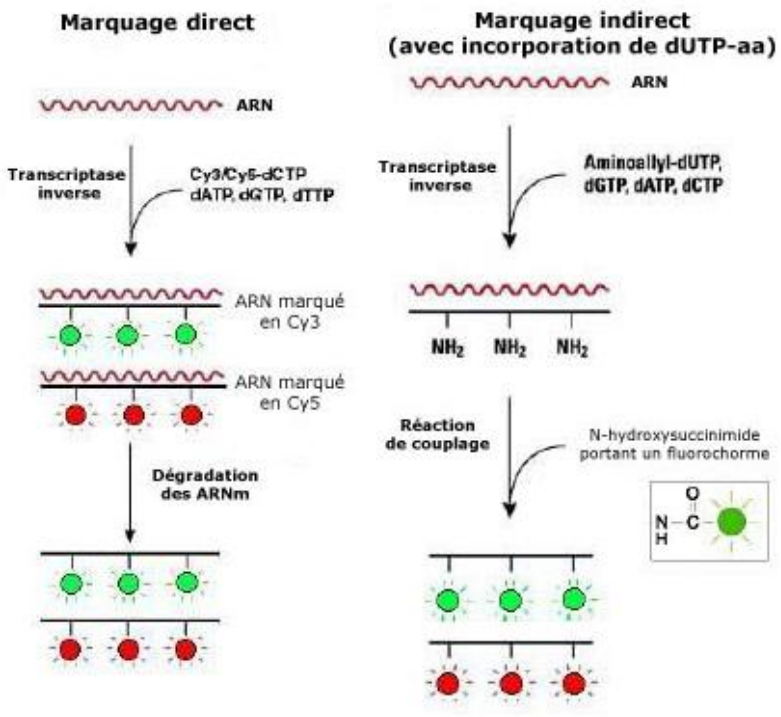


Figure A.2.4 Description des protocoles de marquage direct (à gauche) et indirect (à droite) pour l'incorporation des fluorochromes Cy3 et Cy5 (d'après Yu *et al.*, 2002).

2.3.22 Marquage direct de l'ARN total

Le marquage de l'ARN total permet de limiter la dégradation des ARNm de l'échantillon. Une étude comparative a ainsi montré que cette dégradation était quatre fois moins importante que lors d'une synthèse d'ADNc, principalement en raison de la réduction du nombre de manipulations nécessaires (Mahadevappa et Warrington, 1999). Cette méthode permet donc de travailler avec une quantité initiale de matériel biologique plus faible, ce qui s'avère particulièrement intéressant pour des échantillons de taille limitée (Gupta *et al.*, 2003). Par ailleurs cette méthode permet d'éviter les artefacts qui peuvent être induits au cours de la synthèse d'ADNc (cf. 2.3.21), (De Saizieu *et al.*, 1998). En revanche cette méthode ne permet pas de détecter les ARNm présents en faible quantité, et reste donc réservée à des études ne nécessitant pas une spécificité élevée.

2.3.23 Marquage direct des ARNm

Cette méthode de marquage direct et spécifique des ARNm a été développée essentiellement pour répondre aux problématiques posées par les modèles bactériens. Au cours d'une première étape, les ARN ribosomiaux qui représentent approximativement 90 % de l'ARN total subissent une transcription

inverse utilisant des amorces spécifiques. Les hybrides ARN-ADN obtenus subissent ensuite une digestion spécifique par une ARNase H et l'ADN restant est ensuite éliminé par digestion. L'échantillon obtenu contient donc essentiellement de l'ARNm. Il subit alors une seconde étape de fragmentation puis de marquage par des molécules de biotine fixées à l'extrémité 5' modifiée des ARNm (Harrington *et al.*, 2000). Cette méthode permet donc de limiter les risques d'hybridation aspécifique qui peuvent être importants chez certains organismes.

2.3.3 Hybridation

2.3.31 Principe

Les paramètres influençant la stabilité de l'hybridation ont été relativement bien étudiés, de façon à déterminer les conditions dans lesquelles la stabilité des hybrides est favorisée (Maskos et Southern, 1992). Cependant, la plupart des études ont été réalisées pour des hybridations en milieu liquide, et peu de données sont disponibles concernant l'hybridation d'une cible avec sa sonde homologue fixée sur une matrice solide. Une étape de développement méthodologique est donc essentielle pour déterminer les sondes et les conditions optimales dans lesquelles doit se dérouler l'étape d'hybridation. Si la mise au point des protocoles peut être longue, elle aboutit finalement à la réalisation non plus d'une seule hybridation, mais d'un grand nombre d'hybridations dans des conditions variées. Cette possibilité de multiplier les hybridations permet pour la première fois d'obtenir une vue globale de l'expression des gènes de l'organisme étudié dans des conditions diverses.

La phase d'hybridation est réalisée dans un incubateur durant toute une nuit. Elle est performante pour un très faible volume de solution (de l'ordre de la dizaine de microlitres) bien qu'elle n'aboutisse à la formation d'hybrides que pour 0,1 à 1 % des sondes marquées (Granjeaud *et al.*, 1999). Elle est suivie d'une étape de lavage (destinée à éliminer de la surface de la puce les cibles non hybridées) qu'il est également essentiel d'optimiser pour obtenir le meilleur rapport possible entre signal et bruit de fond.

2.3.32 Choix des conditions d'hybridation

La formation et la stabilité des hybrides dépendent en premier lieu de facteurs intrinsèques aux sondes utilisées. Il est en effet fréquent d'observer des comportements d'hybridation parfois très variables pour les différentes sondes d'un même gène. Ces variations peuvent être attribuées à de nombreux facteurs comme le taux de GC, la structure des acides nucléiques ou encore la localisa-

tion de la sonde sur la séquence du gène (Barczak *et al.*, 2003). Cet aspect essentiel étant discuté dans le chapitre suivant, seuls les aspects expérimentaux sont présentés dans cette partie.

Le premier paramètre important pour l'hybridation concerne la fixation des sondes sur la lame. L'ajout d'une molécule « espaçante » entre les sondes et la lame permet d'éliminer la gène stérique causée par la surface solide de la puce et augmente l'efficacité de la réaction d'hybridation (Watson *et al.*, 1998). L'utilisation d'une molécule « espaçante » présentant une taille supérieure à quarante atomes de carbone permet ainsi de multiplier le rendement d'hybridation par plus de cent (Shchepinov *et al.*, 1997).

En ce qui concerne l'influence du choix de l'extrémité de la sonde fixée à la puce, il semble qu'elle varie suivant l'étude réalisée. Dans un cas, le profil général d'hybridation peut rester le même, que les sondes soient fixées par leur extrémité 3' ou 5' (Mir et Southern, 1999), alors que dans un autre cas, l'extrémité fixée sur la puce influence la formation des hybrides car les bases qui la constituent sont moins accessibles (Southern *et al.*, 1999).

Le second paramètre à prendre en compte est la température d'hybridation (Relogio *et al.*, 2002). En solution, cette température dépend de la température de fusion (Tm) de l'hybride formé entre une cible et sa sonde, c'est-à-dire de la température à laquelle la moitié des acides nucléiques est sous forme double brin. Il s'agit d'une mesure directe de la stabilité de l'hybridation. La température optimale de la réaction d'hybridation se situe généralement entre 5 et 10 °C au-dessous de la valeur du Tm (Rychlik et Rhoads, 1989). Cependant, il est important de noter que ces considérations ne sont valables que lorsque sondes et cibles sont présentes en solution. Or, dans le cas des puces à ADN, l'immobilisation de la sonde limite les possibilités d'hybridation et implique une diminution du Tm de 7 à 8 °C par rapport au Tm calculé en solution (Wallace *et al.*, 1979). Une étude plus récente a permis de déterminer une partie des paramètres thermodynamiques pour des hybridations sur matrice de gel (Kunitsyn *et al.*, 1996). Ces paramètres suivent approximativement une relation linéaire avec ceux des hybridations réalisées en solution, bien que leurs valeurs soient différentes. Il semble donc qu'il soit possible d'utiliser les paramètres thermodynamiques de l'hybridation en solution et notamment la valeur du Tm pour prédire au moins de façon relative la stabilité de l'hybridation sur puces.

La composition de la solution d'hybridation est également liée au choix de la température de l'expérience. La température d'hybridation peut en effet être diminuée de 0,63 °C pour chaque pour cent de formamide ajouté. Cette diminution de température permet de réduire considérablement le détachement des sondes liées de façon non covalente à la puce. Une solution stan-

dard contient 50 % de formamide. Pour des concentrations inférieures, les signaux obtenus deviennent plus faibles et pour des valeurs plus élevées (au-dessus de 70 à 80 %) une perte de spécificité est observée (Relogio *et al.*, 2002).

Le dernier paramètre influant sur le rendement d'hybridation est la répartition des cibles sur la lame de façon à ce que leur concentration soit homogène. Le taux de diffusion sur un support solide étant plus faible qu'en solution (Chan *et al.*, 1995), une molécule de cible n'est pas capable de diffuser jusqu'à sa sonde homologue. Il est donc nécessaire de mélanger et de faire circuler les réactifs dans une chambre fermée. Cependant avec ce type d'hybridation, les variations expérimentales observées dépendent fortement de la position des sondes sur la puce. Des méthodes d'hybridation automatique ont donc récemment été développées pour pallier ce type de problème. Elles permettent d'obtenir des résultats d'hybridation hautement reproductibles. La machine *Discovery®XT System*[5] (*Ventana*, Tucson, AZ, USA) utilise par exemple un système complexe d'application et de mise en circulation de la solution de cibles sur la lame et bien que la cible soit beaucoup plus diluée que dans les méthodes d'hybridation classique, aucune perte de signal n'est observée (Holloway *et al.*, 2002).

2.4 Acquisition des données

Après hybridation, une étape de lecture de la puce permet de repérer les sondes ayant réagi avec l'échantillon testé. Ce repérage des hybridations effectives est une prise d'empreinte qui est ensuite traitée et analysée par un système informatique. Cette lecture est une étape clé. En effet, sa qualité conditionne de façon majeure la précision des données, et donc la pertinence des interprétations (Yang *et al.*, 2002a). Il existe de nombreux systèmes de lecture en fonction du type de marquage utilisé. Compte tenu de notre problématique, seules les particularités de l'acquisition des données pour les mesures de fluorescence sont présentées et discutées.

2.4.1 Acquisition des images

La première étape fondamentale du processus d'acquisition des données est l'obtention des images (Leung et Cavalieri, 2003). Elle consiste à détecter la fluorescence émise à la surface de la lame après excitation des fluoro-

[5] http://www.ventanadiscovery.com/.

phores. Le procédé de détection le plus classiquement utilisé combine un laser, pour exciter les fluorophores, et un microscope confocal (ou scanner) couplé à un tube photo-multiplicateur (PMT) pour analyser les photons émis par les marqueurs (Schena, 1999). Les canaux de lecture correspondant aux longueurs d'onde 635 nm et 532 nm sont utilisés pour lire respectivement la fluorescence du Cy5 et celle du Cy3. Le signal pour chaque fluorochrome est mesuré par la somme des intensités des pixels du plot. Cette somme représente la quantité totale de cibles hybridées sur les sondes (Dudoit *et al.*, 2002).

Les scanners disposent généralement d'options diverses qui permettent d'améliorer la qualité du signal détecté. L'ajustement le plus couramment utilisé consiste à régler le gain du PMT de façon à ne conserver qu'un nombre minimal de plots présentant des pixels saturés (de l'ordre de 1 %). Cette mesure permet de conserver une gamme dynamique étendue pour les intensités de fluorescence (Leung et Cavalieri, 2003) pour laquelle les intensités sont en relation linéaire avec l'abondance des transcrits (Ramdas *et al.*, 2001 ; Taylor *et al.*, 2001). Smyth *et al.* (2003) prétendent même que, tant que la saturation n'est pas atteinte, l'effet du réglage du PMT est pratiquement négligeable sur la qualité des mesures. Leung et Cavalieri (2003) suggèrent donc d'équilibrer systématiquement les valeurs de PMT pour la détection des deux canaux afin de minimiser la variabilité des résultats obtenus. Des logiciels comme *Masliner* (Dudley *et al.*, 2002) ont été développés pour corriger les problèmes de saturation et de non linéarité du signal qui ne peuvent pas être évités par un simple ajustement du PMT.

Certains scanners permettent également d'ajuster la puissance du laser. Il semble préférable d'utiliser une faible puissance de laser (30 %) pour prévenir la diminution précoce des intensités de fluorescence à la surface des lames (*photo-bleaching*), d'autant plus qu'une forte puissance du laser engendre souvent un bruit de fond important (Leung et Cavalieri, 2003). Enfin il est essentiel de choisir une résolution correspondant approximativement à 10 % du diamètre des plots.

Cette étape permet finalement de générer une image 16-bit pour chaque fluorophore. Ces deux images, en niveau de gris, représentent les intensités de fluorescence lues par le scanner qui reflètent le niveau d'expression des gènes dans les deux conditions expérimentales. Elles sont souvent représentées en fausse couleur. Le vert sert ainsi à caractériser l'échantillon marqué en Cy3, le rouge est réservé au Cy5 et le jaune permet d'indiquer que les cibles marquées en Cy3 et en Cy5 sont hybridées en proportion égale sur un plot.

2.4.2 Analyse des images

L'analyse des deux images obtenues a pour but de mesurer pour chacun des plots la quantité de transcrits hybridés. Elle permet également d'estimer le bruit de fond de l'image. De nombreux logiciels ont été développés pour réaliser les trois parties de cette étape d'analyse.

La première partie de l'analyse est la localisation des plots qui permet de préciser les coordonnées de chaque plot sur l'image. Si la technique de dépôt est fiable et si les lames sont parfaitement lavées, il suffit en principe de positionner une grille en respectant le plan des sondes et les caractéristiques du robot qui a déposé les sondes. En pratique, du fait du recouvrement de certains plots, il est généralement nécessaire d'apporter certains ajustements de façon manuelle.

La seconde partie de l'analyse des images est la segmentation. Elle permet de classer les pixels en tant que signal ou bruit de fond dans le voisinage de chacun des plots identifiés précédemment. La forme, la taille et les irrégularités des plots sont les problèmes majeurs qui sont pris en compte par les algorithmes de segmentation (Yang *et al.*, 2002a).

Différentes méthodes de segmentation ont été développées pour la détection du signal. Une méthode simple consiste à définir un cercle caractérisant le plot. Ce cercle peut ensuite être ajusté pour chaque plot, comme dans le logiciel *Genepix*[6] (*Axon Instruments*, Union City, CA, USA). Une approche plus complexe a été développée par Yang *et al.* (2002a) et intégrée dans le logiciel *Spot*[7] développé pour R. Elle est basée sur l'utilisation d'un algorithme mathématique de morphologie qui permet de prendre en compte les irrégularités des plots.

De nombreuses méthodes ont également été développées pour l'étude du bruit de fond. Le bruit de fond peut être déterminé simplement dans une zone limitée encadrant le plot, cependant, certains auteurs proposent des approches plus complexes. Ainsi, Goryachev et Mac Gregor (2001) estiment par exemple l'intensité du bruit de fond sur une large région voisine du plot. Kooperberg *et al.* (2002) proposent une approche bayésienne. Comme pour les valeurs du signal, le logiciel *Spot* estime un bruit de fond « morphologique ». Avec cette estimation, très peu de gènes ont des valeurs d'intensité négatives après soustraction du bruit de fond. Le logiciel *Genepix* calcule quant à lui un

[6] http://www.axon.fr.
[7] http://www.cmis.csiro.au/iap/spot/spotmanual.htm.

bruit de fond dit « médian » (cf. **Figure A.2.5**) avec lequel plus de 30 % des plots d'une puce ont en moyenne des valeurs négatives (Smyth et Speed, 2003).

Finalement la dernière étape de l'analyse d'image est le calcul d'un ensemble de valeurs destinées à caractériser le signal et le bruit de fond. La moyenne, la médiane, l'écart-type des pixels, et bien d'autres valeurs intermédiaires sont ainsi estimés pour le signal comme pour le bruit de fond. Ces valeurs sont utilisées pour étudier la qualité des lames au cours de l'étape de filtration des données et calculer les rapports d'intensité de chaque plot.

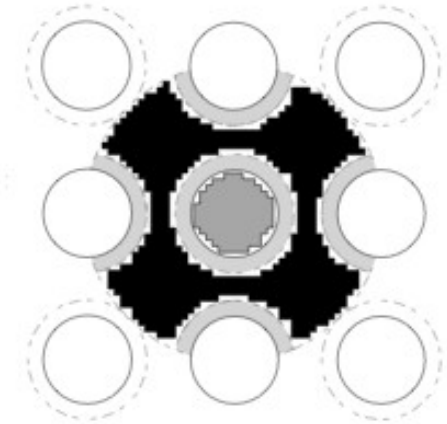


Figure A.2.5 Principe de la segmentation du signal (en gris foncé) et du bruit de fond (en noir) réalisée par le logiciel *Genepix*. Une région d'exclusion localisée entre les zones correspondant au signal et au bruit de fond n 'est pas prise en compte dans le calcul des intensités (d'après le manuel *« GenePix Pro 4.0 array acquisition and analysis software for the Genepix 4000B, Review E »*).

2.4.3 Filtration des données

Les logiciels d'analyse proposent différents critères pour évaluer la qualité des plots, comme la variance de l'intensité des pixels du signal et du bruit de fond, l'aire du plot ou encore son diamètre. Ces critères peuvent être utilisés pour calculer des valeurs permettant une sélection plus efficace comme la circularité des plots, ou le coefficient de variation des intensités. Il est alors facile de repérer les mauvais plots, c'est-à-dire ceux qui présentent une taille anormale, une mauvaise circularité, une forte variance de l'intensité du signal ou du bruit fond, ou encore un faible rapport signal sur bruit. Beaucoup d'auteurs ont donc proposé des méthodes empiriques de filtration qui sont basées sur des seuils de rejet des plots définis *a priori*. D'autres ont proposé des méthodes statistiques d'exclusion (Tseng *et al.*, 2001) ou encore des méthodes de pondération (Smyth et Speed, 2003). Asyali *et al.* (2004) proposent par exemple une approche basée sur l'utilisation d'un modèle bayésien pour déter-

miner les données avec des rapports d'intensités trop faibles et qui doivent donc être éliminés. Cette étape de filtration, qui permet finalement d'éliminer les plots aberrants de l'analyse, a même été automatisée avec le développement de certains logiciels comme celui de Fielden *et al.* (2002) qui s'utilise avec *Genepix*.

L'élimination des mauvais plots permet d'obtenir des résultats de meilleure qualité, cependant certaines analyses statistiques nécessitent l'utilisation de tableaux de données complets. Il est alors nécessaire d'estimer les données manquantes. Pour cela, de nombreux logiciels ont été développés. Ils sont basés principalement sur l'utilisation d'algorithme EM (*Expectation-Maximisation*) (Troyanskaya *et al.*, 2002) ou sur l'utilisation de la méthode des moindres carrés (Bo *et al.*, 2004).

2.5 Transformation des données

Le rapport des intensités de fluorescence est généralement utilisé comme mesure du rapport des concentrations en ARNm des échantillons dans les deux conditions d'étude. Cependant, le rapport d'intensités est influencé par des effets systématiques qui peuvent introduire des biais et qu'il est donc nécessaire d'éliminer avant de pouvoir tirer des conclusions sur le niveau relatif de l'expression des gènes. Le processus d'élimination de ces effets systématiques peut être décomposé en trois étapes qui sont la correction du bruit de fond, la transformation et finalement la normalisation des données.

2.5.1 Notation et représentation des données

L'analyse des données de puces implique l'utilisation d'une notation simplifiée des valeurs d'intensités. Les lettres R (*red*) et G (*green*) servent généralement de référence respectivement aux fluorophores Cy5 et Cy3 et sont les notations utilisées dans l'ensemble de cette thèse. Pour un plot donné, les médianes des intensités des pixels du signal sont notées Rf et Gf et celles des pixels du bruit de fond Rb et Gb. Les valeurs obtenues après correction du bruit de fond sont notées R et G.

Par ailleurs, il est important d'utiliser différentes représentations graphiques de façon à évaluer la qualité des données. Bien que simples et facilement interprétables, ces étapes descriptives sont très informatives et permettent souvent d'orienter les analyses, d'améliorer les protocoles expérimentaux ou même de définir de nouveaux plans d'expériences. La représentation la plus simple des données consiste à placer les valeurs G en abscisses et R en ordon-

nées. Cependant les données d'intensité sont rarement manipulées sans transformation et la transformation la plus couramment employée est celle qui utilise le logarithme en base deux. Il existe plusieurs raisons pour justifier cette transformation. D'une part, la variation du logarithme des intensités est moins dépendante de la grandeur des intensités, et d'autre part, cette transformation permet de se rapprocher d'une distribution symétrique et d'obtenir une meilleure dispersion avec moins de valeurs extrêmes. De plus, elle permet de transformer les rapports de fluorescence en différence et les erreurs multiplicatives en erreurs additives. La base deux du logarithme est préférée à la base naturelle ou la base dix car les intensités des pixels sont comprises entre 0 et (2^{16}–1) pour le niveau de saturation. Dudoit *et al.* (2002) utilisent cette transformation pour proposer un autre type de représentation dans lequel l'axe des abscisses représente la valeur A (*add*) :

$$A = \frac{1}{2}\log_2(RG)$$

et l'axe des ordonnées représente la valeur M(*minus*) :

$$M = \log_2(\frac{R}{G})$$

A et M correspondent respectivement à l'abondance moyenne des transcrits et aux rapports des intensités dans les deux conditions d'étude (cf. **Figure A.2.6**). Sur cette représentation, avec l'utilisation de la base deux pour la transformation logarithmique, il n'existe pas de différence d'expression entre les deux conditions pour M=0 et une induction (ou une répression) d'ordre deux se lit simplement lorsque M=1 (ou pour M=-1) (Smyth et Speed, 2003).

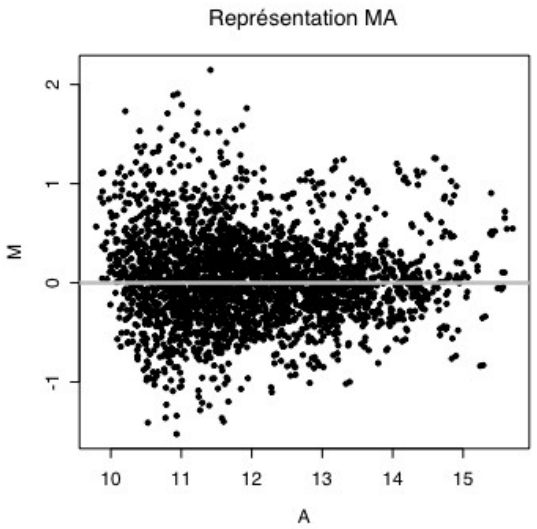


Figure A.2.6 Exemple de représentation MA pour l'ensemble des intensités observées sur une lame.

2.5.2 Correction du bruit de fond

Parallèlement aux représentations des signaux d'intensités des plots, une étude du bruit de fond associé à ces plots peut également être très utile pour juger de la qualité des données. La représentation du bruit de fond en fonction du signal peut en effet révéler une forme structurée qui est le signe d'une répartition non uniforme du bruit de fond.

La simple soustraction des bruits de fond Rb et Gb aux valeurs Rf et Gf peut dans certains cas aboutir à deux situations particulières. Si les niveaux d'expressions sont faibles, les valeurs corrigées R et G peuvent être très petites et générer des ratios M très grands (division par une valeur proche de zéro). De même si les valeurs de bruit de fond sont importantes, les valeurs corrigées peuvent être faibles voire négatives. Or l'apparition de valeurs négatives implique leur disparition au moment de la transformation logarithmique. Pour éviter ce type de problème, de nombreux algorithmes ont été développés (Yang *et al.*, 2002a). Une méthode mixte a par exemple été proposée par Edwards (2003). Une simple soustraction de l'intensité du bruit de fond (i_b) à l'intensité du signal (i_f) est calculée pour tous les plots. Cette valeur est conservée lorsque l'intensité obtenue est supérieure à une valeur seuil δ. Dans le cas contraire, l'intensité totale est calculée avec une fonction de lissage monotone :

$$\begin{cases} (i_f - i_b) > \delta \Rightarrow i = i_f - i_b \\ (i_f - i_b) \leq \delta \Rightarrow i = \delta \cdot \exp(1 - \dfrac{i_b + \delta}{i_f}) \end{cases}$$

2.5.3 Normalisation des données

2.5.31 Introduction

La normalisation est une étape primordiale de l'analyse des données de puces (Bilban *et al.*, 2002). L'intensité de fluorescence de chaque plot à la surface de la puce étant détectée par amplification électronique, aucune mesure absolue de fluorescence ne peut être réalisée. Il est donc nécessaire d'effectuer une normalisation relative des deux intensités de fluorescence avant de procéder au calcul des rapports. Cette étape de normalisation des données permet également de distinguer les sources de variations aléatoires (biologiques et expérimentales) et les sources de variations systématiques (liées aux caractéristiques propres des deux fluorochromes) qui sont souvent importantes et introduisent un biais dans l'estimation des niveaux d'expression. L'objectif de la normalisation est d'éliminer l'influence de ces facteurs qui rendent impossible une interprétation directe des résultats en termes de niveaux d'expression des gènes (Ramdas *et al.*, 2001). En effet, s'il est généralement admis que les inten-

sités R et G sont proportionnelles au niveau d'expression des gènes, cette convention n'a de sens que pour des données normalisées. Les niveaux d'expression des gènes font donc toujours référence aux valeurs R et G normalisées.

2.5.32 Choix des gènes de normalisation

L'étape de normalisation permet de se positionner dans un cadre d'analyse vérifiant certaines hypothèses biologiques. En particulier, il est généralement admis que pour une expérience donnée, un nombre relativement faible (moins de 10 ou 20 %) de gènes sont exprimés de façon différentielle entre les deux conditions d'étude. La validité de cette hypothèse est fortement influencée par le nombre de gènes étudiés sur la puce. Ainsi, il est possible de distinguer trois situations qui influencent le choix de la méthode de normalisation. Pour un nombre important de gènes étudiés, l'hypothèse que la plupart des gènes analysés ne varient pas, ou se répartissent de façon symétrique entre sur et sous exprimés, est réaliste et la normalisation peut être réalisée à partir de l'ensemble des gènes. Pour un nombre moyen de gènes, l'hypothèse précédente est probablement moins réaliste et la normalisation peut être effectuée sur un sous ensemble de gènes déterminés soit *a priori* (gènes de ménage ou cibles de contrôle), soit *a posteriori* (détermination de gènes invariants), soit sur un ensemble mixte composés de cibles de contrôle et de gènes invariants. Enfin si le nombre de gènes étudiés est faible, la normalisation doit être réalisée exclusivement à partir des intensités observées pour des sondes de contrôle prévues sur la puce.

L'utilisation de gènes connus comme étant invariants *a priori* a déjà été présentée (cf. 2.2.22 et 2.3.12). Cependant, il n'est généralement pas possible d'utiliser des gènes de ménage ou des cibles de contrôle. Dans ce cas, il peut être utile de rechercher un ensemble de gènes invariants non connus *a priori*. Cette idée a été proposée en premier par Schadt *et al.* (2001). Le but n'est pas de définir le maximum de gènes invariants mais seulement un nombre suffisant pour une utilisation pertinente. Cette méthode reprise par Tseng *et al.*, (2001) est basée sur l'étude des rangs des intensités de fluorescence R et G. Un gène est considéré comme invariant si ses rangs sont égaux pour les deux fluorochromes. En pratique, pour un gène donné, si le rang de R ne diffère pas de plus d'une différence de d par rapport au rang de G et si l'intensité moyenne n'est pas située parmi les l intensités moyennes les plus fortes ou les l intensités moyennes les plus faibles, alors le plot est considéré comme « invariant ». Les valeurs d et l sont estimées empiriquement en fonction du nombre de données. Lorsque le nombre de plots est très important, la détermination de d se fait par une méthode itérative. L'écart d est calculé sur la base du nombre de gènes sé-

lectionnés à l'étape antérieure, multiplié par un pourcentage *p* qui correspond à l'écart toléré autour d'un rang et la limite *l* est appliquée seulement pour la première itération. Le processus d'itération s'achève lorsque le nombre de plots invariants obtenus entre deux itérations successives est inchangé. Tseng *et al.* (2001) utilisent les paramètres l=25 et p=0,02 pour une étude portant sur 4129 plots.

Après avoir déterminé les gènes sur lesquels la normalisation doit porter, il est important de choisir le type de normalisation le plus adéquat à appliquer parmi les nombreuses méthodes de normalisation développées. Des normalisations successives peuvent être appliquées, de la plus simple à la plus sophistiquée. Il est donc important de commencer avec la méthode la plus simple et de ne passer à une méthode plus complexe que si un critère de qualité particulier est amélioré.

2.5.33 Normalisation globale

Les méthodes de normalisation globale sont les premières qui ont été développées. Elles sont basées sur l'hypothèse que les intensités R et G globales sont liées par un facteur constant. Deux de ces méthodes sont principalement utilisées.

La première est la normalisation dite par la médiane, ou par la moyenne qui permet de centrer la distribution des log ratios sur 0 :

$$\begin{cases} R = kG \\ \log_2 \frac{R'}{G'} = \log_2(\frac{R}{kG}) = \log_2(\frac{R}{G}) - \log_2 k \end{cases}$$

Le paramètre constant ($\log_2$ k) est la médiane ou la moyenne des log ratios pour l'ensemble des gènes.

La seconde est la normalisation linéaire qui a été proposée par Finkelstein *et al.* (2002) : $\log_2 R = a + b(\log_2 G)$.
Les coefficients du modèle linéaire, calculés par itérations successives, sont utilisés pour normaliser $\log_2$R par rapport à $\log_2$G :

$$\begin{cases} \log_2 R' = \log_2 R - a \\ \log_2 G' = b(\log_2 G) \end{cases} \Rightarrow \begin{cases} R' = \frac{R}{a} \\ G' = G^b \end{cases}$$

Une méthode similaire appelée *shift-log* a également été proposée par Newton *et al.* (2001) et par Kerr *et al.* (2002).

L'ensemble de ces solutions de normalisation reste néanmoins d'un intérêt limité en raison de la nature souvent non linéaire des relations entre les intensités observées (Ramdas *et al.*, 2001 ; Schadt *et al.*, 2001).

2.5.34 Normalisation intensité-dépendante

Pour pallier les problèmes que soulève la normalisation globale, Yang *et al.* (2002b) proposent une approche de normalisation dépendante de l'intensité A pour chaque gène (*g*) : $M_g' = M_g - c_g(A)$. Cette méthode, contrairement aux précédentes, permet de prendre en compte la non-linéarité de la relation qui existe généralement entre R et G. La fonction $c_g(A)$ est estimée à l'aide d'une méthode de régression non linéaire. Les auteurs proposent une méthode de régression locale appelée *locally weighted scatterplot smoothing* qui peut être de type *loess* (régression linéaire locale) ou *lowess* (fonction quadratique locale) (Cleveland, 1979). Pour chaque valeur d'intensité A_i, la correction $c(A_i)$ est obtenue par la méthode des moindres carrés pondérés à partir d'un ensemble de *q* points dans le voisinage de A_i. Il est nécessaire d'estimer les paramètres β_0 et β_1 qui minimisent :

$$\sum_{j=1}^{q} w_i(A_g)(M_k - \beta_0 - \beta_1 A_g)^2$$

La fonction de pondération w_i affecte un poids faible aux points (A_g, M_g) pour lesquels la valeur A_g est éloignée de A_i et la valeur M_g est trop importante. Le poids des points en dehors du voisinage est considéré comme nul. La fonction *lowess* est contrôlée par le paramètre f ($0<f\leq1$) qui spécifie la proportion du nuage de points définissant le voisinage pour ajuster la droite des moindres carrés pondérés. Yang *et al.* (2002b) préconisent l'utilisation de f=0,4.

La méthode de normalisation *lowess* permet de prendre en compte l'absence de relation linéaire entre les intensités, mais ne règle pas le problème de l'hétéroscédasticité de. En effet, l'observation des graphes M=f(A) montre très souvent que pour les faibles valeurs de A la variabilité sur les rapports (M) est très forte. Pour pallier ce problème d'hétéroscédasticité certains auteurs ont tout simplement suggéré d'éliminer les données en dessous d'une certaine valeur de signal (Yang *et al.*, 2001). Pour éviter l'élimination d'un nombre parfois élevé de plots, d'autres auteurs ont proposé des transformations visant à stabiliser les variances pour les faibles intensités de signal. Ainsi, Durbin *et al.* (2002) et Huber *et al.* (2002) proposent une transformation *arsinh* des données, sous l'hypothèse d'une relation quadratique entre la variance et l'intensité relative du signal. Cette fonction permet de prendre en compte l'effet-intensité sur la variance des données. Cui *et al.* (2002) affirment cependant que cette transformation fonctionne bien sur des données simulées mais assez mal sur des données réelles. Ils proposent une transformation *linlog* qui combine une transformation linéaire pour les faibles intensités et une transformation logarithmique pour les intensités plus fortes. La transition entre les deux transformations est lissée et prend en compte les avantages des deux méthodes. Les auteurs la justifient par le fait que les erreurs additives sont prédominantes pour les faibles va-

leurs d'intensité, alors que les erreurs multiplicatives dominent pour les fortes valeurs d'intensités.

2.5.35 Normalisation intensité-dépendante par aiguille

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 lavage de la lame. Cependant, 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.* (2002b) proposent d'éliminer l'hétérogénéité spatiale en intégrant une normalisation aiguille dépendante. Cette normalisation est réalisée de la façon suivante : $M_g' = M_g - c_g(A_i)$ où la fonction $c_g(A_i)$ est estimée à l'aide d'une régression *lowess* pour chacune des aiguilles i du robot de dépôt.

Cette méthode permet généralement d'éliminer la plupart des hétérogénéités spatiales (Smyth et Speed, 2003). Dans le cas contraire, il est alors possible d'ajouter une correction sur la surface de la lame à partir des coordonnées (x,y) des plots : $M_g' = M_g - c_g(x,y)$. Il est également possible d'éliminer l'influence spatiale et l'influence de l'intensité dans un même temps en appliquant la correction proposée par Cui *et al.* (2002) : $M_g' = M_g - c_g(A,x,y)$.

Ces méthodes de normalisation sont très robustes pour un petit nombre de gènes différentiellement exprimés. En revanche, elles ne peuvent s'appliquer que sous l'hypothèse que ce nombre de gènes reste très faible par rapport au nombre total de gènes impliqués, et que la répartition des gènes surexprimés et sous-exprimés est symétrique. Dans le cas contraire, il est toujours possible d'appliquer ces méthodes sur des cibles de contrôles (Colantuoni *et al.*, 2002). Cela nécessite néanmoins la répartition de contrôles sur toute la surface de la lame, ce qui est encore rarement le cas.

2.5.36 Normalisation itérative

Lorsque les données contiennent un nombre important de gènes différentiels, il est possible d'utiliser une procédure de normalisation plus robuste (Wilson *et al.*, 2003). Pour cela, une première étape permet de calculer les résidus issus de la régression *lowess*. La médiane des valeurs absolues de ces résidus ou MAD (*median absolute deviation*) est alors utilisée pour déterminer de nouvelles estimations robustes des valeurs de pondération. Une seconde étape permet de calculer les intensités normalisées en utilisant à la fois ces estimations robustes des valeurs de pondération et celles de la régression locale. Ces deux étapes répétées itérativement permettent d'obtenir des données normalisées même en présence d'un nombre élevé de gènes exprimés de façon différentielle.

2.5.37 Normalisation mixte

La normalisation mixte appelée aussi composite a été développée par Yang *et al.* (2002b). Elle est basée à la fois sur une normalisation intensité-dépendante et sur l'utilisation d'un ensemble de cibles de contrôle. Ces cibles sont constituées d'un mélange d'environ 2000 EST de Souris et forment un *MSP* (*Microarray Sample Pool*). Ce *MSP* est utilisé comme de l'ADN génomique (cf. 2.3.12). Il présente l'avantage de ne pas contenir de régions non codantes ce qui limite les risques d'hybridation aspécifique et permet d'augmenter le taux d'hybridation. Des dilutions de ce MSP sont effectuées de façon à obtenir des valeurs d'intensités pour toute la gamme de fluorescence. Les résultats obtenus sur les lames permettent de calculer une courbe de normalisation pour les plots *MSP* et une autre pour l'ensemble des plots de la puce. Ces deux courbes sont ensuite utilisées pour déterminer une courbe dite « composite » qui correspond à une courbe moyenne entre les deux précédentes et pondérée en fonction du nombre de gènes par niveaux d'intensité.

Cette méthode est particulièrement intéressante sur les plots situés aux extrémités de la distribution et sur les plots possédant de faibles valeurs de A. Dans ces deux cas, les valeurs des plots *MSP* permettent une meilleure estimation de la courbe de normalisation en augmentant la densité de points au voisinage des plots concernés. De plus cette méthode offre la possibilité de vérifier que la courbe intensité-dépendante calculée n'est pas biaisée par des gènes exprimés de façon différentielle (Smyth et Speed, 2003).

2.5.38 Normalisation inter-lames

L'ensemble des méthodes proposées dans cette partie est destiné à la normalisation individuelle des différentes lames d'une expérience. Cependant la normalisation peut également être réalisée entre des répétitions de lames et même sur l'ensemble des lames d'une expérience. Pour cela, Zien *et al.* (2001) proposent une méthode complète dite de centralisation qui permet à la fois de normaliser et de calibrer les données de façon à permettre les comparaisons inter-lames. Après élimination du bruit de fond, les rapports des intensités de fluorescence sont calculés pour l'ensemble des plots puis une méthode itérative permet d'éliminer les gènes dont les rapports sont situés aux extrémités de la distribution. Un facteur de calibrage (S_e) est alors calculé pour chaque condition (c). Sa valeur est estimée au maximum de vraisemblance en utilisant les valeurs des médianes des rapports d'intensité pour chaque couple de conditions. Ce facteur permet finalement de transformer les intensités correspondant à chaque condition. D'autres méthodes plus complexes ont également été proposées (Workman *et al.*, 2002). Cependant la mise au point de nouvelles méthodes de normalisation inter-lames reste nécessaire pour permettre la comparaison de toutes les expériences (Schadt *et al.*, 2001), quelles que soient les méthodes uti-

lisées et les laboratoires dans lesquels elles sont effectuées. Cette comparaison reste encore extrêmement difficile compte tenu de la diversité des techniques (préparation des échantillons, type de puces) et des contrôles de normalisation employés.

2.6 Analyse statistique des données

2.6.1 Des données complexes

Les données d'expression sont des données complexes pour l'analyse statistiques en raison de plusieurs caractéristiques inhabituelles.

La plus évidente de ces caractéristiques est certainement la dissymétrie importante des tableaux de données. En effet, le nombre de conditions expérimentales (par exemple le nombre d'échantillons prélevés sur un ensemble de patients) est bien souvent très faible par rapport aux nombres de gènes analysés. Il existe donc une redondance extrême des régresseurs pour caractériser les conditions expérimentales par les gènes, et une sous-paramétrisation pour réaliser la démarche inverse.

L'observation de la distribution des données d'expression montre une absence de normalité. En effet, dans la plupart des organismes cellulaires (des bactéries les plus simples aux eucaryotes les plus organisés), à un instant donné, la grande majorité des gènes s'exprime à un niveau très faible (cf. **Figure A.2.7**). Il est important de noter que le niveau d'expression d'un gène n'est pas lié à l'importance de sa fonction. Les facteurs de transcription, par exemple, sont exprimés à des niveaux très faibles alors que certains, très ubiquitaires, sont à l'origine de changements physiologiques majeurs dans la cellule. La distribution des niveaux d'expression est donc systématiquement dissymétrique et très étirée vers les fortes valeurs d'expression.

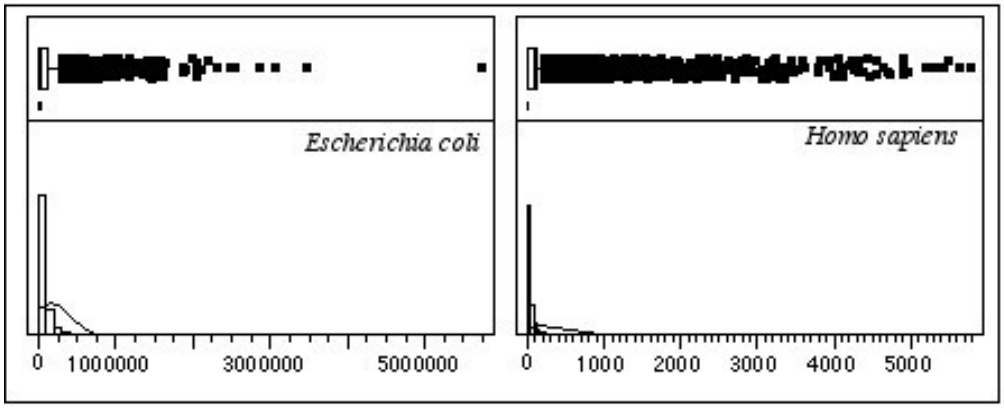


Figure A.2.7 Distribution des niveaux d'expression (unité relative de fluorescence en abscisse) chez la bactérie *Escherichia coli* (5522 gènes) et chez l'homme (12 632 gènes). Les diagrammes (en haut) montrent l'étirement de la distribution vers les fortes valeurs d'expression (ce qui signifie que seul un faible nombre de gènes est fortement exprimé.) La courbe (en bas) représente la distribution normale théorique. Les données sont issues de la base d'expression GEO[8] (numéros d'accession : GSE33 et GSE516).

Enfin, il est important de ne pas oublier que les données d'expression se caractérisent par le fait que les gènes étudiés ne sont pas indépendants. En effet, la plupart des mécanismes biologiques sont régulés au niveau transcriptionnel par des cascades d'activateurs ou d'inhibiteurs, dans lesquelles certains facteurs de transcription sont capables d'activer ou d'inhiber l'expression de plusieurs dizaines de gènes à la fois. Cette dépendance est un facteur très important à prendre en compte lors des analyses statistiques.

2.6.2 Importance du plan expérimental

Une attention particulière concernant les puces à ADN a été accordée aux méthodes d'analyse statistique des données. Cependant il est essentiel de ne pas négliger la définition du plan expérimental qui est à la base même de la qualité de cette analyse (Leung, 2002). Pour cela, il est important de définir les différentes questions posées et de les hiérarchiser entre elles mais aussi de prendre en compte les contraintes matérielles comme le nombre d'hybridations possibles, ainsi que les sources de variabilités techniques et biologiques (Churchill, 2002). Il existe en effet trois sources principales de variation dans les expériences de puces à ADN. La plus importante est la variabilité biologique qui est intrinsèque à tous les organismes et dépend de facteurs génétiques et

[8] http://www.ncbi.nih.gov/geo/.

environnementaux. Elle est normalement prise en compte au moment du choix des échantillons (cf. 2.3.11). La seconde est la variabilité technique qui est introduite au cours des étapes d'extraction, de marquage et d'hybridation. Enfin le dernier type de variabilité est l'erreur de mesure qui est associée à la lecture des signaux de fluorescence (Churchill, 2002).

2.6.21 Unité expérimentale et répétitions

Le premier point d'un plan d'expérience est la définition de l'unité expérimentale, en d'autres termes, la question est de savoir quel est le meilleur échantillon à utiliser pour réduire la variance biologique. Par exemple, pour comparer les ARNm issus de deux sources A et B, il est possible d'utiliser trois extractions et marquages à partir de chacune des sources, puis de les rassembler aléatoirement deux à deux sur une lame et de moyenner les résultats obtenus. Une alternative est de rassembler les ARNm issus respectivement des sources A et B, puis de les subdiviser en trois échantillons pour réaliser finalement trois répétitions techniques. Quelle est la meilleure solution ? Il est difficile de répondre à cette question de façon générale. Le fait de rassembler des échantillons augmente en effet la précision en réduisant la variance des comparaisons d'intérêt, mais au prix de laisser à un seul échantillon une influence forte sur les résultats et avec le risque d'extrapoler des conclusions erronées pour l'ensemble de la population. Aucune expérience n'a actuellement été réalisée pour répondre à cette question en ce qui concerne les mesures de fluorescence. Seule une étude réalisée avec des puces *Affymetrix* (citée dans Yang et Speed, 2003) montre que le fait de rassembler des échantillons n'induit qu'une faible amélioration de la précision et qu'il n'existe pas réellement de biais dans les résultats. Cette étude laisse penser que le gain de précision obtenu en rassemblant des échantillons ne justifie probablement pas les risques potentiels, et qu'il est sans doute préférable de pouvoir déterminer la variation qui existe entre les échantillons, plutôt que de perdre la capacité de le faire. Des études réalisées sur des mesures de fluorescence permettront sans doute de confirmer ou d'infirmer ces conclusions, même si dans certains cas, il n'existe pas d'autre possibilité que de rassembler des échantillons.

Une fois que l'unité expérimentale est définie, il est essentiel de choisir le nombre de répétitions de mesures à réaliser. En effet, les répétitions sont à la base des inférences statistiques dont elles augmentent la précision. Pour les expériences sur puces, il existe trois types de répétitions. Les premières sont les répétitions biologiques qui correspondent à des échantillons provenant de sources différentes. Ces répétitions biologiques sont essentielles pour tirer des conclusions pour la population et non pas simplement pour l'échantillon d'étude. Les secondes répétitions sont les répétitions techniques qui corres-

pondent aux marquages indépendants réalisés à partir d'un même échantillon. Elles permettent d'augmenter la précision des résultats et d'évaluer la variabilité technique du protocole expérimental (Hoheisel et Vingron, 2000 ; Lee *et al.*, 2000). Enfin, la répétition des plots sur la puce permet d'augmenter la précision des mesures.

2.6.22 Associer les échantillons

Compte tenu du coût des expérimentations et du nombre parfois limité d'échantillons, il est souvent difficile de faire toutes les comparaisons possibles d'échantillons au sein d'un plan d'expérience. Il est donc essentiel de définir le plan le plus adapté aux nombres d'échantillons disponibles et à la question posée, mais aussi de veiller à choisir une représentation équilibrée des échantillons et des fluorochromes de façon à compenser partiellement certains biais techniques. Cependant, compte tenu de la nature même des expérimentations sur puces, tous les plans qui peuvent être proposés restent toujours des plans incomplets. Enfin, s'il est souvent impossible de réaliser un tirage aléatoire des échantillons pourtant à la base de la validité des analyses statistiques (Fisher, 1951), il est tout à fait possible de le mettre en place pour les lames. Le dépôt est en effet effectué sur un lot de lames qui sont numérotées dans l'ordre dans lequel elles ont été traitées. Il est donc important de les tirer aléatoirement au moment de l'utilisation (Oleksiak *et al.*, 2002).

2.6.221 Plan en flip-flop

Le plan dit en flip-flop est un plan simple et efficace pour la comparaison directe de deux échantillons (Kerr et Churchill, 2001b). Le principe est de réaliser un marquage réciproque de chaque échantillon par les deux fluorochromes. Sur ce dispositif en carré latin, différentes associations des deux échantillons sont possibles suivant le type de répétitions choisi (cf. **Figure A.2.8**). Dans tous les cas, ce plan permet d'éliminer les biais liés aux comportements différents des deux fluorochromes, mais induit une confusion d'effet entre le traitement, le fluorochrome et la lame (Dudoit *et al.*, 2002).

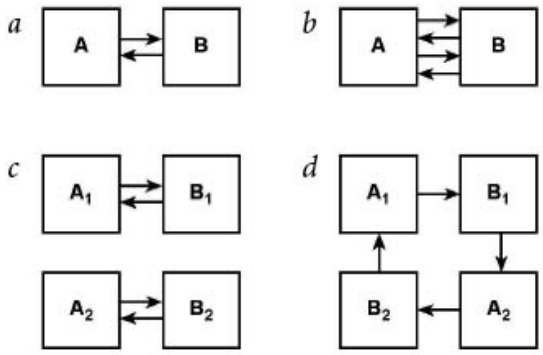
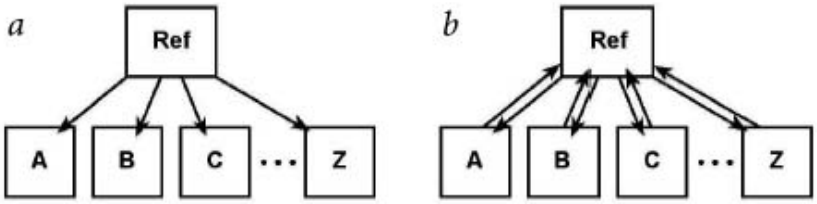


Figure A.2.8 Exemple de plan de comparaison direct de deux échantillons (A et B). Plusieurs associations sont possibles : en flip-flop (a), en flip-flop avec répétitions techniques (b), en flip-flop avec répétitions biologiques (c) et en boucle (d) (d'après Churchill, 2002).

2.6.222 Plan avec référence

Dans un plan avec référence, toutes les comparaisons sont effectuées par rapport à un échantillon de référence (cf.**Figure A.2.9**). La moitié des mesures est donc réalisée sur cet échantillon de référence qui ne présente en général que peu d'intérêt biologique. Cette pratique nécessite de plus l'utilisation de deux fois plus d'échantillons d'intérêt biologique (Jin *et al.*, 2001) et les comparaisons d'intérêt ne sont jamais réalisées sur la même lame. En dépit de cette relative inefficacité, ce plan présente quelques avantages. En effet, le chemin à parcourir entre deux échantillons ne nécessite jamais plus (et jamais moins) de deux étapes. Toutes les comparaisons sont donc faites avec la même efficacité. Enfin, ce plan peut se révéler intéressant lorsque la référence a une forte signification biologique (Churchill, 2002).

Figure A.2.9 Exemples de plan avec référence. Le plan classique utilise une lame



pour chaque échantillon (A,B, C…Z) (a). Une variation possible est de coupler ce plan avec des flip-flop pour chaque échantillon (b) (d'après Churchill, 2002).

2.6.223 Plan en boucle

Il s'agit d'une extension du plan en flip-flop qui permet de comparer deux échantillons (cf. **Figure A.2.8 d**) et d'une alternative au plan de référence

pour l'étude de plusieurs échantillons (cf. **Figure A.2.10**) (Yang et Speed, 2002). Ce type de plan peut également être utilisé pour les expériences destinées à tester l'influence simultanée de plusieurs facteurs. Dans ce cas, il est essentiel d'associer les comparaisons d'échantillons qui présentent le plus grand intérêt sur une même lame. En revanche, ce plan ne permet pas, contrairement au plan avec référence, d'ajouter un nouvel échantillon dans l'analyse ou de recommencer une lame dans le cas d'une hybridation ratée.

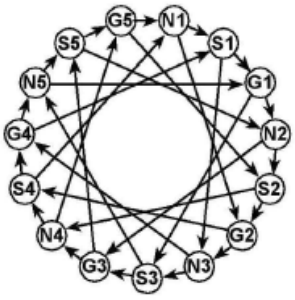


Figure A.2.10 Exemple de plan en boucle pour trois échantillons différents (N,S et G) (d'après Churchill, 2002).

2.6.3 Détection des gènes différentiellement exprimés

La sélection des gènes différentiellement exprimés nécessite deux étapes. La première est le choix et l'utilisation d'un test statistique qui permette le classement des gènes en fonction des probabilités obtenues. La seconde passe par la définition d'un seuil de significativité qui distingue les gènes exprimés de façon différentielle.

2.6.31 Comparaison de deux conditions

2.6.311 Comparaison intra-lame

Initialement, la détection de gènes exprimés différentiellement était simplement basée sur la valeur des log ratios calculés entre les deux conditions comparées. Avec cette méthode appelée parfois « *fold change* », tous les gènes dont le rapport était au-dessus d'une valeur choisie arbitrairement étaient considérés comme exprimés de façon différentielle (Schena *et al.*, 1996 ; Derisi *et al.*, 1997). Des valeurs comprises entre trois et cinq étaient considérées comme significatives en comparaison à des témoins positifs ou aux gènes de ménage (Schena *et al.*, 1995 ; Schena *et al.*, 1996). Cependant, ce test n'est pas un test statistique et il n'existe pas de valeur seuil qui puisse être associée à un intervalle de confiance pour désigner les gènes différentiellement exprimés. De plus, cette méthode peut être biaisée par la qualité des données. Ainsi, un gène faiblement exprimé peut être identifié comme exprimé de façon différentielle,

simplement parce que son rapport est plus élevé que celui des gènes fortement exprimés (problème de la division par une valeur très proche de zéro) (Newton *et al.*, 2001 ; Rocke et Durbin, 2001).

Avec l'amélioration de la qualité des images obtenues, liée à l'amélioration des performances de chacune des étapes du processus de fabrication et d'utilisation des puces à ADN, des méthodes plus fines et basées sur des analyses statistiques ont été développées. Ces méthodes de comparaison des profils d'expression de deux échantillons hybridés sur une lame unique sont toutes basées sur la modélisation statistique de R et de G. Elles diffèrent principalement par les hypothèses de distribution qui sont faites sur les intensités pour définir une règle permettant la détection des gènes différentiellement exprimés.

Chen *et al.* (1997) proposent une règle basée sur différentes hypothèses de distribution des intensités, dont la normalité. Sapir et Churchill (2000) suggèrent d'identifier les gènes différentiellement exprimés en utilisant les probabilités calculées sous l'hypothèse d'un modèle mixte pour les log ratios normalisés. Une des limitations de ces deux méthodes est qu'elles ignorent l'information concernant les quantités de transcrits (disponible avec le produit des intensités RG). Deux autres méthodes reconnaissant ce problème ont donc été développées pour définir une règle de décision prenant en compte l'abondance des transcrits. L'approche de Roberts *et al.* (2000) est basée sur l'hypothèse que les intensités R et G sont indépendantes et normalement distribuées avec des variances dépendant de la moyenne. Quant à Newton *et al.* (2001), ils considèrent un modèle hiérarchique sous lequel les intensités R et G suivent une loi gamma et suggèrent d'identifier les gènes différentiellement exprimés en utilisant le calcul du rapport des chances d'avoir une expression différentielle pour chaque gène. Le calcul de ce rapport des chances dépend de (R+G) et (RG) de façon à prendre en compte l'abondance des transcrits. Chacune de ces méthodes produit une règle de décision dépendant du modèle qui permet de visualiser dans le plan (log R, log G) deux courbes (cf. **Figure A.2.11**). Les gènes situés à l'extérieur de ces courbes sont considérés comme différentiellement exprimés.

Ces modèles sont intéressants lorsque pour des raisons diverses (lames inutilisables, matériel biologique en quantité limitée) seule une lame est disponible pour l'étude. Leurs hypothèses sont cependant très conservatrices. De plus, compte tenu de la particularité des expériences de puces à ADN, ces modèles risquent de détecter essentiellement des variations systématiques et aléatoires inhérentes à la nature même des données (Dudoit *et al.*, 2002). Enfin, lorsque d'autres lames sont disponibles, les modèles d'erreur associés ne per-

mettent pas de prendre en compte les données répétées dans le calcul des probabilités d'expression différentielle.

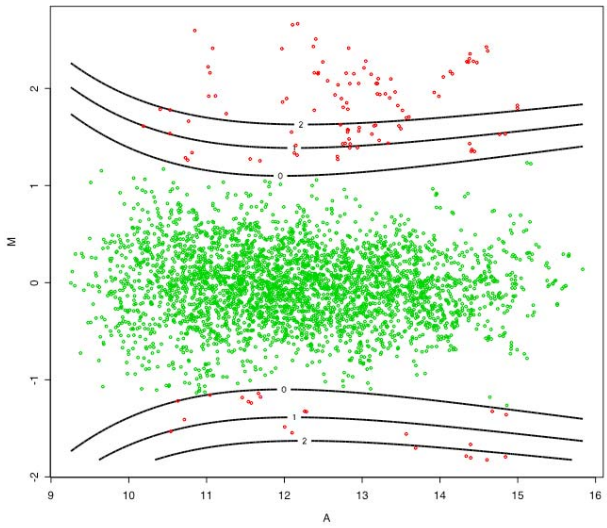


Figure A.2.11 Exemple du graphe obtenu pour une analyse de Newton. Les points en rouge sont les points considérés comme différentiellement exprimés.

2.6.312 Comparaison inter-lames

Approche classique : test de t

L'approche classiquement utilisée pour déterminer les gènes exprimés de façon différentielle dans un plan expérimental comparant deux conditions est l'utilisation d'un test de t réalisé pour chaque gène. Si les deux conditions sont appelées 1 et 2 pour les N gènes de l'étude, le niveau d'expression du gène g dans chacune des deux conditions est appelé y_{g1} et y_{g2}. Les véritables niveaux d'expression du gène g sont notés μ_{g1} est μ_{g2}. L'hypothèse que le gène n'est pas différentiellement exprimé dans les deux conditions est appelée $H_{0,g}$ et l'hypothèse alternative $H_{1,g}$. Ces hypothèses peuvent s'écrire de la façon suivante pour les conditions 1 et 2 :

$$H_{0,g} \Leftrightarrow \mu_{g1} = \mu_{g2}$$

$$H_{1,g} \Leftrightarrow \mu_{g1} \neq \mu_{g2}$$

Pour identifier l'ensemble des gènes qui sont exprimés différentiellement, il suffit alors de réaliser les tests d'hypothèses pour chacun des N gènes et de conserver les gènes pour lesquels l'hypothèse nulle est rejetée. Pour réaliser ces N tests, la statistique usuelle est :

$$t_g = \frac{\overline{y_{g2}} - \overline{y_{g1}}}{\sigma_g}$$

Pour chaque gène, le numérateur de cette statistique est simplement la différence des moyennes des niveaux d'expression dans les deux conditions et le dénominateur représente l'écart-type estimé de la différence des moyennes. Le test est ensuite réalisé sous l'hypothèse d'indépendance et de normalité des niveaux d'expression du gène dans les deux conditions. La loi de la variable aléatoire T_g, sous l'hypothèse nulle est une loi de Student et correspond au test traditionnel de comparaison de moyenne.

La réalisation d'un test de t pour chaque gène est intéressante car elle n'impose pas une homogénéité des variances. Cependant, la puissance du test peut être faible si le nombre de répétitions pour chaque condition est restreint, ce qui est souvent le cas dans les expériences de puces à ADN. Une autre limitation concerne la stabilité de la variance de chaque gène. Ainsi, lorsque la variance estimée pour un gène est petite, la valeur de t_g peut être importante même pour un faible différentiel d'intensité. Pour pallier ce problème, certains auteurs proposent d'estimer une variance globale pour l'ensemble des gènes (Arfin *et al.*, 2000 ; Tanaka *et al.*, 2000). Cette estimation ne peut cependant être réalisée que sous l'hypothèse forte d'homogénéité des variances, qui est rarement vérifiée dans les expériences de puces à ADN. Enfin il est essentiel de ne pas oublier que la réalisation d'un nombre élevé de tests nécessite de prendre en compte les problèmes spécifiques des tests multiples (cf. 2.6.4).

Approche bayésienne

L'utilisation de l'approche classique, décrite précédemment, montre que les variances sont difficiles à estimer et soumises à des fluctuations lorsque le nombre de répétitions est faible. Si des estimations plus fiables peuvent être obtenues en combinant les données de l'ensemble des gènes, elles restent biaisées lorsque le principe d'homoscédasticité est violé. Des versions modifiées et empiriques du test de t classique ont donc été proposées.

La modification la plus simple à mettre en œuvre est celle de la version SAM (*Signifiance Analysis of Microarrays*) du test de t, connue également sous le nom de test S (Tusher *et al.*, 2001). Pour pallier le problème lié à l'estimation des petites variances, une constante positive a (correspondant au quatre-vingt dixième percentile de la distribution des écarts-types de tous les gènes) est ajoutée au dénominateur du test classique :

$$t_g = \frac{\overline{y_{g2}} - \overline{y_{g1}}}{\sigma_g + a}$$

Baldi et Long (2001) proposent une autre approche avec un test de t régularisé. Pour cela, le calcul du dénominateur du test de t classique est basé sur une estimation pondérée de la variance qui prend en compte à la fois l'estimation de la variance gène spécifique et celle de la variance moyenne globale. Ces deux approches sont généralement regroupées sous le terme « approche bayésienne ». Cependant, elles ne mettent pas véritablement en jeu l'utilisation des modèles de Bayes, contrairement à la dernière approche proposées par plusieurs auteurs et appelée test B (Efron et Tibshirani, 2002 ; Lonnstedt et Speed, 2002). Dans cette méthode, une proportion définie *a priori* de gènes différentiellement exprimés (par exemple 1 %) permet l'estimation de la variable B qui correspond au logarithme du rapport des chances d'expression différentielle et d'expression non différentielle.

Ces trois approches sont implémentées dans différents logiciels. Le logiciel *SAM* (Tusher *et al.*, 2001) propose les tests S et B et le logiciel *CyberT* le test de t régularisé (Baldi et Long, 2001). Si elles permettent dans la plupart des cas d'améliorer les problèmes liés aux estimations de la variance, elles restent soumises, comme le test de t classique, aux problématiques des tests multiples.

Une étude des différentes méthodes de comparaison inter-lames montre qu'elles aboutissent parfois à des résultats très différents, aussi bien en ce qui concerne le nombre et la nature des gènes que leur classement (Pan, 2002).

2.6.32 Comparaison de plus de deux conditions

Lorsque le plan expérimental est destiné à l'étude de plusieurs conditions et de leur interaction, l'analyse de variance (ou ANOVA) qui permet d'apprécier l'effet de variables qualitatives (les facteurs) sur une variable numérique (le niveau d'expression des gènes) est la méthode qui est naturellement choisie (Kerr *et al.*, 2000).

2.6.321 Modèle global d'analyse de variance

Le modèle global d'analyse de variance permet de représenter le niveau d'expression des gènes comme une combinaison spécifique de différents facteurs. Pour une expérience sur puces à ADN, les principaux facteurs qui influencent les mesures d'expression sont la lame (A pour *array*), le fluorophore (D pour *dye*), la condition expérimentale (V pour *variety*) et le gène (G pour *gene*). Il est alors possible de proposer le modèle suivant pour représenter les niveaux d'expression : $y_{ijkgr} = \mu + A_i + D_j + V_k + G_g + VG_{kg} + \varepsilon_{ijkgr}$. Avec les termes suivants :

- y_{ijkgr} représente le logarithme des niveaux d'expression,
- μ est l'effet global,
- A_i est l'effet de la $i^{ième}$ lame (I lames),
- D_j est le $j^{ième}$ niveau de l'effet fluorophore D (2 couleurs),
- V_k correspond au $k^{ième}$ niveau de l'effet variété V (K conditions),
- G_{kg} traduit l'effet du $g^{ième}$ gène pour le facteur G (N gènes),
- VG_{kg} représente l'interaction entre la $k^{ième}$ variété et le $g^{ième}$ gène (il s'agit donc du terme d'intérêt),
- ε_{ijkgr} représente les erreurs de mesures indépendantes qui suivent une loi normale centrée de variance σ^2.

Ce modèle est le plus simple possible dans la mesure où seuls les facteurs principaux et l'interaction d'intérêt d'ordre 2 VG_{kg}, qui traduit les variations d'expression du gène g dans la condition k, sont considérés (Kerr et Churchill, 2001a). Des modèles plus compliqués avec d'autres interactions peuvent également être envisagés. Cependant, l'ensemble des facteurs et des interactions de tout ordre n'est pas toujours estimable car les plans sont toujours incomplets. En effet, selon le plan d'expérience envisagé, certains de ces facteurs et de ces interactions sont partiellement ou totalement confondus. Par ailleurs, ce modèle est un modèle à effet fixe dans lequel tous les facteurs ont un nombre fini de niveaux. En raison de la nature de certains effets, il semble néanmoins raisonnable de considérer certains facteurs (la lame par exemple) comme des facteurs à effets aléatoires. Dans ce cas le modèle d'analyse de variance utilisé devient un modèle mixte (Wolfinger *et al.*, 2001).

Kerr *et al.* (2000) ont proposé dans un premier temps d'appliquer directement cette analyse de variance pour des données brutes, sans réaliser de normalisation et en considérant les termes A, D et V comme des facteurs de normalisation. Cependant, l'étude des résidus estimés par ce type de modèle montre que certains effets non linéaires n'autorisent pas une approche aussi directe. Il semble donc préférable d'effectuer une étape de normalisation avant d'appliquer l'analyse de variance. Par ailleurs, ce modèle d'analyse de variance ne permet d'obtenir une représentation des niveaux d'expression que pour l'ensemble des mesures de l'expérience. Une nouvelle approche proposée initialement par Dudoit *et al.* (2002) a récemment été développée pour extraire la partie du modèle liée aux aspects de normalisation et permettre une analyse de variance gène-spécifique (Cui et Churchill, 2003).

2.6.322 Modèle d'analyse de variance gène-spécifique

Une nouvelle approche est basée sur l'utilisation de deux modèles successifs d'analyse. Le premier modèle correspond à une étape de normalisation

qui permet de prendre en compte les effets des facteurs lame (A) et fluorochrome (D) : $y_{ijgr} = \mu + A_i + D_j + AD_{ij} + r_{ijgr}$.
Cette étape génère des résidus r_{ijgr} à partir des logarithmes des mesures d'intensité. Ces résidus sont utilisés pour modéliser les effets gène spécifique dans la seconde étape d'analyse. Le modèle gène spécifique s'écrit donc : $r_{ijgr} = G + VG_{ij} + DG_j + AG_i + \varepsilon_{ijr}$.

Cette décomposition permet d'optimiser de façon importante les calculs à effectuer. En effet, la méthode usuelle de construction des modèles d'ANOVA par les moindres carrés impose une étape d'inversion de la matrice d'expérience. Pour une expérience sur puces à ADN, cette matrice a une dimension de l'ordre de la dizaine de milliers et son inversion directe est délicate, voire impossible à réaliser dans certains cas. L'utilisation de deux étapes d'analyse permet donc de réduire les matrices d'expériences. De plus, dans une expérience de puce, le même jeu de gènes est représenté par définition sur l'ensemble des lames. Le facteur gène (G) est donc équilibré pour les facteurs lame (A) et fluorochrome (D). Cet équilibre permet finalement d'obtenir des estimations identiques pour le modèle d'analyse de variance global et le modèle en deux étapes (Wu *et al.*, 2002).

2.6.323 Tests de F

Le test de F est le test classique de l'ANOVA. Il s'agit simplement d'une généralisation du test de t pour la comparaison de plus de deux conditions. Comme pour le test de t, plusieurs variations du test de F existent pour le modèle d'ANOVA à effets fixes. Le test de F classique est appelé test F1. Mais il est également possible de calculer une variance commune à l'ensemble des gènes de l'expérience pour réaliser le test F3. Ce test est plus puissant que le test classique, mais il peut être biaisé si de nombreux gènes sont faiblement exprimés. Dans ce cas, il est préférable d'utiliser le test F2. Ce test est l'analogue du test de t régularisé et se base sur l'utilisation d'une combinaison pondérée de la variance globale et des variances spécifiques de chaque gène (Wu *et al.*, 2002 ; Cui et Churchill, 2003). Des abaques sont disponibles pour la réalisation du test F1, en revanche les valeurs correspondant aux tests F2 et F3 doivent être calculées par permutations.

Pour le modèle fixe, il n'existe qu'un seul terme de variance pour l'ensemble des facteurs et les effets des différents facteurs sont donc testés contre cette variance. Pour le modèle mixte en revanche les facteurs ne sont pas toujours testés contre les mêmes variances et le choix de l'une d'entre elles dépend de la question posée. Construire les tests statistiques se révèle plus compliqué que dans le cas des modèles fixes, mais peut être réalisé à l'aide de logi-

ciels spécifiques comme avec des bibliothèques disponibles pour SAS (SAS®MICROARRAY[9]) ou pour R (MAANOVA[10]).

Tous les tests statistiques présentés précédemment permettent d'obtenir une liste de gènes différentiellement exprimés qui peut être classée en fonction des valeurs de M, de t ou de F. Tenir compte dans ce classement de la valeur moyenne de A peut être un bon indicateur de l'importance à accorder aux valeurs observées (Yang et Speed, 2003). Mais même ainsi, il est difficile de placer un seuil de significativité permettant de retenir les gènes dont l'expression différentielle est la plus pertinente. Quels que soient le modèle et le test qui sont utilisés, la réalisation d'un test d'hypothèse pour chaque gène soulève le problème lié à l'utilisation des tests multiples.

2.6.4 Choix d'un seuil de significativité

Le nombre important d'hypothèses à tester simultanément sur les données de puces à ADN implique l'étude des deux erreurs associées aux tests multiples. La première est l'erreur *FWER* (*Family Wise Error Rate*) qui correspond à la probabilité de rejeter au moins une hypothèse nulle alors qu'elles sont toutes vraies. Il s'agit de l'équivalent pour les tests multiples de l'erreur de première espèce (α) associée à un test simple. La probabilité de commettre l'erreur *FWER* augmente de façon dramatique avec le nombre N de tests. Le second type d'erreur associée aux tests multiples est l'erreur *FDR* (*False Discovery Rate*), introduite par Benjamini et Hochberg (1995). Cette erreur correspond à la proportion d'hypothèses qui sont rejetées à tort.

Dans le cadre des tests simples, l'acceptation ou le rejet du test est basé sur le calcul de la valeur critique ou probabilité. Cette notion peut être généralisée dans le cadre des tests multiples avec le calcul de probabilités ajustées qui dépendent du type d'erreur à contrôler (Satagopan et Panageas, 2003).

2.6.41 Contrôle de l'erreur FWER

La méthode de Bonferroni (Bland et Altman, 1995) est la méthode la plus connue et la plus utilisée pour la calcul des probabilités ajustées. La probabilité ajustée est la probabilité obtenue pour un test simple dont l'erreur de première espèce est fixée à α/N. Cette méthode est simple, tout comme les nombreuses adaptations qui ont été proposées. Elle présente cependant l'inconvénient d'être difficilement applicable lorsque le nombre de tests est su-

[9] http://www.sas.com/industry/pharma/mas.html.
[10] http://www.jax.org/staff/churchill/labsite/software/anova/rmaanova.

périeur à dix et lorsqu'il existe des corrélations entre les statistiques de test. Cette méthode est donc beaucoup trop conservative pour les milliers de test liés aux expériences de puces (Satagopan et Panageas, 2003). Par ailleurs, il existe une corrélation forte entre les tests statistiques qui est liée à la co-régulation des gènes et à une dépendance des erreurs de mesure des niveaux d'expression. Pour pallier ce problème, Westfall et Young (1993) ont proposé une procédure dans laquelle les probabilités sont estimées par permutation. Cette procédure nécessite cependant un nombre important de données pour limiter la dispersion.

2.6.42 Contrôle de l'erreur FDR

Dans tous les cas, les méthodes contrôlant l'erreur *FWER* sont trop conservatrices pour des applications où peu de gènes sont différentiellement exprimés. Cependant, l'absence de correction implique souvent le risque d'obtenir un nombre élevé de faux positifs. En supposant par exemple que toutes les erreurs de première espèce α sont égales à 0,5 pour 1000 gènes étudiés, alors le nombre de faux positifs attendus est de 1000*0,05=50. Reiner *et al.* (2003) proposent un compromis en adoptant une démarche de contrôle du *FDR*, c'est-à-dire de la proportion de faux positifs parmi les gènes attendus comme différentiellement exprimés. De nombreuses procédures de contrôle du *FDR* peuvent être utilisées. La plupart sont basées sur l'utilisation d'algorithmes de re-échantillonnage comme la procédure proposée par Benjamini et Hochberg (1995). Actuellement de nombreux logiciels, tels que SAM (*Signifiance Analysis of Microarrays*) (Tusher *et al.*, 2001) ou Qvalue (Storey, 2004), permettent le calcul des probabilités ajustées en utilisant le contrôle du *FDR*. Toutes ces approches présentent l'intérêt de s'affranchir de l'hypothèse d'indépendance des gènes. Par ailleurs, elles ne nécessitent pas un nombre élevé de répétitions et sont moins conservatives que les procédures contrôlant le *FWER*.

Il semble donc que les procédures basées sur l'estimation du *FDR* offrent une bonne alternative à celles basées sur le *FWER* pour les analyses des données de puces. Cependant, des développements sont encore nécessaires pour que ce type de procédure s'applique parfaitement aux problématiques associées aux données d'expression, particulièrement en ce qui concerne leur aspect beaucoup trop conservatif (Yang et Speed, 2003).

2.6.5 Détermination de profils d'expression

Le but de cette partie n'est pas de dresser une liste exhaustive des nombreuses méthodes actuellement disponibles pour l'analyse des profils d'expression. Il s'agit plutôt de montrer comment certaines techniques d'analyse parmi les plus utilisées peuvent être applicables au problème de la

classification des données d'expression. De plus, les méthodes de classification sont les premières approches qui ont été développées pour l'analyse des données issues des puces à ADN. Aujourd'hui, les avancées récentes en technique d'apprentissage, et particulièrement les réseaux de neurones et les arbres de décision, ont démontré le potentiel de ces méthodes pour l'analyse des données d'expression (Holloway *et al.*, 2002).

2.6.51 Classification

Dans la majorité des cas, la première méthode d'analyse retenue est la méthode de classification hiérarchique. Cette technique, introduite par Eisen *et al.* (1998) permet de visualiser les données en utilisant des algorithmes de classification. Ces méthodes de classification sont utiles lorsque le but est de découvrir des groupes de gènes dans les données d'expression en l'absence d'informations extérieures.

Pour un plan d'expérience incluant plusieurs répétitions et différentes conditions expérimentales, les résultats sont organisés dans une matrice dite de niveau d'expression. Une ligne de cette matrice correspond à un gène unique, tandis que chaque colonne représente une condition particulière. Cette matrice est ensuite représentée sous forme graphique. La classification intervient alors pour regrouper et identifier les gènes présentant des profils d'expression similaires et qui peuvent être associés dans un même processus cellulaire (cf. **Figure A.2.12**).

La principale difficulté posée par l'utilisation de ce type de méthode est liée au fait que chaque processus biologique peut impliquer un nombre relativement faible de gènes et que la majorité de ceux qui sont présents sur la puce constitue un bruit de fond qui risque de masquer l'effet du sous-ensemble intervenant. De plus, chaque gène ne peut être placé dans la plupart des méthodes que dans un seul sous-ensemble de la classification, ce qui implique qu'un gène ne peut être associé qu'à un seul processus biologique pour une condition donnée. Une approche simple pour trouver les couples gène/condition qui conduisent à des classes « significatives » est de constituer des sous-matrices à partir des données initiales et d'appliquer la procédure classique de classification à chacune d'entre elles. Le nombre de sous-matrices augmentant exponentiellement avec la quantité de couples étudiés, de nombreuses approches heuristiques basées sur des procédés itératifs ont été développées (Getz *et al.*, 2000).

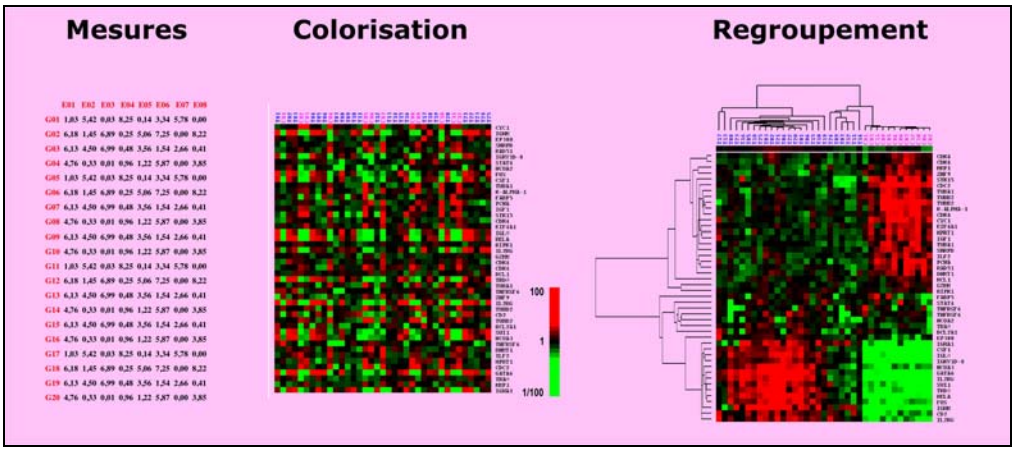


Figure A.2.12 Représentation graphique issue d'une analyse par classification hiérarchique. À partir des tableaux de mesures dans lesquels chaque colonne correspond à une expérience et chaque ligne à un gène (à gauche), chaque valeur est représentée par une couleur qui est le reflet qualitatif et quantitatif du rapport des fluorescences (au centre). Les gènes dont le niveau d'expression est inchangé sont représentés en noir, ceux dont le niveau augmente en rouge et les autres en vert. Une échelle de couleurs permet de quantifier les nuances de vert et de rouge. Finalement, les gènes ayant le même profil d'expression sur plusieurs expériences sont regroupés (à droite) (d'après Bertucci *et al.*, 2002).

D'un point de vue statistique, Alon *et al.* (1999) recommandent l'utilisation du coefficient de corrélation de Pearson comme mesure de la similarité des données entre gènes ou entre conditions expérimentales. Ce coefficient est en effet en accord avec la notion biologique de gènes co-exprimés et permet de mettre en évidence une similarité de « forme » entre deux profils d'expression, sans pour autant tenir compte des différences d'amplitude (Eisen *et al.*, 1998). La valeur du coefficient est donc élevée entre deux gènes qui sont impliqués par le même processus, même si chacun d'entre eux subit une activation (ou une répression) différente au cours du processus. Cependant, il présente l'inconvénient de donner une valeur de corrélation positive plus significative pour deux gènes fortement exprimés que pour deux gènes faiblement exprimés, car il ne prend pas en compte les niveaux absolus d'expression. Par ailleurs, lorsque gènes et conditions expérimentales sont pris simultanément en compte, il est préférable de calculer une distance euclidienne.

Ces méthodes de classification se révèlent particulièrement intéressantes pour déterminer la fonction des gènes inconnus en mettant en évidence des phénomènes de co-régulation entre des gènes de fonctions connues et des gènes dont les fonctions sont à déterminer (Tavazoie *et al.*, 1999). Cependant, l'expression de certains gènes n'est pas régulée au niveau transcriptionnel (Van

Helden *et al.*, 2000b). Dans ce cas, des expériences de délétion de gène menées par la suite peuvent apporter de plus amples détails sur chaque gène pris individuellement au sein d'un groupe co-régulé (Golub *et al.*, 1999) (cf 2.8.3).

Les méthodes de classification sont également très utilisées pour l'étude des séries de mesures temporelles et la reconstruction des voies métaboliques. Cependant, il existe encore de nombreuses difficultés pour ce type d'analyse. Elles sont liées autant à des raisons pratiques (erreurs de mesure, distance trop importante entre deux temps de la cinétique) qu'à des raisons conceptuelles (les changements relatifs de niveaux d'expression de deux gènes peuvent relever de la coïncidence plus que de la causalité). Il est donc essentiel de relier les résultats obtenus avec des connaissances *a priori* sur les interactions géniques (Darel et Képès, 2002). Mais au-delà de cette analyse par regroupement, le but est d'inférer le réseau génétique, c'est-à-dire de retrouver les interactions régulatrices sous-jacentes aux données d'expression en utilisant des modèles (Wolkenhauer, 2002).

2.6.52 Analyse en composante principale

Parallèlement aux méthodes de classifications, il est possible d'utiliser des méthodes comme l'analyse en composante principale (ACP) pour étudier les profils d'expression. Le travail de Chu *et al.* (1998) représente l'un des premiers exemples d'application de l'analyse en composante principale à des données temporelles d'expression. Alter *et al.* (2000) obtiennent des résultat similaires en utilisant une décomposition en valeur singulière (DVS) tout en montrant une plus grande créativité dans la présentation des résultats. Enfin, des modèles dynamiques de l'expression des gènes ont également été proposés à partir d'analyse en composante principale. Holter *et al.* (2000) estiment ainsi des matrices de translation indépendantes du temps pour montrer que des profils d'expression complexes peuvent être représentés par un petit nombre de sous-ensembles qui caractérisent les profils d'expression temporels.

Il est également possible de coupler ACP et méthode de classification au sein d'une même analyse. En effet, une des caractéristiques des données d'expression est que le nombre d'échantillons étudiés est typiquement plus petit que le nombre de gènes représentés sur la puce, ce qui viole les hypothèses classiques de la statistique (Satagopan et Panageas, 2003). Pour regrouper des échantillons présentant des caractéristiques similaires, il est donc impossible d'utiliser des modèles classiques, contrairement à ce qui est réalisé pour les gènes. Pour pallier ce problème, Ghosh et Chinnaiyan (2002) proposent un algorithme basé sur l'utilisation de modèles mixtes pour permettre la classification simultanée des gènes et des échantillons. Pour cela, une première étape utilisant une ACP permet de réduire la dimension des gènes, puis une seconde étape

permet la classification des échantillons. De la même manière, Culhane *et al.* (2002) proposent une analyse entre groupes de données en utilisant une méthode d'analyse de correspondance à la place d'un classement individuel des données.

Les méthodes de classification, tout comme les analyses en composante principale, permettent donc de classer les gènes par fonctions similaires ou en groupe de co-régulation et les échantillons pour un même état cellulaire ou encore un phénotype biologique identique. Ces résultats peuvent ensuite être reliés à des observations indépendantes et des connaissances extérieures, concernant par exemple les régulateurs, pour reconstruire le réseau métabolique (Alter *et al.*, 2000).

2.7 Relier les profils d'expression aux voies métaboliques

Les gènes n'interviennent jamais seuls dans un organisme mais dans des réseaux, et souvent en cascade. L'analyse des données issues des puces à ADN dans une perspective de modélisation des voies métaboliques, peut donc permettre la compréhension des systèmes vivants. Cette recherche des principes, qui dictent le comportement dynamique des systèmes de régulation de la cellule, nécessite une intégration des approches théoriques et expérimentales à des niveaux variés (Neves et Iyengar, 2002). Pour cela trois méthodes sont utilisées.

La première est une extension naturelle de l'analyse par classification. En effet, plusieurs gènes associés ensemble peuvent être co-régulés ou impliqués dans une même voie métabolique. L'analyse complémentaire des promoteurs associés à ce groupe de gènes peut souvent révéler les schémas de régulation (Pilpel *et al.*, 2001). Pour cela il existe de nombreux logiciels disponibles comme le panel d'outils *RSAT* (*Regulatory Sequence Analysis Tools)* développés par Van Helden *et al.* (2000a).

La seconde approche est une approche dite d'ingénierie inverse des voies métaboliques (D'haeseleer *et al.*, 2000 ; De Jong, 2002). Il s'agit d'identifier le réseau de régulation à partir des données transcriptomiques. Pour cela, il est possible de perturber de façon systématique et ciblée l'organisme étudié, au moyen de mutations ou de traitements chimiques par exemple (Hughes *et al.*, 2000). Il peut également être utile de réaliser des séries temporelles de mesures (Tavazoie *et al.*, 1999). Ces expériences sont conduites avec l'hypothèse que les perturbations ou les données temporelles permettront d'observer de nouvelles modifications des profils d'expression, qui sont néces-

saires à la reconstruction de l'architecture du réseau sous-jacent. Des méthodes diverses ont été proposées pour inférer les réseaux à partir de ce type de données. Il est possible d'utiliser une approche booléenne dans laquelle l'expression des gènes est modélisée par une valeur logique binéaire (Thieffry, 1999 ; Akutsu *et al.*, 2000 ; Thieffry et De Jong, 2002). L'utilisation d'un réseau bayésien est une autre approche permettant la modélisation des interactions entre les gènes (Husmeier, 2003). Pour cela les probabilités d'occurrence des différents modèles d'interaction entre gènes sont calculées (Friedman *et al.*, 2000). Cette approche couplée à la recherche de promoteurs a par exemple été utilisée pour reconstruire avec succès les voies métaboliques de la levure (Segal *et al.*, 2003). Lorsque des données temporelles sont disponibles, il peut également être intéressant d'utiliser des systèmes d'équations différentielles pour représenter les interactions métaboliques. Des logiciels comme *GEPASI* (Mendes, 1993 ; Mendes et Kell, 1998) permettent ce type de modélisation cinétique.

Enfin la dernière approche concerne la construction de cartes métaboliques couplant les données d'expression aux voies métaboliques (Leung et Cavalieri, 2003). Le logiciel *Osprey* permet ainsi de représenter graphiquement les gènes et leurs interactions avec des codes couleur associés à leur fonction (Breitkreutz *et al.*, 2003). Le logiciel libre *GenMAPP* (*Gene MicroArray Pathway Profiler*) permet de visualiser les gènes appartenant à un même groupe de co-régulation sur les voies métaboliques (Doniger *et al.*, 2003). Enfin le logiciel *Pathway Processor* permet de visualiser les voies métaboliques qui sont les plus affectées par des changements au niveau transcriptionnel (Grosu *et al.*, 2002).

L'inférence des réseaux métaboliques à partir des données issues des puces à ADN sera certainement l'un des défis majeurs des années à venir. Cependant l'évolution rapide de ce domaine d'étude, qui voit chaque jour l'apparition de méthodes nouvelles toujours plus complexes, perd parfois les chercheurs dans la compréhension des nouveaux logiciels disponibles et des méthodes innovantes qui leurs sont associées (Leung et Cavalieri, 2003).

2.8 Et après ?

2.8.1 Valider les données

Le processus de validation des données peut être divisé en trois étapes. La première est le contrôle de la qualité des données expérimentales, la seconde

est la confirmation indépendante des données et la troisième est une réflexion concernant l'universalité des résultats obtenus (Chuaqui *et al.*, 2002).

Le contrôle de la qualité des expériences est permis par l'intégration de répétitions au moment de la définition du plan expérimental. L'analyse de leur cohérence permet de valider la qualité des lames réalisées, d'évaluer les éventuelles variations expérimentales, voire d'éliminer certains plots ou certaines lames de l'analyse (cf. 2.6.21).

La confirmation indépendante des résultats peut être réalisée de plusieurs façons. La méthode *in silico* permet de comparer les résultats obtenus avec les informations disponibles dans la littérature et les bases de données. Elle offre l'opportunité de vérifier les données sans réaliser de nouvelles expérimentations. En ce qui concerne les méthodes de validation basées sur la conduite de nouvelles expériences, elles sont généralement menées avec les techniques de RT-PCR (Ye *et al.*, 2001), de PCR en temps réel et de *northern blotting* (Taniguchi *et al.*, 2001). La PCR en temps réel est la méthode qui est le plus souvent employée car elle permet de réaliser des mesures quantitatives très précises pour un ARN messager en particulier (Wang *et al.*, 2003). En plus de valider les niveaux d'expression avec les taux d'ARN messagers, il est également possible d'évaluer les quantités de protéines correspondantes par des méthodes immuno-chimiques ou par immunoblotting (Burton *et al.*, 2002).

Finalement la question de l'universalité des résultats est de savoir si les données obtenues correspondent à une description essentielle d'un état biologique donné. Cette validation est réalisée par la comparaison des résultats avec des données obtenues dans d'autres conditions physiologiques, ou pour des échantillons, des tissus ou encore des organismes différents (Chuaqui *et al.*, 2002).

2.8.2 Formaliser et stocker les données

Le développement rapide des puces à ADN a abouti à une accumulation sans précédent des données dans les laboratoires du monde entier. Leur stockage est donc devenu crucial (Stoeckert *et al.*, 2002) et nécessite le développement d'outils locaux et d'entrepôts publics de données accessibles à l'ensemble de la communauté scientifique. Pour pouvoir échanger et comparer ces données, même lorsqu'elles sont issues de technologies différentes, il est nécessaire de mettre en place une standardisation et un formalisme rigoureux au sein même de ces bases de données.

2.8.21 Formaliser les connaissances associées au transcriptome

Avant de stocker puis d'échanger les données d'expression entre les différents laboratoires, il est essentiel de s'assurer que tous décrivent bien les

mêmes objets. En effet, l'idéal serait que chaque gène ait un identifiant unique et que l'information correspondant à ses fonctions soit disponible dans les bases de données. Cependant la réalité est encore assez éloignée de cette situation (Holloway *et al.*, 2002). Pour améliorer les comparaisons entre données, il est important de travailler tant au niveau de l'annotation syntaxique des séquences, qui a pour but l'identification des gènes et de leurs éléments régulateurs, qu'au niveau de l'annotation fonctionnelle, qui s'attache à la détermination de leur fonction. Face à la quantité de séquences introduites chaque jour dans les bases de données, ces processus ont été fortement automatisés.

L'annotation syntaxique est un problème extrêmement complexe, surtout chez les eucaryotes dont les génomes contiennent une quantité élevée d'introns dans les séquences codantes et une multiplicité des systèmes de régulation (Lewis *et al.*, 2000). Un effort d'homogénéisation et de standardisation énorme a été fourni par la mise en place de bases de données. Malgré cela, les descriptions hasardeuses, contradictoires et parfois fausses rendent très difficilement exploitables ces informations par une machine (Iliopoulos *et al.*, 2003). De plus, ces données sont en pleine construction et l'arrivée massive d'informations en provenance des centres de séquençage génère une fluctuation très importante des informations (notamment au niveau des identifiants des gènes).

La principale méthode d'annotation fonctionnelle automatique consiste à rechercher pour chaque nouvelle séquence une homologie avec une autre séquence déjà annotée. Il faut rester conscient des problèmes induits par cette automatisation et rester lucide sur la qualité de l'annotation effectuée. En effet, le problème ne réside pas seulement dans la comparaison de séquences entre elles, mais surtout dans l'interprétation des résultats de cette comparaison. D'une part, les critères permettant de déterminer l'homologue le plus proche sont variés, et d'autre part certaines similitudes sont fortuites et des protéines très similaires peuvent posséder des fonctions très différentes (comme le montre l'exemple des gènes paralogues).

Afin que ces données en constante évolution soient utilisées de la même façon par l'ensemble de la communauté scientifique des vocabulaires contrôlés (ou ontologies[11]) organisés en base de connaissances ont été développés. Le consortium *Gene Ontology*[12] (Ashburner *et al.*, 2000) qui s'est mis en place en 1998 propose ainsi de produire une ontologie, applicable à tous les organismes, même si les connaissances associées aux gènes peuvent évoluer, et si

[11] http://smi-web.stanford.edu/projects/bio-ontology/.
[12] http://www.geneontology.org.

le rôle des protéines peut changer. Il existe par exemple un vocabulaire commun entre des bases de données aussi différentes que *FlyBase* (pour la Drosophile), *Saccharomyces Genome Database* (SGD), *Mouse Genome Database* (MGD), ce qui contribue au développement plus aisé d'outils d'intégration (Camon *et al.*, 2003).

2.8.22 Stocker les données

Une démarche similaire à celle qui a été utilisée pour formaliser les connaissances associées au transcriptome a été entreprise afin de faire face au manque de standard pour représenter et échanger les données d'expression (Brazma *et al.*, 2001). Elle a abouti à la création du document *MIAME* (*Minimum Information about Microarray Experiment*) défini par le groupe *MGED* (*Microarray Gene Expression Data Group*[13]). Le document *MIAME* permet de conserver les informations concernant le plan d'expérience, le plan de la puce, le protocole d'extraction et de préparation des échantillons, les conditions d'hybridation, les mesures d'intensités (images, quantification, filtration) et le type de normalisation utilisée. Bien qu'il ne soit pas encore entièrement formalisé, il sert de référence dans la communauté scientifique utilisant la technologie des puces à ADN et il est devenu essentiel pour publier dans de nombreux journaux (Holloway *et al.*, 2002). Des implantations XML de ce format (MAML et MAGE-ML) ont également été développées pour faciliter les échanges de données.

Il existe à présent de nombreux entrepôts publics pour les données transcriptomiques, comme la base *ArrayExpress*[14] de l'Institut de Bioinformatique Européen (*EBI*). Ils permettent de stocker les informations essentielles d'une expérience de puce sous le format *MIAME*. De plus, de nombreuses initiatives ont vu le jour dans les différents centres de recherche pour le stockage local des données. La plate-forme transcriptome de la Génopole Rhône-Alpes a par exemple développé la base de données SI-TRANS[15] (Système d'Information pour le Transcriptome) qui possède une interface facilitant la saisie des informations par les différents utilisateurs. Cette base est actuellement disponible et les premiers tests d'utilisation sont en cours.

[13] http://www.mged.org.
[14] http://www.ebi.ac.uk/arrayexpress/.
[15] http://liris.cnrs.fr/~sitrans.

2.8.3 Intégrer des données de nature différente

Des données génomiques, transcriptomiques, protéomiques et métaboliques deviennent disponibles pour de nombreux organismes. Il est donc nécessaire d'intégrer ces données de nature différente (Voit et Riley, 2003). Cet enjeu est d'autant plus crucial que les données issues des puces à ADN ne contiennent généralement pas assez d'information pour permettre la modélisation des réseaux de gènes de façon complète.

Les premières données qu'il semble naturel d'associer aux données transcriptomiques sont celle du protéome, en raison notamment de l'existence de nombreuses régulations post-traductionnelles. En effet, il n'existe dans certains cas qu'une corrélation partielle entre les concentrations de transcrits et de protéines (Greenbaum *et al.*, 2002 ; Ideker *et al.*, 2001b). Une étude comparant les quantités d'ARN messager obtenues à partir de données *SAGE* (*Serial Analysis of Gene Expression*) et les quantités de protéines obtenues sur gel 2D pour 150 protéines de la levure montre ainsi que, pour des quantités de transcrits restant identiques entre deux conditions, les quantités de protéines correspondantes peuvent être multipliées par vingt. Inversement, pour certaines protéines dont les quantités semblent identiques, une multiplication par trente de la quantité de transcrits est parfois observée (Gygi *et al.*, 1999). Cependant, une autre étude réalisée à plus grande échelle (sur 1400 protéines) laisse envisager une corrélation satisfaisante, bien qu'imparfaite, entre les abondances d'ARN messagers et de protéines (Futcher *et al.*, 1999). Si cette étude offre une certaine validation de l'utilisation de l'abondance des transcrits pour prédire l'abondance des protéines cellulaires, ou au moins leur abondance relative (4000 molécules de protéines sont représentées par une molécule unique d'ARN messager), elle montre également l'intérêt de coupler les données du transcriptome aux données du protéome. Cette combinaison de l'expression des gènes et des quantités de protéines a d'ailleurs été utilisée avec succès chez *Escherichia coli* pour prédire la fonction de nouveaux gènes et les comportements métaboliques (Corbin *et al.*, 2003).

Le second type de données à associer aux données transcriptomiques sont les données métaboliques disponibles dans des bases de données publiques telles que *KEGG* (Kanehisa *et al.*, 2002 ; Ogata *et al.*, 1999), *Ecocyc* et *MetaCyc* (Karp *et al.*, 2000). Des travaux ont déjà été réalisés dans cette voie chez *Escherichia coli* pour proposer un modèle mathématique dans lequel les associations entre gènes, protéines et réactions métaboliques sont représentées (Reed *et al.*, 2003).

Ce type d'étude combinant des données transcriptomiques, protéomiques et métaboliques est un outil puissant pour formuler de nouvelles hypothè-

ses nécessaires à une approche de biologie des systèmes (Urbanczyk-Wochniak *et al.*, 2003). Ces collections de données d'expression, associées aux données protéomiques et à des cartes métaboliques, ont été intégrées dans certaines bases de connaissances de façon à permettre l'interprétation des résultats à un niveau physiologique (Kanehisa *et al.*, 2002) et l'identification des principaux mécanismes impliqués dans les modifications métaboliques d'un organisme.

2.8.4 Puces et art !

Les exemples précédents mettent en avant la quantité de travail à réaliser après les expérimentations à proprement parler pour valider, stocker, échanger et finalement comparer les données obtenues avec les puces à ADN. Mais que faire des non moins nombreuses puces dont les données ne sont pas exploitables ? Il semble que certains aient trouvé la solution en développant un nouveau mouvement artistique autoproclamé *art arrays* (cf. **Figure A.2.13**)…

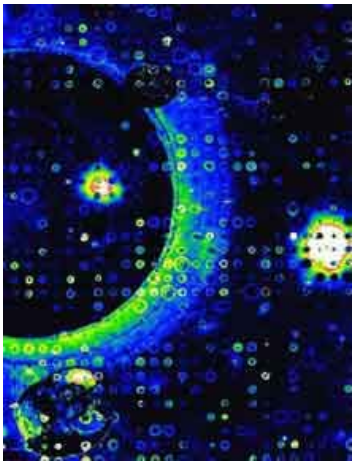


Figure A.2.13 Un exemple de l'*art arrays* intitulé « Cosmiqlow[16] » mais qui pourrait également s'appeler « la beauté de l'erreur ».

2.9 Les objectifs de la thèse du point de vue méthodologique

Les puces à ADN constituent une panoplie d'outils qui regroupe un vaste ensemble de systèmes différant par le type de support utilisé, la nature, la densité et le mode de synthèse des sondes, mais aussi par les conditions d'utilisation et de lecture des lames (Case Green *et al.*, 1998). Toutes ces tech-

[16] D'autres œuvres sont disponibles sur le site : http://perso.wanadoo.fr/waka/galerie/gl_index.htm.

niques présentent l'immense avantage de permettre l'analyse simultanée de plusieurs milliers de gènes, ce qui leur confèrent la capacité d'analyser l'expression des gènes d'un système biologique dans toute sa complexité. Cependant, à la lecture de ce chapitre, le néophyte pourrait avoir une vision très pessimiste de la technologie des puces à ADN avec cette énumération de contraintes, de biais et de limitations qui sont associés à chacune des étapes. Il est important de conserver à l'esprit que cette technologie, en plein développement, constitue une véritable révolution dans le domaine de la biologie moléculaire et le nombre croissant de travaux montrant une réelle pertinence dans l'analyse globale des génomes révèle bien son potentiel extraordinaire.

Leur évolution future dépendra beaucoup des solutions apportées aux deux principaux défis soulevés par leur utilisation. Le premier concerne le manque de standardisation des techniques et de formalisation des résultats, qui permettraient pourtant la confrontation de travaux réalisés indépendamment par différents laboratoires. La standardisation des techniques passe bien sûr par une standardisation des protocoles expérimentaux (aidée par le développement des documents MIAME), mais surtout par une planification expérimentale appropriée. La formalisation des résultats passe par l'établissement d'ontologies dédiées à la description des données d'expression et au développement d'outils d'analyse et de représentation. Le second enjeu repose sur le développement bioinformatique d'outils facilitant l'analyse de données, mais aussi la conception des puces.

Dans cette perspective, deux objectifs méthodologiques ont été fixés pour cette thèse. Le premier est le développement d'un logiciel destiné à optimiser l'une des étapes les plus délicates de la conception d'une puce : la détermination de sondes oligonucléotidiques intégrant l'ensemble des contraintes de l'étude. L'utilisation de ce logiciel est la première étape de conception d'une puce dédiée à *Buchnera*. Le second objectif est de définir les protocoles optimaux de préparation du matériel biologique et les méthodes d'utilisation, d'acquisition et d'analyses statistiques de cette puce dédiée, pour finalement permettre l'étude de l'expression des gènes du métabolisme des acides aminés chez *Buchnera*.

Partie B
Développements méthodologiques

1

1 Le logiciel ROSO

« L'homme n'est qu'un roseau, le plus faible de la nature ; mais c'est un roseau pensant. »
Blaise Pascal[17]

Un des aspects fondamentaux de la conception des puces à ADN concerne le choix des sondes à déposer sur les lames. En ce qui concerne la puce *Buchnera*, l'utilisation de sondes oligonucléotidiques a été préférée à celle de sondes d'ADNc fabriquées à partir de produits de PCR. En effet, leur production aurait été beaucoup plus difficile à envisager au laboratoire, notamment en l'absence de banques de gènes clonés pour la bactérie. Il était donc nécessaire de choisir des sondes pour l'ensemble du génome de *Buchnera* qui répondent aux deux contraintes principales associées aux choix multiples de sondes et définies par Bains dès 1994. Les sondes choisies doivent être spécifiques de chaque gène et compatibles entre elles en termes de rendement d'hybridation. Or, au moment de la mise en place du projet d'analyse du transcriptome de *Buchnera* (septembre 2000), il n'existait pas véritablement de logiciel intégrant ces deux critères. En effet, les seuls logiciels disponibles ne permettaient de choisir des sondes que pour l'étude d'un seul gène : *HYBsimulator* (Hyndman *et al.*, 1996), *Oligo®* (*MedProbe*, Saint Hanshaugen, Norvège) ou encore *Primer master* (Proutski et Holmes, 1996). Le seul logiciel qui semblait réellement capable de déterminer des sondes de façon simultanée pour des milliers de gènes était le logiciel commercial *ArrayDesigner®* (*Premier Biosoft International*, Palo Alto, CA, USA), Cependant, ce dernier n'était pas accessible gratuitement à la communauté scientifique et ne permettait pas encore de gérer l'étude de la spécificité en plus de celle des rendements d'hybridation des sondes.

Le premier objectif de cette thèse a donc tout naturellement été de concevoir le logiciel ROSO (Recherche et Optimisation de Sondes Oligonucléotidiques) qui intègre l'ensemble des paramètres d'optimisation des sondes oligonucléotidiques, pour doter la plateforme transcriptome Rhône-Alpes d'un vé-

[17] ascal, B. Pensées. Gallimard, Paris (1670).

ritable outil d'aide à la conception des puces à ADN. Ce travail se situe au centre d'une double problématique qui intègre une question directement liée à la biologie concernant les nombreux critères de choix d'une « bonne sonde » oligonucléotidique et une question bioinformatique abordant l'implémentation de ces différents paramètres au sein d'un logiciel.

Au sein de ce chapitre, la présentation de ROSO est divisée en deux parties. La première décrit le développement du logiciel et la seconde présente l'utilisation de ROSO à partir de l'exemple du choix des sondes dédiées à *Buchnera*.

1.1 De la problématique du choix des sondes au développement du logiciel ROSO

1.1.1 Matériel et méthodes : comment choisir une « bonne sonde » ?

L'objectif de cette partie est de définir les critères essentiels pour le choix multiple de sondes (Mitsuhashi, 1996). Le premier de ces critères est lié à la spécificité des sondes. En effet, il est nécessaire de rechercher au sein de chaque gène des régions qui possèdent le plus faible taux d'identité entre elles, l'idéal étant de sélectionner des régions uniques dans le génome. Cette spécificité permet de limiter les phénomènes d'hybridation aspécifique sur les sondes. Le second critère concerne le rendement d'hybridation. Toutes les sondes doivent en effet posséder un même profil thermodynamique, c'est-à-dire une même température de fusion (ou Tm), qui permet de définir la température de la réaction d'hybridation (Santalucia *et al.*, 1996). Enfin le troisième critère est basé sur la structure des sondes qui influence la stabilité des hybridations. Il s'agit d'une part de vérifier l'absence de structures secondaires stables sur la sonde (homodimère et épingle à cheveux) qui gêneraient la réaction d'hybridation et d'autre part de définir différents critères assurant une stabilité optimale des hybrides cibles/sondes.

1.1.11 Étude de la spécificité

La recherche de séquences homologues n'est pas un problème nouveau en bioinformatique. En effet, de nombreux outils ont été développés et notamment pour répondre aux besoins de la phylogénie. Il est donc possible d'utiliser ces outils pour la recherche de régions identiques entre les différents gènes d'un organisme. Par défaut, les régions qui restent sont des régions spécifiques dans lesquelles il est possible de choisir des sondes.

Pour effectuer cette recherche le logiciel d'alignement *BLAST* (*Basic Local Alignment Search Tool*) (Altschul *et al.*, 1990 et 1997) a été retenu en raison de sa rapidité et de la disponibilité du parseur associé *DUBLASTN* développé par Laurent Duret au Laboratoire de Biométrie et Biologie Evolutive (BBE) de l'Université Claude Bernard de Lyon.

1.1.111 Présentation des logiciels BLAST et DUBLASTN

Le logiciel *BLAST* est un programme de recherche de similarité développé au NCBI. La conception de l'algorithme est basée sur un modèle statistique établi d'après les travaux de Karlin et Altschul (1990 et 1993). L'unité fondamentale de *BLAST* est le *HSP* (*High-scoring Segment Pair*). Un *HSP* correspond à une région de similitude la plus longue possible entre deux séquences, ayant un score supérieur ou égal à un score seuil. Un deuxième score *MSP* (*Maximal-scoring Segment Pair*) a été défini comme étant le meilleur score obtenu parmi tous les couples possibles que peuvent produire deux séquences. Les méthodes statistiques sont alors appliquées pour déterminer la signification biologique des *MSP,* et par extrapolation, la signification des scores *HSP* obtenus lors de la comparaison. Ce logiciel possède en fait quatre programmes distincts de comparaison de séquences avec les bases de données. *BLASTN* (séquence nucléique contre base nucléique), *BLASTP* (séquence protéique contre base protéique), *BLASTX* (séquence nucléique traduite en six phases contre base protéique), et *TBLASTN* (séquence protéique contre base nucléique traduite en six phases). Pour la recherche de spécificité qui nous intéresse, seul *BLASTN* est utilisé. Ce logiciel et les utilitaires associés sont disponibles sur le site de téléchargement du NCBI[18] pour presque tous les systèmes d'exploitation.

La stratégie de *BLASTN* consiste à rechercher tous les mots de longueur W dans la séquence (W=11 par défaut), comparer ces mots avec les séquences de la banque afin d'identifier les homologies exactes (les *hits*), puis réaliser une extension du segment, quand cela est possible. Cette extension est réalisée à partir du mot commun dans les deux directions le long de chaque séquence, de manière à ce que le *score* cumulé puisse être amélioré. L'extension est arrêtée dans les trois cas suivants :

- le *score* cumulé descend d'une quantité x donnée par rapport à la valeur maximale qu'il avait atteint,
- le *score* cumulé devient inférieur ou égal à zéro,
- la fin d'une des deux séquences est atteinte.

[18] ftp://ftp.ncbi.nih.gov/*BLAST*/executables/.

La signification des alignements est ensuite évaluée statistiquement en fonction de la longueur et de la composition de la séquence, de la taille de la banque et de la matrice de *scores* utilisée. Le fichier de sortie contient une liste de séquences ayant un alignement significatif, c'est-à-dire les séquences ayant une valeur *e* inférieure au seuil fixé lors de la recherche. Ces résultats n'ont *a priori* aucune signification biologique. Pour chacune de ces séquences, deux paramètres sont calculés pour évaluer la qualité de l'alignement. Le premier est le *score S'* qui est dérivé du score brut de l'alignement. Il a été normalisé et peut donc être utilisé pour comparer des *scores* provenant de recherches différentes. Le second est la valeur *e* qui correspond au nombre d'alignements différents qui peuvent potentiellement exister dans les banques avec un *score* supérieur ou égal à *S'* (c'est-à-dire la probabilité d'observer au hasard ce score à travers la banque de séquences considérée). L'alignement est donc d'autant plus significatif que la valeur *e* est faible.

Les fichiers contenant les séquences à étudier doivent être au format FASTA. Il s'agit d'un format très courant dans lequel une description simple de la séquence est donnée sur une première ligne qui est suivie des lignes contenant la séquence. La ligne de description se distingue des lignes de séquence en étant toujours précédée du symbole « > » (cf. **Figure B.1.1**).

```
>refgb|NM_005455|NM_005455 Homo sapiens zinc finger protein 265 (ZNF265), mRNA
ATGTCGACCAAGAATTTCCGAGTCAGTGACGGGGACTGGATGCAAAA AATGTGGAAATGTAAACTTTGCTAGAAGAA
CCAGGTAATCGATGTGGTCGGGAGATGTCGACCAAGAATTTCCGAGTCAGTGACGGGGACTGGATGCAAAA AATGTG
GAAATGTAAACTTTGCTAGAAGAA

>NM_004441
GAATTCCACATGCACACCCACACCCACGCGCGCCCGCACC-GCCCCACGCGCACACACTCCTGCCCCAGGTAATCGAT
GTGG
```

Figure B.1.1 Exemple de deux séquences au format FASTA.

La base contre laquelle sont testées les séquences d'intérêt nécessite un formatage qui est réalisé avec l'outil *formatdb* et l'option de formatage -p F (pour des séquences nucléotides). L'utilisation à proprement parler de l'outil *BLASTN* est alors possible avec différents paramètres (cf. **Tableau B.1.1**). Leurs valeurs par défaut permettent d'arrêter les extensions d'alignement et masquent une partie des résultats de façon à convenir aux recherches les plus classiques d'homologie (Mcginnis et Madden, 2004).

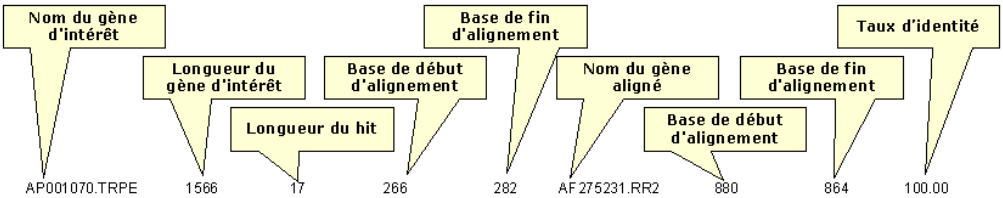
Il est important de noter que par défaut, les filtres détectant les séquences de faible complexité ne sont pas activés (paramètre -F F). Or la complexité des séquences est un facteur important pour les phénomènes d'hybridation. En effet, une faible complexité (une séquence présentant par exemple la répétition AATAATAAT) favorise une hybridation imparfaite entre

la sonde et sa cible et accroît les risques d'hybridation croisée avec les autres cibles (Bozdech *et al.*, 2003 ; Wootton et Federhen, 1996).

Tableau B.1.1 Récapitulatif des principaux paramètres du logiciel *BLAST* avec leurs valeurs par défaut. La recherche par défaut est réalisée sans éliminer les séquences de faibles complexité (-F F), sur les deux brins d'ADN (-S 3), avec un mot de 11 bases (-W 11) et des valeurs de gain et de pénalité au moment de l'extension respectivement de 1 et –3 (–r 1 -q –3). La taille de la base est adaptée automatiquement à la base fournie en entrée (-z 0) et les résultats pour lesquels la valeur de e est inférieure à 10 (-e 10) ne sont pas affichés.

Symbole	Signification	Valeur par défaut
-F	Filtre de faible complexité	F
-S	Brin sur lequel est effectuée la recherche	3
-W	Taille du mot recherché	11
-r	Gain pour un appariement	1
-q	Pénalité pour un mésappariement	-3
-e	*Expectation value*	10
-z	Taille de la base	0

Le parseur *DUBLASTN* permet de lancer *BLASTN* directement et possède trois critères permettant de définir les critères d'affichage des *hits* repérés par *BLAST* en fonction des intérêts de l'utilisateur. Le fichier de sortie obtenu est beaucoup plus facilement manipulable que les fichiers de sortie classiquement proposés par *BLASTN* pour un grand nombre de séquences (cf. **Figure B.1.2**). Les deux premiers critères permettant d'adapter l'affichage sont la taille minimale du *hit* et le taux d'homologie minimal. Lorsque la taille du gène d'intérêt est inférieure à la longueur minimale du *hit*, le *hit* est conservé uniquement lorsqu'il possède une taille supérieure à un certain pourcentage de la longueur du gène d'intérêt. Ce pourcentage est le troisième critère à définir.



AP001070.TRPE	1566	17	266	282	AF275231.RR2	880	864	100.00
AP001070.TRPE	1566	17	266	282	AF275231	1387	1371	100.00
AP001070.TRPE	1566	16	131	146	SPI298678	185	200	100.00
AP001070.TRPE	1566	16	128	143	AF275251.RR2	489	474	100.00
AP001070.TRPE	1566	16	128	143	AF275251	1001	986	100.00
AP001070.TRPE	1566	16	128	143	AF275250.RR2	489	474	100.00
AP001070.TRPE	1566	16	128	143	AF275250	953	938	100.00
AP001070.TRPE	1566	20	266	285	AF275247.RR2	858	877	95.00
AP001070.TRPE	1566	18	128	145	AF275247.RR2	493	476	94.44
AP001070.TRPE	1566	20	266	285	AF275247	1315	1334	95.00
AP001070.TRPE	1566	18	128	145	AF275247	950	933	94.44
AP001070.TRPE	1566	20	128	147	AF275245.RR2	497	478	95.00
AP001070.TRPE	1566	20	128	147	AF275245	1006	987	95.00
AP001070.TRPE	1566	16	128	143	AF275244.RR2	507	492	100.00
AP001070.TRPE	1566	16	128	143	AF275244	1019	1004	100.00
AP001070.TRPE	1566	16	128	143	AF275243.RR2	507	492	100.00
AP001070.TRPE	1566	16	128	143	AF275243	1017	1002	100.00
AP001070.TRPE	1566	16	128	143	AF069113	88	103	100.00
AP001070.TRPE	1566	20	128	147	AF069103	514	495	95.00
AP001070.TRPE	1566	20	128	147	AF069100	516	497	95.00
AP001070.TRPE	1566	20	128	147	AF069097	577	558	95.00
AP001070.TRPE	1566	16	343	358	AB035886.RR2	933	948	100.00

Figure B.1.2 Exemple d'un fichier de sortie de *DUBLASTN*.

1.1.112 Détermination des paramètres d'utilisation de BLASTN et DUBLASTN

Sur le principe de base présenté précédemment, la recherche de similarité peut donc être adaptée à la problématique de l'utilisateur à l'aide des différents paramètres. Le site de téléchargement du NCBI comporte une documentation précise sur l'utilisation de *BLASTN* et ses options, mais aucun détail n'est fourni en ce qui concerne l'efficacité de la recherche en fonction des valeurs des paramètres. Il est d'ailleurs indiqué qu'il n'est pas possible de fournir un guide théorique clair de l'utilisation des paramètres les plus appropriés et que leurs valeurs par défaut ont été sélectionnées de façon empirique au cours des années d'utilisation. Pour réaliser l'étude de similitude de façon optimale, il a été nécessaire de définir les valeurs les plus adaptées des différents paramètres de *BLAST* et de *DUBLASTN*.

Des études expérimentales montrent que le seuil de sensibilité d'hybridation des cibles et des sondes en termes d'identité minimale de séquence est de 70 % sur 20 paires de bases. En effet, pour ces valeurs, le risque d'hybridation aspécifique n'est pas significatif (Richmond *et al.*, 1999). De plus Hughes *et al.* (2001) ont démontré qu'une identité de 70 % quelle que soit la taille des sondes était suffisante pour réduire l'hybridation aspécifique au niveau du bruit de fond. Kane *et al.* (2000) ont également montré que des séquences de 50 bases avec des taux de similitude de 75 %, ou une séquence d'au moins 15 bases identiques peuvent être responsables d'une hybridation aspécifique.

Il est donc nécessaire de déterminer des paramètres permettant de détecter ce seuil de sensibilité minimal de 70 % d'identité sur 20 paires de bases. Pour cela, différents types de fichiers de séquences au format FASTA ont été créés avec le logiciel R. Ces fichiers contiennent tous des ensembles de séquences simulées de 1000 bases. Les fichiers *bloc35*, *bloc50* et *bloc100* contiennent des paires de séquences avec un fragment similaire de respectivement 35, 50 et 100 bases situé en position 500. Les séquences de ce fragment possèdent des taux de divergence compris entre 5 et 60 % et le reste des deux séquences ne contient pas d'identité détectable. Les deux fichiers *mut1* et *mut2* contiennent des paires de séquences avec un taux de divergence croissant de 1 à 60 %. Les paires ne sont pas homologues entre elles. Enfin, les fichiers *grad1* et *grad2* contiennent deux couples de séquences dont le taux d'identité décroît de façon graduelle le long de la séquence (de 1 à 50 % de divergence avec un pas de 2 %) par tranche de 42 bases.

Un ajustement empirique des paramètres de *BLASTN* a abouti aux choix des valeurs suivantes : W=7, z=10^6 et r=2 qui permettent de détecter des séquences identiques de 100, 50 et 35 bases présentant jusqu'à 70 % d'identité de façon correcte (cf. **Tableau B.1.2**). Ces valeurs permettent également de détecter des similitudes comprises entre 80 et 60 % sur des séquences de 1000 bases (cf. **Tableau B.1.3**) et des similitudes graduelles sur pratiquement l'ensemble de la séquence (cf. **Tableau B.1.4**), contrairement aux valeurs par défaut avec lesquelles l'extension de l'alignement est arrêtée de façon précoce.

De la même façon, les valeurs des paramètres de *DUBLASTN* suivantes ont été retenues : 30, 0,80 et 0,70 (respectivement taille, taux d'identité et pour les séquences de taille inférieure à 30 bases, pourcentage minimum de la séquence ayant une identité d'au moins 80 %). Elles permettent de ne conserver que les *hits* au-dessus du seuil de sensibilité.

Tableau B.1.2 Résultats des recherches de *BLASTN* avec les fichiers de séquences contenant une sous-séquence de 35, 50 et 100 bases avec différents taux de similitude. Les colonnes 3 et 4 présentent les résultats obtenus avec les paramètres par défaut et les colonnes 5 et 6 ceux qui sont obtenus avec les paramètres ajustés. ND signifie que la sous-séquence n'a pas été détectée.

Fichiers	Taux d'identité réel (%)	Taille (pb) et taux d'identité (%) (paramètres par défaut)		Taille (pb) et taux d'identité (%) (paramètres ajustés)	
bloc35	95	36	97	36	97
	90	30	97	34	94
	85	36	86	36	86
	80	17	94	39	82
	70	ND	ND	37	72
bloc50	95	50	98	53	96
	90	42	95	46	93
	85	36	97	43	93
	80	38	87	52	82
	75	25	92	39	84
	70	11	100	39	82
	65	ND	ND	22	81
bloc100	95	100	96	100	96
	90	101	90	101	90
	85	79	91	103	85
	80	60	87	103	80
	75	56	86	94	77
	70	28	89	95	72
	65	13	100	57	77

Tableau B.1.3 Résultats des recherches de *BLASTN* avec les fichiers contenant des séquences de 1000 bases avec différents taux d'identité. À partir de 75 % d'identité, les paramètres par défaut ne permettent plus une détection correcte de la séquence de 1000 bases (détection uniquement d'une séquence de 72 bases à 95 % d'identité). Pour des taux d'identité inférieur 70 % seuls les paramètres ajustés permettent de détecter la séquence et sont donc présentés dans le tableau.

Fichiers	Taux d'identité réel (%)	Taille (pb) et taux d'identité (%) (paramètres ajustés)	
mut1	70	1000	70
	65	730	68
	61	608	65
	60	486	63
mut2	70	1000	73
	65	815	71
	60	695	70

Tableau B.1.4 Résultats des recherches de *BLASTN* avec les fichiers contenant des séquences de 1000 bases avec une divergence graduelle toutes les 42 paires de bases comprise entre 1 et 50 %.

Fichiers	Taille (pb) et taux d'identité (%)) (paramètres par défaut)		Taille (pb) et taux d'identité (%) (paramètres ajustés)	
grad1	520	89	812	82
grad2	433	91	854	80

1.1.12 Étude du rendement d'hybridation entre deux brins d'ADN

Actuellement, le modèle le plus complet pour étudier les rendements d'hybridation des sondes sur les cibles est le modèle thermodynamique dit du plus proche voisin. Au sein d'une molécule d'ADN double brin, ce modèle définit des interactions de voisinage entre deux nucléotides successifs qui permettent de prendre en compte à la fois la nature et la place des nucléotides (Williams *et al.*, 1994). Ces paires de nucléotides sont associées à des valeurs distinctes d'enthalpie, d'entropie et d'énergie libre pour l'association des deux brins d'ADN (Doktycz *et al.*, 1995). Récemment Santalucia (1998) a proposé des valeurs unifiées de ces énergies calculées d'après l'ensemble des études menées auparavant (cf. annexes 1.2). Ces valeurs sont également connues pour huit types de mésappariements possibles : G.T (Allawi et Santalucia, 1997), A.C (Allawi et Santalucia, 1998a), G.A (Allawi et Santalucia, 1998b), C.T (Allawi et Santalucia, 1998c), G.G, C.C, A.A et T.T (Peyret *et al.*, 1999). Elles ont été obtenues à partir de l'étude expérimentale des courbes de fusion (cf. **Figure B.1.3**) de différents oligonucléotides dont les séquences contenaient toutes les possibilités d'interactions de voisinage en position centrale. Des études complémentaires ont montré que le marquage des nucléotides ne modifie pas la valeur de ces énergies. Ce modèle thermodynamique peut donc s'appliquer à l'analyse des hybrides formés entre sondes et cibles marquées, qui sont utilisées pour les expériences menées sur puces à ADN (Griffin et Smith, 1998).

1.1.121 Détermination de la température de fusion

La température de fusion ou Tm (*Temperature of melting* ou plus exactement *Temperature of mid-transition*) (Mergny et Lacroix, 2002) est une mesure directe de la stabilité de l'association de deux brins d'acides nucléiques en solution (cf. **Figure B.1.3**). Il s'agit de la température à laquelle la moitié des acides nucléiques est sous forme double brin. En l'absence d'agents déstabilisants, comme la formamide ou l'urée, le Tm dépend principalement du taux de GC, de la concentration en acides nucléiques et de la concentration en sels. Ces trois facteurs augmentent en effet le Tm en stabilisant l'hybride formé.

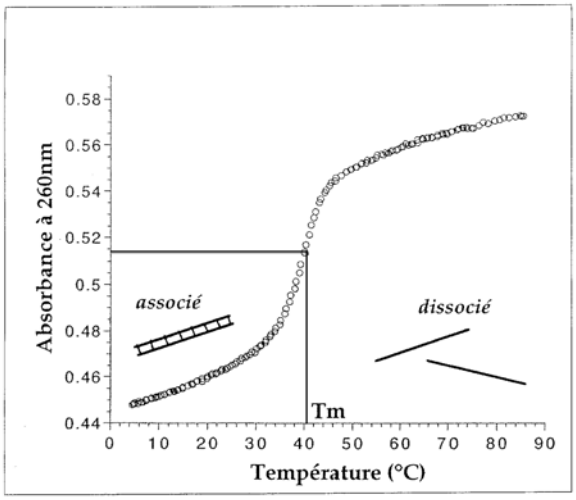


Figure B.1.3 Exemple de la courbe de fusion obtenue pour un oligonucléotide (d'après Mergny et Lacroix, 2002). La dénaturation d'une molécule d'ADN s'accompagne d'une augmentation de l'absorption lumineuse à 260 nm appelée aussi effet hyperchrome (Chester et Marshak, 1993). Cette dénaturation s'effectue généralement dans une zone de température restreinte dont le point médian correspond à la température de fusion ou Tm.

En fonction de la taille de la molécule étudiée, deux modèle différents sont utilisés pour le calcul du Tm.

Pour les petites molécules, le modèle thermodynamique du plus proche voisin offre la meilleure précision pour le calcul de la température de fusion car il permet de prendre en compte à la fois la nature et la place des nucléotides au sein de la séquence :

$$Tm(°C) = \frac{\Delta H}{\Delta S + R\ln(C_T)} + 16{,}6\log\left(\frac{[K^+]}{1+0{,}7[K^+]}\right) - 273{,}15$$

R représente la constante des gaz parfaits (R=1,987 cal. K-1.mol-1) et C_T la concentration totale en acides nucléiques appariés (à remplacer par $C_T/4$ pour les séquences présentant des mésappariements) (Freier *et al.*, 1986).

En accord avec le modèle du plus proche voisin, le calcul de l'enthalpie (ΔH) et de l'entropie (ΔS) d'hybridation peut être décomposé en plusieurs termes additifs (Breslauer *et al.*, 1986 ; Santalucia, 1998) :

$$\Delta H_{total} = \Delta H_{initiation} + \Delta H_{symétrie} + \sum_x \Delta H_x$$

$$\Delta S_{total[Na^+ 1M]} = \Delta S_{initiation} + \Delta S_{symétrie} + \sum_x \Delta S_x$$

L'introduction d'un paramètre d'initiation a été proposée par Sugimoto *et al.* (1996) puis reprise par Santalucia (1998). Sa valeur dépend de la présence

ou non de bases G et C dans la séquence. Quant au terme de symétrie, il intervient uniquement lorsque l'hybride formé est parfaitement homologue (cf. annexes 1.1). Enfin $\sum \Delta H_x$ (et ΔS_x) représentent la somme des x interactions de voisinage de la séquence.

L'enthalpie est indépendante de la concentration en sels. En revanche, il existe un terme correctif pour l'entropie. Il permet de prendre en compte une concentration en ions sodium différente de 1M (Schütz et Von Ahsen, 1999) pour une séquences contenant N nucléotides :

$$\Delta S_{total} = \Delta S_{total\left[Na^{+}1M\right]} + 0,368(N-1)\ln\left[Na^{+}\right]$$

Lorsque les séquences étudiées dépassent cinquante nucléotides, elles ne présentent rarement que deux états de transition (double ou simple brin), ce qui réduit considérablement la précision des prédictions de la méthode du plus proche voisin (Von Ahsen *et al.*, 1999). Il est donc préférable d'utiliser la formule de Baldino *et al.* (1989) qui prend en compte le taux de GC (%GC), le taux de mésappariements entre les deux brins (%mésappariements) et le nombre N de nucléotides de la séquence :

$$Tm(°C) = 81,5 + 0,41(\%GC) + 16,6\log\left(\frac{\left[K^{+}\right]}{1+0,7\left[K^{+}\right]}\right) - (\%mésappariements) - \frac{500}{N}$$

1.1.122 Énergie libre de formation des structures secondaires

Les sondes peuvent former des structures secondaires qui risquent de gêner la réaction d'hybridation. Ces structures sont d'autant plus stables que la valeur de leur énergie libre de formation est négative (Southern *et al.*, 1999). L'énergie libre est une mesure de la stabilité de l'hybridation entre deux brins qui dépend de la température (T) (Griffin et Smith, 1998) et de la concentration en sels du milieu (Peyret *et al.*, 1999). Le modèle thermodynamique du plus proche voisin permet de calculer les énergies libres d'hybridation intra ou inter-séquences de façon précise pour une concentration en sodium de 1M et de corriger la valeur obtenue en fonction du nombre N de nucléotides de la séquence pour des concentrations en sodium différentes :

$$\Delta G_{total\left[Na^{+}1M\right]} = \Delta H - T\Delta S = \Delta G_{initiation} + \Delta G_{symétrie} + \sum_{x} \Delta G_{x}$$

$$\Delta G_{total} = \Delta G_{total\left[Na^{+}1M\right]} - 0,0114\,N\ln\left[Na^{+}\right]$$

En pratique, les sondes peuvent former des épingles à cheveux (la sonde se replie sur elle-même) et des homodimères (association tête-bêche de deux sondes identiques) qui sont à éviter pour une réaction d'hybridation. Pour que la formation d'une épingle à cheveux soit possible, sa tige doit posséder au

minimum deux liaisons entre bases homologues et sa boucle doit contenir au moins trois bases (Groebe et Uhlenbeck, 1988 ; Rychlik et Rhoads, 1989). En ce qui concerne les homodimères, leur formation est possible uniquement s'il existe au moins deux liaisons successives entre bases homologues.

Figure B1.4 Exemple de formation d'un homodimère et d'une épingle à cheveux.

1.1.123 Autres critères thermodynamiques d'intérêt pour le choix des sondes

De nombreuses données expérimentales montrent qu'il est possible de définir d'autres critères d'intérêt influençant la stabilité de l'hybridation. Ces critères sont beaucoup moins importants que la valeur du Tm et l'absence de structures secondaires. Quatre d'entre eux ont cependant été retenus pour le développement du logiciel ROSO afin de départager des sondes répondant déjà aux deux critères précédents.

Énergie libre de formation des pentamères situés aux extrémités des sondes

Les pentamères situés aux extrémités de la sonde semblent jouer un rôle important dans l'accrochage des cibles et influencent la stabilité de l'hybride qui en résulte. En effet, des ancrages GC aux extrémités des sondes stabilisent la formation des hybrides et permettent d'obtenir de meilleurs rendements d'hybridation (Mitsuhashi, 1996). Ces ancrages sont associés aux valeurs les plus négatives des énergies libres d'hybridation des pentamères situés aux extrémités des sondes.

Taux de GC

La stabilité d'un ADN double brin dépend du nombre de liaisons hydrogène formées. Les séquences riches en GC sont donc les plus stables (Von

Ahsen *et al.*, 1999). En pratique un taux de GC compris entre 40 et 65 % semble assurer un rendement d'hybridation optimal et limite les risques d'hybridation aspécifique (Luebke *et al.*, 2003 ; Talla *et al.*, 2003).

Nature des bases situées aux extrémités des sondes

Pour des sondes de composition en bases identiques, les rendements d'hybridation les plus faibles sont observés lorsqu'une base A (ou T) est présente à chacune des deux extrémités et réciproquement les rendements les plus élevés sont obtenus avec des bases G (ou C) (Southern *et al.*, 1999 ; Xia *et al.*, 1998). En effet, au cours de la réaction d'hybridation un processus dit « de fermeture éclair » (*zippering process*) débute sur la base située à l'extrémité de la sonde et semble facilité par la présence d'une base G ou C (Maskos et Southern, 1993). De plus il a été montré expérimentalement que la présence de telles bases limite les risques de dénaturation (Williams *et al.*, 1994).

Absence de séquences contenant trois G ou trois C successifs

La présence, au sein d'une séquence, de trois bases G (ou C) successives est à l'origine d'un biais dans le calcul des énergies d'interaction car la valeur obtenue est beaucoup plus faible que celle qui est mesurée expérimentalement. Ce type de séquence est donc à éviter lorsque la méthode du plus proche voisin est appliquée (Santalucia *et al.*, 1996 ; Williams *et al.*, 1994).

1.1.13 Choix informatiques

Le langage de programmation qui a été utilisé est le langage C. Il est né en 1972 des travaux de Brian Kernighan et Dennis Ritchie et a été normalisé en 1989 par le comité de l'*ANSI* (*American National Standards Institute*). Son principal avantage est sa très grande portabilité liée à l'absence de dialectes et à la disponibilité de compilateurs pour la plupart des machines. Le travail de développement et de compilation a été réalisé sur les plateformes Windows, Unix (MacOSX), Irix (Silicon Graphics) et Solaris (Sun). Son second atout réside dans sa puissance, liée à l'existence de bibliothèques quasi-normalisées, mais aussi à la souplesse de ses algorithmes et à l'extrême richesse de ses expressions.

Plusieurs fichiers sources et un fichier *Makefile* ont été utilisés au cours du développement. Ils permettent de créer l'exécutable *roso*. Les deux premiers fichiers nommés *roso.h* et *constantes.h* (h pour *header*) correspondent à l'interface et contiennent les définitions et les déclarations de tous les objets : en-têtes des fonctions, types, variables, et structures pour le premier et constantes pour le second. Le troisième, appelé *roso.c* correspond à la réalisation et contient les objets privés et le corps des fonctions (cf. **Figure B.1.5**).

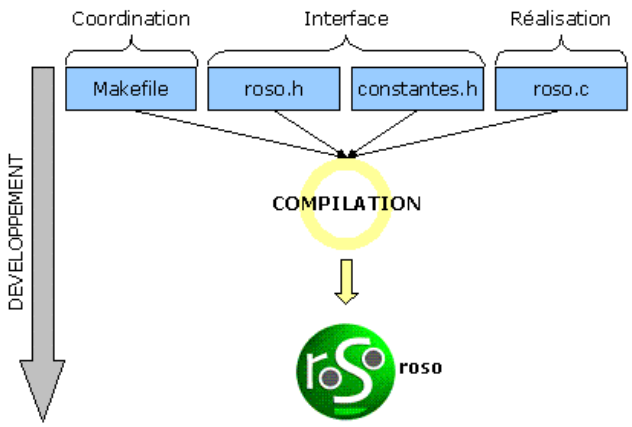


Figure B.1.5 Présentation des différents fichiers utilisés au cours du développement. Ils sont nécessaires à la création de l'exécutable *roso*.

1.1.2 Résultats : la conception

La conception du logiciel ROSO a été réalisée en deux étapes successives. Une étape préliminaire a permis de définir l'architecture générale du programme avec un découpage en unités fonctionnelles formant les différents modules du logiciel. Une étape de conception détaillée a ensuite été réalisée avec l'écriture des algorithmes et des fonctions. Pour faciliter la lecture, cette partie est présentée avant tout du point de vue de l'utilisateur, et laisse de côté les aspects techniques qui intéressent plutôt le concepteur. Cette présentation suit donc la chronologie d'utilisation de ROSO qui est divisée en trois parties distinctes. Une partie préliminaire permet la création des fichiers nécessaires à partir des séquences de l'utilisateur. La partie suivante permet la recherche de régions spécifiques dans les séquences et la dernière permet la recherche de sondes optimales.

1.1.21 Les fichiers d'entrée

L'utilisation de ROSO requiert un ou deux fichiers contenant des séquences au format FASTA. Le premier fichier contient les séquences des gènes d'intérêt, c'est-à-dire les séquences que l'utilisateur souhaite représenter par des sondes sur la puce. Le second est facultatif et permet de définir les séquences externes qui sont des séquences avec lesquelles l'utilisateur ne souhaite aucune hybridation aspécifique avec les sondes. Par exemple, lorsque le fichier de séquences d'intérêt contient un sous-ensemble de gènes de l'organisme étudié, il est préférable d'ajouter un fichier de séquences externes contenant le reste des gènes de l'organisme, et éventuellement les pseudogènes et les régions in-

tergéniques. Ces séquences peuvent provenir soit d'une entrée manuelle (téléchargement d'un fichier de séquences personnel), soit de l'utilisation d'une des bases de données disponibles sur le site d'utilisation de ROSO[19].

Un fichier supplémentaire permet à l'utilisateur de définir ses préférences pour la détermination de ses sondes. Ce fichier est facultatif. Il est en effet créé automatiquement sur le site Web lorsque l'utilisateur remplit le formulaire des paramètres. En ce qui concerne l'utilisation en ligne de commande, un menu permet d'entrer les paramètres de façon plus conviviale que sous la forme d'un fichier extérieur. L'utilisateur peut ainsi choisir la taille, l'orientation (complémentaire inverse ou identique) et la localisation générale des sondes. Il peut choisir leur nombre et dans le cas de sondes multiples, autoriser ou non les chevauchements entre elles. Il peut définir, s'il le souhaite, un intervalle de Tm, les seuils de rejet des structures secondaires, la température d'hybridation, la concentration en cibles et les concentrations ioniques (sodium et potassium) de la solution d'hybridation. Des valeurs par défaut sont proposées pour l'ensemble de ces paramètres de façon à faciliter l'utilisation de ROSO par les néophytes.

1.1.22 Préparation des jeux de données

Cette étape préliminaire met en jeu le logiciel *BLAST* et son parseur *DUBLASTN* qui permettent d'effectuer une première comparaison des séquences d'intérêt entre elles, puis une seconde comparaison de ces séquences d'intérêt aux séquences de la base externe. Elle aboutit à la création des fichiers qui sont nécessaires à l'utilisation de ROSO au sens strict. Ces fichiers supplémentaires sont le fichier contenant les *hits* issus de la comparaison des séquences d'intérêt entre elles et éventuellement le fichier contenant les hits issus de la comparaison des séquences d'intérêt avec les séquences externes (cf. **Figure B.1.6**).

[19] http://pbil.univ-lyon1.fr/roso.

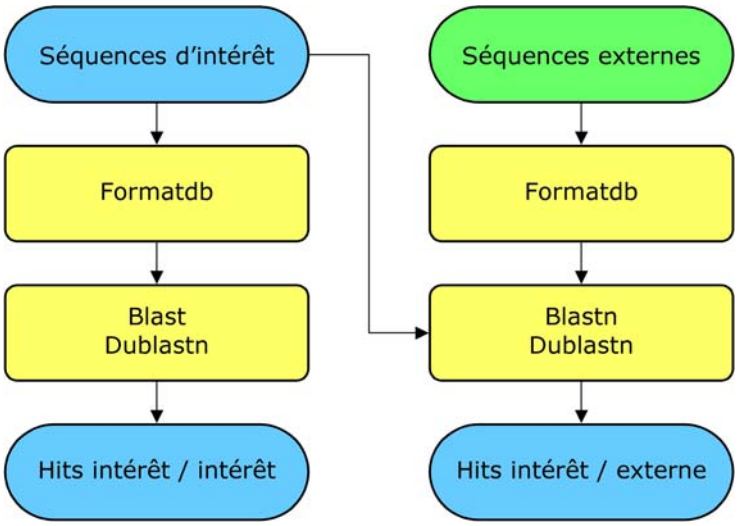


Figure B.1.6 Organigramme de l'étape préliminaire. Les fichiers en bleus sont les trois fichiers d'entrée des étapes de recherche de sondes à proprement parler.

1.1.23 Recherche de sondes spécifiques

Le fichier de séquences d'intérêt et les deux fichiers de *hits* obtenus au cours de l'étape préliminaire, sont utilisés pour la première étape de recherche qui porte sur l'étude de la spécificité des sondes.

Un des aspects cruciaux du choix des sondes concerne la redondance dans les fichiers de séquences. En effet, il est très fréquent qu'un même gène soit présent plusieurs fois dans une base sous des noms différents. De plus, les fichiers d'entrée peuvent contenir des familles de gènes dans lesquelles il est impossible de définir une région spécifique pour chaque gène. Pour éliminer cette redondance et déterminer les éventuelles familles multigéniques, la première étape d'analyse est donc la recherche des gènes identiques dans le fichier de séquences d'intérêt. Le critère d'identité qui a été retenu est une identité de 98 % sur au moins 1000 paires de bases. Ce taux est ramené à 95 % dans le cas de séquences EST en raison de la qualité moins élevée du séquençage. Lorsque le gène possède une taille inférieure à 100 paires de bases, ces taux d'identité doivent concerner au moins 70 % de la longueur du gène. ROSO détermine également les gènes paralogues en utilisant un seuil de 90 % d'identité sur au moins 1000 paires de bases. Parmi les ensembles de gènes identiques, celui possédant la séquence la plus longue est retenu pour l'étude et le ou les noms des gènes identiques lui sont associés.

Les portions de séquences contenant des bases indéterminées (la lettre N remplace le nucléotide) et celles dont la localisation n'est pas en accord avec les choix de l'utilisateur sont ensuite éliminées. Si l'utilisateur le souhaite, les

séquences de faible complexité (séquences de quatre nucléotides successifs identiques comme AAAA) sont également éliminées. Ces séquences favorisent en effet les risques d'hybridation aspécifique et sont de plus difficiles à synthétiser.

Les séquences restantes sont utilisées pour la recherche de sondes spécifiques. Pour cela, l'ensemble des séquences est découpé en différentes régions en fonction des taux d'identité lus dans les fichiers de *hits*. En cas de chevauchements de plusieurs *hits*, le taux d'identité le plus élevé est associé au nucléotide (cf. **Figure B.1.7**). Lorsqu'un nucléotide ne correspond à aucun hit (région spécifique dans la séquence), il est associé à un taux d'identité de 60 %, ce qui correspond à la limite de sensibilité détectée par *BLAST*.

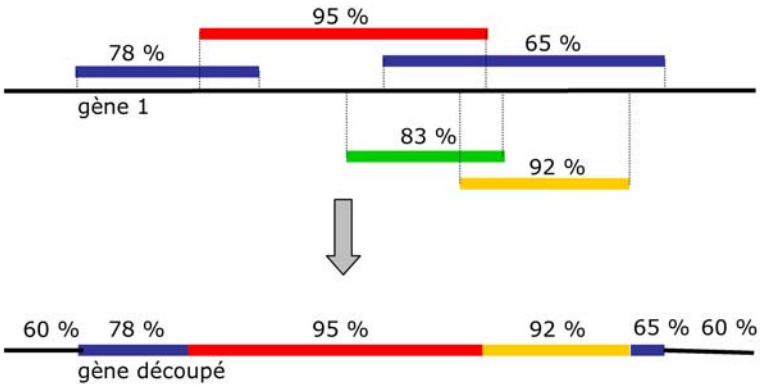


Figure B.1.7 Exemple de découpage en régions pour le gène 1 en fonction des taux d'identité des différents hits.

Ce premier module s'achève avec la détermination du taux d'identité des sondes qui est calculé comme la moyenne arithmétique des taux d'identité des nucléotides qui la composent. Ces taux sont compris entre 60 % pour les régions spécifiques et 100 % pour les régions strictement homologues. Le résultat de cette étude de spécificité est donc une liste contenant toutes les sondes possibles avec leur taux d'identité associé (cf. **Figure B.1.8**).

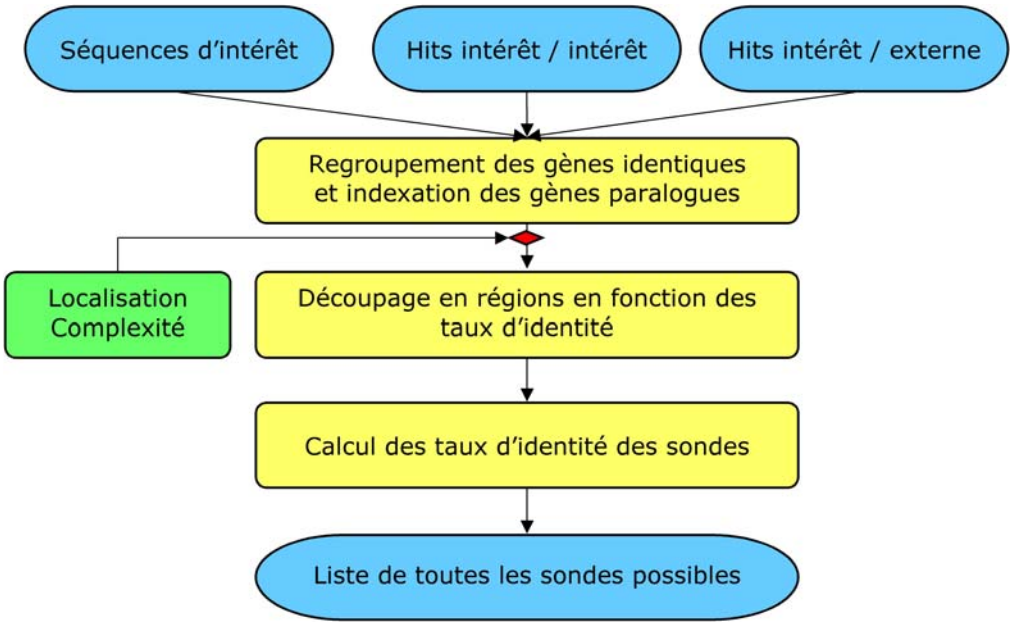


Figure B.1.8 Organigramme du premier module de ROSO dédié à la recherche de sondes spécifiques.

1.1.24 Recherche de sondes optimales

Le second module de ROSO permet d'éliminer les sondes non valides pour les expérimentations sur puces et ne conserve pour chaque gène que la ou les sondes optimales en fonction du nombre de sondes demandé par l'utilisateur (cf. **Figure B.1.9**). La recherche commence avec les sondes possédant les plus faibles taux d'identité. La première étape prend en compte le Tm, qui influence le rendement d'hybridation des cibles sur leurs sondes. Seules les sondes qui possèdent une valeur de Tm en accord avec l'intervalle de températures défini par l'utilisateur sont retenues. La deuxième étape permet d'éliminer les sondes formant des structures secondaires stables, en fonction de la valeur de leurs énergies libres de formation. À la fin de cette première analyse, si un gène ne possède pas de sonde, le taux d'identité est incrémenté de 1 % de façon à analyser de nouvelles sondes moins spécifiques. Cette itération est répétée jusqu'à l'obtention d'au moins une sonde pour chaque gène. Les sondes obtenues subissent ensuite un processus final d'optimisation en deux temps. Les sondes sont d'abord retenues dans l'intervalle de Tm le plus restreint possible de façon à obtenir des valeurs homogènes, puis l'analyse de quatre critères de stabilité permet de retenir uniquement la ou les meilleures sondes possibles pour chaque gène. Chacune de ces étapes est décrite dans les paragraphes qui suivent.

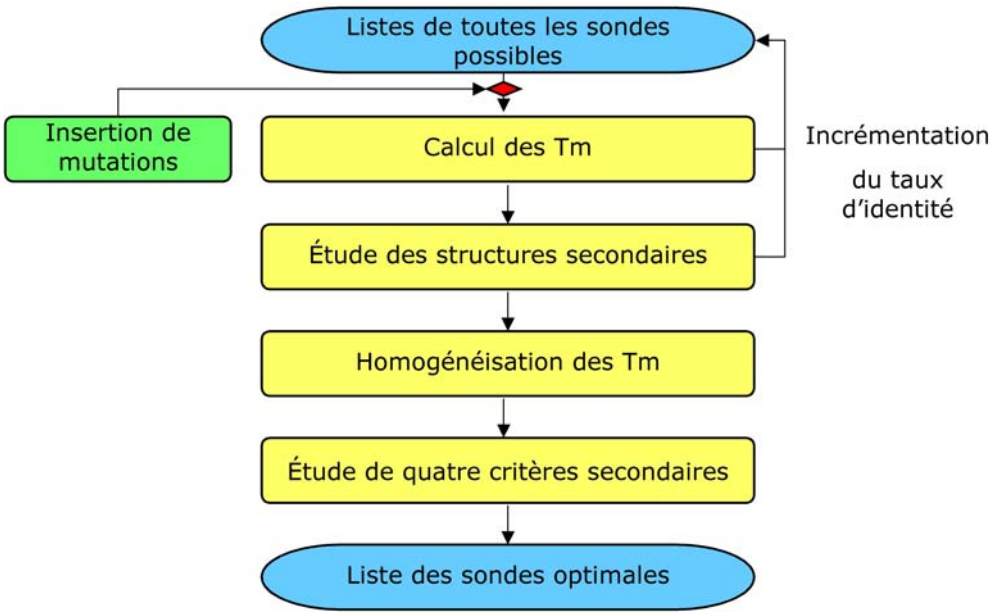


Figure B.1.9 Organigramme du second module de ROSO dédié à la recherche de sondes optimales.

1.1.241 Étude de la température de fusion des sondes

La valeur de la température de fusion est le critère de choix principal des sondes (Chen et Zhu, 1997). Dans ROSO, son calcul est basé sur deux formules différentes suivant la taille des sondes. Dans le cas de sondes de moins de 50 nucléotides, l'enthalpie, l'entropie puis la température de fusion de la première sonde de la séquence sont calculées à l'aide du modèle thermodynamique du plus proche voisin et des paramètres unifiés de Santalucia (1998). Ces paramètres ont été intégrés dans une table afin de faciliter leur lecture (cf. annexes 1.2). Le cadre de lecture est ensuite déplacé d'un nucléotide dans le sens de lecture et les valeurs calculées précédemment sont adaptées à la nouvelle sonde. Les sondes sont sélectionnées lorsqu'elles possèdent une valeur de Tm appartenant à l'intervalle défini par l'utilisateur. Pour les sondes dont la taille est égale ou dépasse 51 nucléotides, la formule du taux de GC est utilisée, mais la même démarche est adoptée.

1.1.242 Étude des structures secondaires des sondes

ROSO recherche parmi les sondes possédant la meilleure spécificité (taux d'identité le plus faible) la présence de structures secondaires stables. Pour chaque sonde, toutes les conformations sont étudiées et leurs énergies libres de formation sont calculées à la température d'hybridation définie par l'utilisateur.

Pour l'étude de la formation des épingles à cheveux, l'ensemble des conformations possibles est testé au moyen de l'algorithme proposé par Rychlik et Rhoads (1989) (cf. **Figure B.1.10**). Lorsque la formation d'une épingle à cheveux est possible, son énergie libre est calculée. Pour chaque sonde, la valeur d'énergie libre la plus faible (correspondant à la structure la plus stable) est retenue. Lorsque cette valeur est au-dessous d'une valeur limite fixée par l'utilisateur ou paramétrée par défaut à 0 kcal.mol^{-1}, la sonde est capable de former une structure secondaire stable et elle est éliminée de la liste des sondes potentielles (Hyndman *et al.*, 1996). Le même principe est appliqué pour la recherche des homodimères mais avec un seuil de rejet à –6 kcal.mol^{-1}.

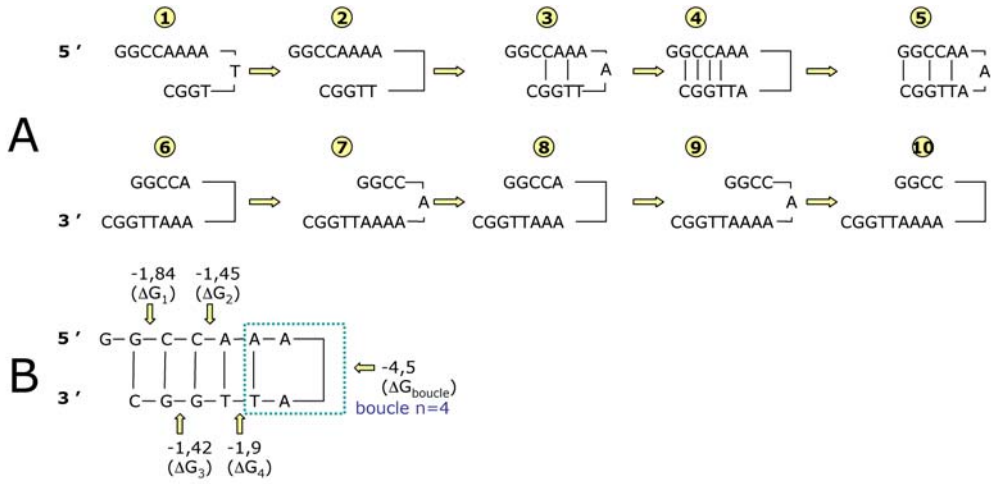


Figure B.1.10 Exemple de l'utilisation de l'algorithme de recherche des épingles à cheveux et du calcul de l'énergie libre associée à la conformation la plus stable.
A : Algorithme de recherche des conformations favorables à la formation des épingles à cheveux. L'extrémité 5' est progressivement déplacée sur le brin 3'. Le nombre de bases nécessaires à la formation de l'épingle à cheveux est testé pour chaque conformation. Si la conformation contient deux liaisons homologues successives et une boucle avec au moins trois bases, la formation de l'épingle à cheveux est possible (conformations 4 et 5).
B : Lorsque la formation de l'épingle est possible, l'énergie libre de la conformation est calculée. Par exemple pour la conformation 4 (d'après Griffin et Smith, 1998) : $\Delta G=(\Delta G1+\Delta G2+\Delta G3+\Delta G4)+\Delta G$ initiation$+\Delta G$ boucle$=(-1,84-1,42-1,45-1,9)+0,98+4,5=(-1,13)$ kcal.mol-1 (avec par exemple $\Delta G1=\Delta H1-T\Delta S1$ à la température d'hybridation T).

Le calcul des énergies libres de formation des structures secondaires est basé, tout comme pour le calcul du Tm, sur le modèle thermodynamique du plus proche voisin. Ce calcul nécessite l'utilisation de l'ensemble des paramètres thermodynamiques disponibles aussi bien pour les couples de bases parfai-

tement homologues que pour les couples contenant une base mésappariée. Ces paramètres ont été intégrés dans une table à double entrée de façon à optimiser l'algorithme de lecture. Pour construire l'ensemble de cette table, la valeur 0 a été utilisée lorsque les valeurs thermodynamiques n'étaient pas disponibles, ce qui est le cas de pratiquement tous les mésappariements doubles (cf. annexes 1.3). Pour les épingles à cheveux, il est nécessaire d'utiliser une valeur supplémentaire qui correspond à la formation de la boucle. Ces valeurs ont été déterminées expérimentalement, et, tout comme les autres paramètres, ont été rassemblés dans une table afin de faciliter leur utilisation (cf. annexes 1.4). L'ensemble de ces valeurs thermodynamiques est rassemblé dans le fichier texte *roso.dpx* utilisé par l'exécutable *roso*, ce qui permet ainsi leur mise à jour régulière en fonction des nouvelles publications.

1.1.243 Homogénéisation des températures de fusion des sondes

Il est essentiel de réduire au maximum la variabilité entre les Tm des différentes sondes. Pour cela, deux matrices sont utilisées. Une première matrice [gènes, Tm] recense pour chaque gène le nombre de sondes correspondant à toutes les valeurs de Tm. Elle permet de choisir la valeur de Tm pour laquelle le nombre de sondes disponibles est le plus élevé. Cette valeur A est utilisée comme valeur centrale du nouvel intervalle réduit de Tm. Une seconde matrice [gènes, Tm] contient pour chaque gène la valeur 0, si aucune sonde ne possède la valeur de Tm, et 1 dans le cas contraire. Sur cette matrice, le premier intervalle de Tm [A, A+1] est testé. Si pour cet intervalle tous les gènes ne possèdent pas au moins une sonde (présence d'un 0 dans la sous-matrice), l'intervalle est élargi de façon symétrique et le nouvel intervalle [A-1, A+1] est testé. Cette procédure itérative est arrêtée lorsqu'un intervalle de Tm permet de conserver au moins une sonde pour chaque gène (absence de 0 dans la sous-matrice).

1.1.244 Choix final des sondes

Les critères qui correspondent d'une part à la valeur de Tm et d'autre part à l'absence de structures secondaires sont deux critères obligatoires pour la conservation des sondes. Toutes les sondes les vérifiant sont donc potentiellement utilisables pour une expérimentation sur puces à ADN. Pour ne retenir qu'une sonde (ou le nombre de sondes choisi par l'utilisateur) quatre critères de stabilité de l'hybridation ont été retenus. Les trois premiers s'expriment de façon booléenne (vrai ou faux) et sont les suivants :

- un taux de GC compris entre 40 et 65 %,
- la présence d'un G ou d'un C aux deux extrémités de la sonde,
- l'absence de 3 G ou de 3 C successifs.

Le dernier critère est une valeur numérique qui correspond à l'énergie libre d'hybridation des pentamères situés aux deux extrémités de la sonde (calculée à la température d'hybridation).

Ces quatre critères sont estimés pour l'ensemble des sondes sélectionnées précédemment. Une note comprise entre 0 et 3 est attribuée à chaque sonde en fonction du nombre de critères booléens vérifiés. Les sondes ayant obtenu la meilleure note sont ensuite départagées en fonction du résultat du quatrième critère. Pour cela les valeurs d'énergie libre d'hybridation des deux pentamères sont ajoutées et la sonde possédant la somme la plus petite (formation de l'hybride cible/sonde le plus stable) est retenue comme étant la sonde optimale.

1.1.245 Fonctionnalité annexe : insertion manuelle de mutations dans les sondes

ROSO contient une fonctionnalité supplémentaire qui permet l'insertion de mutations dans les sondes obtenues. L'utilisateur peut ainsi choisir des témoins de spécificité, c'est-à-dire des sondes dont la séquence diffère de celle du gène cible par une ou plusieurs bases situées en position centrale. L'observation d'une hybridation sur l'un de ces témoins permet de quantifier les phénomènes d'hybridation aspécifique. L'utilisateur choisit successivement le gène, la sonde et la base qu'il souhaite modifier, puis propose une base de remplacement. La valeur de Tm correspondante est calculée en utilisant les formules et les tables permettant de prendre en compte les mésappariements.

1.1.25 Les fichiers de sortie

Après de nombreuses discussions avec différents utilisateurs, plusieurs types de fichiers de sortie ont été jugés utiles. ROSO permet donc d'obtenir les fichiers suivants :

- un fichier de configuration, qui rappelle à l'utilisateur tous les paramètres de sa requête,
- la liste des sondes au format FASTA, qui peut être utilisée pour réaliser de nouveaux *BLAST* de contrôle,
- un fichier de données statistiques contenant les histogrammes de répartition des sondes en fonction des valeurs de Tm et des taux d'identité,
- un fichier contenant tous les détails des résultats pour les sondes (conformations des structures secondaires, valeurs des enthalpies et des entropie, etc.) comme pour les gènes (valeur minimale et maximale des Tm des sondes pour chaque gène, etc.),
- un fichier au format *excel* (sans doute le fichier le plus important !) qui regroupe l'ensemble des sondes et leurs caractéristiques prin-

cipales (nom du gène, numéro de la sonde, position sur le gène, nom des gènes identiques et paralogues, taux de GC, Tm, énergies libres de formation des épingles à cheveux et des homodimères) et qui peut être utilisé pour la commande des oligonucléotides,

- un fichier au format FASTA des gènes pour lesquels aucune sonde n'est disponible et un autre contenant les gènes dont les sondes possèdent des taux de similitude supérieure à 80 % (valeur modifiable par l'utilisateur), c'est-à-dire ayant une mauvaise spécificité. Ces deux fichiers peuvent être utilisés pour une autre utilisation de ROSO avec de nouveaux les paramètres, de façon à obtenir tout de même une sonde, même non optimale pour chaque gène (cf. 1.2 pour une explication détaillée de cette démarche d'optimisation en plusieurs étapes).

1.1.26 L'interface du Pôle Bioinformatique Lyonnais (PBIL)

L'objectif initial de ce travail de développement était d'offrir un outil de choix des sondes pour les puces à ADN à la plateforme transcriptome Rhône-Alpes. Afin de rendre ROSO accessible à l'ensemble de la communauté scientifique et faciliter son utilisation, une interface Web a été développée en PHP (cf. **Figure B.1.11**). ROSO est donc utilisable en ligne sur le site du PBIL (devenu récemment le PRABI) : http://pbil.univ-lyon1/roso.

Figure B.1.11 Page d'accueil pour l'utilisation en ligne de ROSO sur le site du PBIL.

Pour l'interface à proprement parler, une page expliquant le principe et un manuel d'aide ont été réalisés pour faciliter la compréhension des utilisateurs. Des pages de formulaire permettent également de simplifier la saisie des

données et offrent un paramétrage par défaut aux néophytes. Quant à la mise en ligne de ROSO, elle a nécessité de résoudre des problèmes particuliers comme la sauvegarde des fichiers de séquences des utilisateurs dans un répertoire unique (numéro IP), la sauvegarde des fichiers de configuration à partir des formulaires, et les lancements successifs et automatiques de trois exécutables (*formatdb*, *DUBLASTN* et *roso*). Compte tenu de la durée de l'ensemble du processus (plusieurs heures pour les requêtes les plus longues), il a également été essentiel de gérer les files d'attente pour ne pas saturer le serveur, d'envoyer les résultats par mail, et surtout de permettre aux utilisateurs d'annuler leur requête en cas d'erreur.

1.1.3 Discussion

1.1.31 Validations et performance

1.1.311 Validation sur des données simulées

La première évaluation de ROSO a été réalisée sur des données simulées avec le logiciel R. Il s'agit de fichiers contenant des sondes aléatoires de 15 à 70 bases pour des taux de GC compris entre 30 et 70 %. Les valeurs de Tm et d'énergies libres de formation des structures secondaires (épingles à cheveux et homodimères) obtenues avec ROSO ont été comparées avec les résultats du logiciel *Oligo6.68®* (*Molecular Biology Insights*, Cascade, USA). Un écart relatif moyen de 1 % est observé sur l'ensemble des comparaisons. Il correspond vraisemblablement à l'utilisation de valeurs thermodynamiques plus récentes dans ROSO, qui diffèrent légèrement de celles qui sont employées par *Oligo6.68®*.

Une seconde validation a été réalisée sur ces jeux de données simulées pour le calcul des énergies libres de formation des épingles à cheveux avec le logiciel *MFOLD* (Zuker, 2003) qui utilise un algorithme de recherche des conformations possibles beaucoup plus élaboré que ROSO. L'ensemble des sondes a été testé pour trois températures d'hybridation : 25 °C, 42 °C et 65 °C (cf. **Figure B.1.12**). Le seuil de rejet de *MFOLD* est de 0 kcal.mol^{-1}. Avec ce seuil, 82 % des sondes acceptées par *MFOLD* sont également acceptées par ROSO, mais 13 % des sondes acceptées par ROSO sont rejetées par *MFOLD* (faux positifs). En ramenant le seuil de rejet à –1,5 kcal.mol^{-1} pour ROSO, ce taux de faux positifs n'est plus que de 3,5 %. En revanche, ROSO ne retient plus que 71 % des sondes acceptées par *MFOLD*. Diminuer le nombre de faux positifs implique donc une augmentation du nombre de faux négatifs. Pour pallier ce problème, une démarche d'optimisation en plusieurs étapes a été déve-

loppée et permet l'obtention de plusieurs sondes pour certains gènes en fonction de l'importance accordée à la formation des structures secondaires (cf. 1.2).

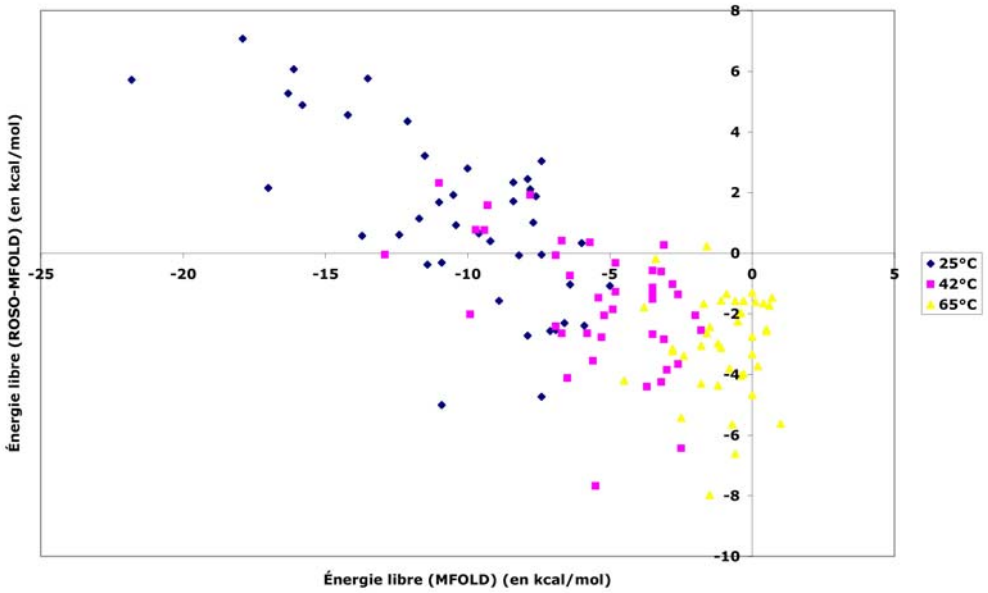


Figure B.1.12 Graphe présentant les valeurs d'énergie libre de formation des épingles à cheveux obtenues avec le logiciel *MFOLD* en fonction des écarts entre valeurs calculées par ROSO et par *MFOLD*) pour les sondes de 70 mers. Ces valeurs ont été calculées pour trois températures différentes d'hybridation : 25 °C, 42 °C et 65 °C.

1.1.312 Validation sur des données réelles

La seconde partie de la validation a porté sur l'utilisation du logiciel ROSO pour choisir des sondes adaptées à l'étude du transcriptome de *Buchnera* mais aussi aux problématiques de différents utilisateurs. Des jeux de sondes ont ainsi été définis pour le génome complet de la bactérie *Ralstonia solanacearum* (Laboratoire Interactions, Plantes, Micro-organisme, INRA de Toulouse), des EST de poulet (Laboratoire de Biologie Moléculaire de la Cellule, ENS de Lyon) et plus de 22000 séquences humaines (Projet Génopôle) (cf. **Tableau B.1.5**). Ils ont permis d'adapter les paramètres d'entrée et les fichiers de sortie de ROSO pour prendre en compte certaines spécificités liées aux organismes d'étude ou à la taille des données. Ainsi, le paramètre de localisation des sondes sur les gènes est l'une contrainte qui a été imposée par le traitement du jeu de données de *Ralstonia solanacearum* et la nécessité d'un fichier de sortie de statistiques est apparue avec le jeu de séquences humaines.

Tableau B.1.5 Premières utilisations du logiciel ROSO.

Organisme	Nombre de gènes	Tailles des sondes (en mers)
Gallus gallus	216	60
Kluyveromyces lactis	47	50
Ralstonia solanacearum	5129	70
Homo sapiens	22322	55
Buchnera aphidicola	619	35

Enfin des fichiers d'entrée ont été créés à partir du génome de la bactérie *Escherichia coli* K12 pour réaliser une étude des performances du logiciel. Les fichiers de séquences d'intérêt comptent un nombre variable de gènes de la bactérie et le fichier de séquence externe contient l'ensemble du génome. Ils ont servi pour déterminer des sondes de 70 nucléotides. Les tests ont été réalisés sur une station *Silicon Graphics® O2®* sous UNIX avec un processeur de type MIPS R12000™ 300 MHz. Pour 100 séquences, l'utilisation de *DUBLASTN* nécessite 5 minutes et celle de ROSO au sens strict 10 minutes, soit une durée totale de 15 minutes. Pour 1000 séquences, chacun des deux processus dure 1 heure. Enfin pour les 4400 séquences *DUBLASTN* nécessite 4 heures 20 et ROSO 2 heures.

1.1.32 Choix de l'utilisation de BLAST

Contrairement à la majorité des algorithmes de recherche et d'optimisation de sondes, le logiciel *BLAST* est utilisé pour rechercher des similitudes entre des régions entières de séquences et non pas directement sur les sondes découpées dans les séquences. L'utilisation de *BLAST* sur les sondes présente l'avantage d'augmenter la rapidité de la recherche mais aboutit à une perte de sensibilité qui peut être importante. En effet, l'algorithme *BLAST* recherche tous les mots pour une taille W définie puis étend la recherche de similitude aux bases contiguës. L'utilisation des sondes aboutit donc à une perte d'information lorsque le mot est présent en dehors d'une sonde potentielle ou même partiellement sur la séquence de la sonde. Pour cette raison, ROSO conserve les séquences complètes des gènes pour la recherche de spécificité. Cette étape étant coûteuse en temps, elle a été découplée du processus d'optimisation pour être réalisée une seule fois. Le processus d'optimisation, qui est beaucoup plus rapide, peut en revanche être répété de façon à rechercher les meilleurs critères d'optimisation, voire obtenir des sondes ayant des caractéristiques différentes pour un même gène.

De nombreux logiciels sont dédiés à la recherche de similitude. Le logiciel FASTA (Pearson et Lipman, 1988) et l'algorithme de Smith et Waterman (1981) sont également basés sur une recherche de mots et offrent une meilleure

sensibilité que *BLAST* (Ning *et al.*, 2001). Cependant leur utilisation est difficile à envisager compte tenu de la taille généralement importante des fichiers de séquences utilisés. En effet, leur sensibilité est obtenue au prix d'une recherche nécessitant beaucoup plus de temps et de ressources. Depuis peu, d'autres types d'algorithmes ont également été développés. L'utilisation d'arbres de décision (*suffix tree*) qui comptent les occurrences de chaque mot dans une séquence permet ainsi de localiser les mutations et les répétitions significatives (Wang et Seed, 2003). D'autres algorithmes utilisent des tables de hachage pour organiser les séquences dans un tableau indexé par des clés. Cette approche est implémentée dans le logiciel récent *SSAHA* (Ning *et al.*, 2001). Ces méthodes alternatives pourraient également être utilisées pour la recherche de spécificité car elles présentent l'avantage d'être beaucoup plus rapides que celles basées sur la recherche de mots. Elles nécessitent cependant une quantité très importante de mémoire.

Les recherches de similarité, présentées précédemment, sont souvent utilisées pour définir des régions spécifiques dans les génomes. Elles n'ont cependant pas la capacité de prédire les comportements d'hybridation. D'autres approches ont donc été développées et notamment l'utilisation du calcul de l'énergie libre d'hybridation d'une cible sur sa sonde. Son utilisation semble fournir de bons résultats, néanmoins d'avantage de données expérimentales sont nécessaires pour définir de façon pertinente les risques d'hybridation aspécifique (Held *et al.*, 2000 ; Matveea *et al.*, 2003). En effet, l'énergie libre d'hybridation dépend d'une part de la présence éventuelle de structures secondaires sur les cibles, qui ne sont généralement pas prises en compte, et d'autre part de la concentration en cibles, qui est très rarement connue (Li et Stormo, 2001).

Cette approche thermodynamique ne permet cependant pas de prendre en compte la redondance qui existe dans les fichiers d'entrée, et l'étude de la complexité des séquences. Pour ces deux aspects, une étude de similarité reste donc nécessaire. Pour éliminer la redondance, Wang et Seed (2003) utilisent comme critère un taux de similitude de 96 % sur la longueur de la séquence étudiée (moins cinq bases). Ils montrent que pour des taux de similitude compris entre 90 et 98 %, la variation du nombre de classes de gènes considérés comme identiques est inférieure à 5 %. L'utilisation dans ROSO de seuils à 98 et à 95 % respectivement pour les gènes et les EST semble donc tout à fait indiquée. En ce qui concerne l'étude des séquences de faible complexité, l'algorithme utilisé dans ROSO se contente d'éliminer les séquences contenant quatre nucléotides successifs identiques. Lorsque ROSO est utilisé sans fichier de séquences externes, il semble donc préférable d'éliminer au préalable les sé-

quences répétées du fichier de séquences d'intérêt. Il est possible d'utiliser pour cela un logiciel comme DUST[20] (Hancock et Armstrong, 1994) qui masque ce type de séquence. Son algorithme est intégré dans le logiciel *BLAST* (option –F) et peut donc être utilisé en même temps que la recherche de similitude. Mais actuellement, l'information de masquage n'est pas conservée par le parseur *DUBLASTN*.

1.1.33 Choix de l'utilisation du modèle thermodynamique du plus proche voisin

Les valeurs d'énergie du modèle thermodynamique du plus proche voisin les plus récemment publiées (Santalucia, 1998) sont utilisées dans ROSO et il semble que ces valeurs évolueront peu (Dirks *et al.*, 2004). En solution et pour des hybrides parfaits, elles permettent d'obtenir une précision de 2 % entre valeurs calculées et mesures expérimentales (Schütz et Von Ahsen, 1999). En revanche, l'utilisation de ce modèle pour les hybridations sur puces à ADN est soumise à certaines limitations.

En effet, ce modèle et ses paramètres ont été définis pour des hybridations réalisées en solution. Or dans les expérimentations sur puce, la sonde est immobilisée sur la lame. Il est donc probable que les entropies de la sonde et de l'hybride cible-sonde sont inférieures à celles qui sont calculées en solution. Il semble cependant que cet écart reste faible, d'autant plus que les valeurs des quelques paramètres obtenus pour des hybridations réalisées sur matrice de gel sont corrélées linéairement avec celles obtenues en solution (Kunitsyn *et al.*, 1996). La détermination expérimentale de certaines valeurs de Tm pour des sondes immobilisées confirme d'ailleurs cette hypothèse (Held *et al.*, 2000). En l'absence de résultats expérimentaux détaillés concernant l'hybridation sur puces le modèle thermodynamique du plus proche voisin reste donc satisfaisant pour prédire, au moins d'un point de vue qualitatif, le comportement d'hybridation sur les puces. Pour cette raison, le calcul du Tm dans ROSO est utilisé essentiellement de façon qualitative. Il permet en effet de rechercher les sondes dans l'intervalle le plus réduit possible, sans considération sur les valeurs absolues. En ce qui concerne les aspects quantitatifs, des études récentes suggèrent l'utilisation du modèle d'adsorption de Langmuir pour décrire l'hybridation sur puce en fonction de la séquence des sondes (Hekstra *et al.*, 2003). Ce modèle relie l'intensité de fluorescence à la concentration en cible. Il permet donc de déterminer les concentrations absolues en cibles et, surtout, de comparer les intensités absolues obtenues pour des sondes et des gènes différents.

[20] http://www.ncbi.nlm.nih.org.

1.1.331 Calcul du Tm

Le modèle thermodynamique du plus proche voisin met en jeu deux types de concentrations pour le calcul du Tm : la concentration en acides nucléiques appariés et la concentration en sels (sodium et potassium).

En ce qui concerne la concentration en acides nucléiques, deux problèmes se posent. Le premier concerne la nature des acides nucléiques à considérer. Les sondes n'étant pas en solution, leur concentration n'est pas connue. Cependant, sachant qu'elles sont déposées en large excès par rapport aux cibles, la concentration en acides nucléiques appariés est proche de la concentration moyenne en cibles. Cette valeur ne permet néanmoins pas de prendre en compte les variations de concentration individuelle de chaque cible. Par conséquent, la température d'hybridation est largement inférieure au Tm des gènes fortement exprimés et réciproquement pour ceux faiblement exprimés. Il existe donc une amplification des écarts d'intensité de fluorescence entre les gènes fortement et faiblement exprimés. Le second problème concerne la valeur numérique à utiliser par défaut. Une estimation a donc été effectuée à partir du protocole utilisé. Le volume d'hybridation de la solution de cibles déposée sur la puce est d'environ 50 µl pour 15 µg d'ARN. La masse moléculaire moyenne d'un nucléotide est de 335 $g.mol^{-1}$ (Lehninger, 1977) et la taille moyenne d'un ARN messager est supposée comprise entre 1000 nucléotides chez les bactéries et 2000 nucléotides chez les eucaryotes. Sachant que l'échantillon contient environ 3 % d'ARNm et que le rendement d'un marquage avect synthèse d'ADNc est de l'ordre de 30 %, la concentration totale en ARNm cible déposée est d'environ 10^{-6} M. Cette concentration, qui est également utilisée par d'autres logiciels (Li et Stormo, 2001 ; Rouillard *et al.*, 2003), a donc été retenue pour le paramétrage par défaut du logiciel. Enfin, même s'il est vrai que la valeur du Tm dépend beaucoup de la concentration (Mergny et Lacroix, 2002), l'erreur reste acceptable (Santalucia et Turner, 1997). En effet, une erreur de 10 % sur l'estimation de la concentration C n'induit qu'une erreur de 1 % sur le terme $ln\ C$ pour une concentration en acides nucléiques de l'ordre de 10^{-5}M.

Pour les concentrations en ions sodium et potassium, il est intéressant de conserver leurs termes dans le calcul du Tm car des études expérimentales montrent que la sensibilité du Tm à la concentration en sels est la même que les acides nucléiques soient en solution ou immobilisés (Meunier-Prest *et al.*, 2003). La valeur par défaut qui a été retenue est de 1 M car elle est utilisée par de nombreux autres logiciels (Rouillard *et al.*, 2003). Une étude expérimentale récente a montré de plus que la meilleure corrélation entre valeurs expérimentales et théoriques est observée lorsqu'une concentration en sels de 1M est employée pour le calcul du Tm. Cela reste vrai même lorsque les expériences sont réalisées avec une concentration en sels de 100 mM (Matveea *et al.*, 2003).

Le modèle thermodynamique du plus proche voisin inclut d'autres paramètres plus complexes à prendre en compte, mais qui permettent une étude plus fine de la stabilité des hybridations (Mathews *et al.*, 1999). L'hybridation d'une cible sur sa sonde aboutit généralement à la présence de bases non appariées aux extrémités de la cible. La présence de la première de ces bases libres (*dangling end*) stabilise de façon significative l'hybride formé (Lane *et al.*, 1992). Il semble que les bases libres suivantes interviennent également sur la stabilité (Williams *et al.*, 1994), cependant les valeurs thermodynamiques qui sont publiées concernent uniquement le premier nucléotide (Bommarito *et al.*, 2000). Ces valeurs montrent que la meilleure stabilisation est obtenue avec une base A située en position 5' de la cible (Southern *et al.*, 1999). L'explication de cette stabilisation n'est pas encore clairement établie, mais il s'agit vraisemblablement d'interactions de type dipôle/dipôle induit ou de type dipôle induit/dipôle induit et d'effets de solvants (Gellman *et al.*, 1996). Par ailleurs, la présence de bases libres semble limiter les possibilités de réaction de dissociation des deux brins (Lane *et al.*, 1992). Ces paramètres thermodynamiques ne sont actuellement pas utilisés dans ROSO car, contrairement aux paramètres classiques du modèle, ils commencent tout juste à être explorés et leurs valeurs numériques subiront certainement des modifications importantes dans les années à venir (Dirks *et al.*, 2004).

1.1.332 Étude des structures secondaires

L'utilisation du modèle thermodynamique du plus proche voisin pose le problème, tout comme pour le calcul du Tm, de l'étude de sondes immobilisées. Cependant, d'autres limitations interviennent. La première est liée à une absence de données en ce qui concerne les mésappariements doubles. Les seules valeurs connues sont en effet celles de GG.TT et de GT.TG (Schütz et Von Ahsen, 1999). Cette absence de données est gérée par l'utilisation de la valeur 0 dans les tables d'énergies d'hybridation de ROSO. Comme cette valeur nulle remplace une valeur positive, les valeurs d'énergie libre de formation des structures secondaires calculées sont inférieures à la réalité (cf. 1.1.311) avec pour conséquence l'élimination possible de sondes valides. Cette solution a tout de même été retenue car, compte tenu des connaissances actuelles, il semble préférable d'éliminer des sondes valides plutôt que de risquer de conserver des sondes formant des structures secondaires. La seconde limitation du modèle concerne l'influence de la localisation des mésappariements dans les hybrides intra ou inter-sondes. En effet, une analyse statistique du comportement de fusion des oligonucléotides indique que leur contribution aux valeurs d'énergie peut diminuer de près de 90 % (Borer *et al.*, 1974) lorsqu'ils sont localisés sur une des trois dernières bases situées aux extrémités (Santalucia, 1998). Cet effet semble lié à des interactions stériques défavorables qui déstabilisent les mésap-

pariements internes (Peyret *et al.*, 1999). L'utilisation des paramètres classiques aux extrémités des structures secondaires, lorsqu'il existe des mésappariements, peut donc également aboutir au calcul d'une valeur d'énergie libre inférieure à la réalité.

L'algorithme de recherche des épingles à cheveux de ROSO n'utilise pas certains paramètres complexes permettant de prendre en compte les torsions imposées par le repliement de la molécule comme les valeurs de liaisons coaxiales (Walter *et al.*, 1994). De même, de nombreuses possibilités de conformation, comme la présence de boucles internes dans la molécule ne sont pas prises en compte par ROSO. En revanche le logiciel *MFOLD* (Zuker *et al.*, 1999), basé sur l'utilisation de chaînes de Markov cachées (Pervouchine *et al.*, 2003), prend en compte ce type de paramètres et de conformations. Il utilise des paramètres qui ont été déterminés par Santa Lucia, mais qui n'ont pas été publiés (Zuker, 2003). Ces nouvelles valeurs, tout comme les précédentes, risquent toutefois de subir des modifications importantes (Dirks *et al.*, 2004). Une fois publiées, ces valeurs pourront être intégrées sans problème dans ROSO car les paramètres thermodynamiques ne sont pas intégrés dans le code source mais sont rassemblés dans un fichier texte *roso.dpx*. Ce fichier est fourni aux utilisateurs avec l'exécutable et peut donc évoluer sans modification du code.

ROSO n'étudie que les structures secondaires potentielles des sondes. D'autres logiciels prennent en compte celles qui sont formées par les cibles (Wang et Seed, 2003). Plusieurs auteurs suggèrent effectivement l'importance de la formation possible de structures secondaires dans les cibles (Mir et Southern, 1999 ; Sohail *et al.*, 1999). Néanmoins, l'étude expérimentale de Luebke *et al.* (2003) indique qu'il n'est pas nécessaire d'étudier les structures secondaires des cibles, et que l'étude des sondes est suffisante pour prédire la qualité des hybridations. Ils proposent de plus d'associer la valeur d'énergie libre de formation des épingles à cheveux (hybridation intra-sonde) à celle d'hybridation cible/sonde et d'utiliser la différence entre les deux pour prédire la qualité des sondes. Une autre étude statistique démontre que l'analyse des structures secondaires des cibles ne donne aucune indication pour prédire la qualité des hybridations sur puces. En revanche, les résultats obtenus sur un grand nombre de sondes montrent que l'intensité d'hybridation est inversement proportionnelle à la valeur absolue des énergies libres de formation des épingles à cheveux sur les sondes et soulignent l'intérêt d'étudier également la formation éventuelle d'homodimères entre les sondes (Matveea *et al.*, 2003). En effet, malgré une forte corrélation entre l'évaluation thermodynamique des homodimères et celle des épingles à cheveux, les auteurs montrent que l'utilisation des deux critères offre une meilleure discrimination des sondes en

termes de rendement d'hybridation que l'utilisation du seul critère de formation des épingles à cheveux.

1.1.34 Un problème d'optimisation multicritère

Le développement du logiciel ROSO a montré que le choix des sondes pour les hybridations sur puce ne peut être qu'un compromis entre les nombreux critères retenus (Nielsen *et al.*, 2003). Pour ROSO, la démarche d'optimisation qui a été retenue est une démarche par raffinements successifs, avec un critère obligatoire d'exclusion des sondes formant des structures secondaires. Cependant, ce type de recherche est typiquement un problème d'optimisation multicritère. Des algorithmes ont d'ailleurs été développés pour choisir, en présence de critères multiples, une alternative parmi un nombre infini de solutions. Le logiciel *OligoDesign* utilise ce type d'algorithme de choix en calculant un paramètre de logique floue (Tolstrup *et al.*, 2003). Cette méthode permet à l'utilisateur de choisir parmi un ensemble de solutions optimales celle qui est la mieux adaptée à ses objectifs expérimentaux. Dans ROSO par exemple, il serait possible d'utiliser une démarche analogue pour l'étude des critères secondaires. Deux approches sont possibles. La première est basée sur la création d'une combinaison linéaire des quatre critères facultatifs pondérés, puis sur la minimisation de la fonction résultante. L'inconvénient de cette approche est qu'il est difficile, compte tenu des connaissances actuelles, de définir cette pondération. La seconde possibilité est l'utilisation d'un algorithme génétique qui permet de définir un nuage de solutions dont l'enveloppe convexe constitue l'ensemble de solutions optimales. Cependant, la mise en œuvre de ce type de démarche reste conditionnée à la validation expérimentale de la plupart des critères utilisés. Une autre solution a donc été envisagée pour ROSO. Il s'agit d'une démarche d'optimisation itérative, qui permet d'obtenir plusieurs sondes pour les « gènes à problème », et surtout qui offre la possibilité d'une optimisation plus fine que l'utilisation d'une pondération pour l'ensemble des gènes. Une explication générale de cette démarche est proposée sur le site de ROSO et un exemple est détaillé dans la partie 1.2.

1.1.35 Extensions possibles

1.1.351 Application aux sondes d'ADNc

Malgré son nom, ROSO peut être utilisé sans problème pour choisir des sondes d'ADNc, aussi bien que d'autres logiciels comme *OliD* (Talla *et al.*, 2003) ou *Primex* (Lexa et Valle, 2003) développés spécifiquement pour cette problématique.

ROSO peut également déterminer une partie des critères de choix des amorces spécifiques qui sont nécessaires à la production des sondes d'ADNc

par PCR. Quelques développements sont néanmoins nécessaires, et pourraient être envisagés, pour prendre en compte toutes les caractéristiques propres au choix des amorces. Cependant, il existe déjà de nombreux logiciels pour la recherche des amorces spécifiques comme *DOPRIMER* (Kämpke *et al.*, 2001), *GenomePRIDE* (Haas *et al.*, 2003), *PRIMEGENS* (Xu *et al.*, 2002) qui utilise *PRIMER 3* (Rozen et Skaletsky, 2000) ou encore *ProbeWiz* (Nielsen et Knudsen, 2002).

1.1.352 Application aux sondes destinées aux génotypage

L'utilisation des puces à ADN pour le criblage de mutations impose des contraintes un peu particulières pour le choix des sondes. Le principe est relativement simple. Le Tm pour un hybride présentant des mésappariements est généralement réduit de 1 à 1,5 °C pour 1 % de mésappariement. Pour une sonde de 20 nucléotides, un mésappariement interne peut donc abaisser le Tm de 5 à 7,5 °C. Cet intervalle est suffisant pour discriminer des hybrides parfaits et des hybrides présentant un mésappariement. Avec des sondes plus grandes en revanche, le Tm n'est pas modifié de façon significative. Pour réaliser une étude de génotypage, il est donc essentiel de choisir des sondes de 20 nucléotides, en plaçant la mutation qui existe sur la cible au centre de la séquence de la sonde correspondante. Rechercher des sondes spécifiques de 20 bases avec ROSO nécessiterait simplement une adaptation des paramètres de *BLAST*. En revanche, repérer et placer la base mutée sur la cible au centre de la sonde impose l'utilisation de nouveaux algorithmes. Les algorithmes du logiciel libre *PROBE* (Pozhitkov et Tautz, 2002) permettent de modéliser la stabilité relative des sondes avec un mésappariement. Pour cela, le mésappariement est considéré comme « un point faible », dont la localisation est modélisée par une fonction de probabilité permettant de prendre en compte les contributions différentielles des positions centrales et terminales. Cet algorithme pourrait éventuellement être intégré dans ROSO.

Cependant, il existe déjà un nombre important de logiciels pour ce type de problématique. *DNAPROBE* (Drummond et Stamper, 1999) utilise les alignements de séquences protéiques comme le propose Nash (1993) pour définir des sondes dédiées à l'analyse des familles multigéniques. En ce qui concerne les études phylogénétiques et la recherche de signatures spécifiques de certaines espèces notamment bactériennes, les logiciels comme *ARB* (Ludwig *et al.*, 2004) ou *PROBEMER* (Emrich *et al.*, 2003) permettent de choisir des sondes spécifiquement sur les ARN ribosomiaux. De même, Zhang *et al.* (2002) ont défini une base de signatures dédiées aux études sur les ARN 16S. Bien que les mésappariements ne soient pas systématiquement localisés au centre des sondes, ces bases disponibles librement pourraient éventuellement être intégrées dans ROSO.

1.1.353 Application aux sondes multi-transcriptomes

Pour réduire le coût associé à la conception des puces à ADN, le logiciel *OligoWiz* (Nielsen *et al.*, 2003) permet de choisir des sondes destinées à l'étude de différents organismes pour la conception de puces multi-transcriptomes. ROSO ne permet pas de choisir des sondes qui peuvent être utilisées pour plusieurs organismes. Il serait néanmoins possible d'envisager d'intégrer, dans l'étape de recherche des sondes spécifiques, un critère permettant, par exemple, de retenir uniquement les régions communes à deux organismes. Cependant, des bases dédiées au stockage et au suivi des collections de sondes oligonucléotidiques ont récemment été développées pour pallier ce problème de sondes communes à différents organismes. Ainsi, le logiciel *ProbeLynx* (Roche *et al.*, 2004) permet une actualisation des caractéristiques des sondes en fonction des mises à jour d'annotations. Cela permet bien sûr de bénéficier des dernières annotations disponibles, mais également de contrôler les risques d'hybridation aspécifiques, et surtout de vérifier qu'une sonde peut être utilisée pour un autre organisme. De même le logiciel *ProbeMatchDB* permet d'identifier les sondes équivalentes entre espèces (Wang *et al.*, 2002).

1.1.354 Application aux autres acides nucléiques

Seuls les paramètres thermodynamiques concernant l'hybridation entre deux molécules d'ADN sont actuellement entrés dans le fichier *roso.dpx*. Cependant la structure des algorithmes de lecture utilisés dans ROSO offre la possibilité d'ajouter très facilement les paramètres d'hybridation entre ADN et ARN. Les utilisateurs de cibles d'ARN peuvent donc également utiliser ROSO.

Par ailleurs, il est possible d'utiliser d'autres types d'acides nucléiques pour la synthèse des sondes. Le comportement des sondes d'acide peptidonucléique (APN, petit peptide portant des bases azotées) est très proche de celui des oligonucléotides d'ADN, notamment en ce qui concerne les hybridations entre séquences complémentaires. Or leur utilisation sur les puces à ADN s'avère intéressante pour de nombreuses raisons. En effet, des études expérimentales ont montré que les hybridations entre APN et ADN offrent une sensibilité et une spécificité plus élevées que les hybridations mettant en jeu deux molécules d'ADN (Maughan *et al.*, 2001). De plus, l'hybridation peut se dérouler dans une solution dépourvue de sels. Cette absence d'ions permet d'éviter la formation de structures secondaires pour les cibles comme pour les sondes (Hoheisel, 1997). Il serait donc possible d'ajouter les paramètres thermodynamiques pour les APN dans le fichier *roso.dpx* (Griffin et Smith, 1998).

Enfin, l'utilisation d'une nouvelle classe d'ARN bicyclique appelée *LNA* (*Locked Nucleic Acid*) pour la synthèse des sondes est actuellement en développement. Il s'agit de dérivés d'ARN dans lesquels le cycle ribose est

contraint par un lien méthylène. Cette restriction de conformation leur confère une affinité exceptionnelle pour leurs cibles complémentaires d'ARN et d'ADN (Tolstrup *et al.*, 2003). De plus, l'utilisation de sondes *LNA* permet d'augmenter de façon significative la spécificité et la sensibilité des sondes lorsqu'il existe plus de 90 % d'identité entre deux cibles (Braasch et Corey, 2001). Enfin, l'utilisation de ce type de sondes permet l'hybridation de cibles non marquées, qui sont simplement détectées par visualisation des phosphates qui sont absents des molécules de sondes (Brandt *et al.*, 2003 ; Jacob *et al.*, 2004). La plupart des paramètres thermodynamiques concernant le comportement d'hybridation de ce type de sondes ont été déterminés récemment. Ils pourraient également être ajoutés dans le fichier *roso.dpx*, comme dans le logiciel *OligoDesign,* qui est spécifiquement dédié à la détermination de ce type de sondes (Tolstrup *et al.*, 2003).

1.1.36 Comparaison avec d'autres logiciels

Parallèlement au développement de ROSO et à son installation sur le site du PBIL, de nombreux logiciels de détermination des sondes oligonucléotidiques ont été développés pour répondre aux besoins des utilisateurs. Les logiciels académiques les plus connus et ayant fait l'objet d'une publication sont présentés avec leurs principales caractéristiques dans le **Tableau B.1.6**. Il existe de plus des logiciels comme celui de Tolonen *et al.* (2002) dédié spécifiquement aux choix des sondes destinées aux puces *Affymetrix*. Il offre un couplage entre optimisation des sondes et minimisation du nombre d'étapes nécessaires à la synthèse *in situ*. Cet algorithme permet de réduire le coût de production des puces, mais malheureusement au détriment d'un processus de sélection rigoureux des sondes. Enfin, il existe également des logiciels commerciaux dont les caractéristiques sont peu détaillées comme le logiciel *ArrayDesigner*[21] (*Premier Biosoft International*, Palo Alto, CA).

L'originalité majeure du processus d'optimisation de ROSO par rapport aux logiciels présentés dans le **Tableau B.1.6** est essentiellement liée à la séparation de l'étape d'analyse de spécificité (basée sur l'utilisation de *BLAST* qui nécessite beaucoup de temps pour un nombre important de séquences) et de l'étape de sélection des sondes optimales. Cette séparation en deux temps permet à l'utilisateur de réaliser plusieurs étapes d'optimisation, en introduisant différents critères de sélection par raffinement progressif. Des modèles de recherche ont été définis à partir de différents types de séquences d'entrée et notamment sur celles de *Buchnera* (cf. 1.2). Ainsi, en présence de gènes pour les-

[21] http://www.premierbiosoft.com/dnamicroarray/dnamicroarray.html.

quels il est difficile d'obtenir une sonde optimale, ROSO permet de calculer rapidement plusieurs solutions en accordant des importances variables aux différents paramètres. Deux utilisations de ROSO sont donc possibles. Pour les néophytes, il est en effet possible d'utiliser ROSO avec l'ensemble des paramètres par défaut puisque même l'intervalle des Tm est réduit automatiquement. Au contraire, d'autres logiciels (Mrowka *et al.*, 2002 ; Chang et Peck, 2003) calculent les paramètres de toutes les sondes possibles et laissent à l'utilisateur le choix final des sondes. En ce qui concerne ROSO, les utilisateurs qui souhaitent s'impliquer dans la démarche d'optimisation ont la possibilité de définir plusieurs étapes d'optimisation, qui prennent en compte leurs contraintes matérielles et biologiques. Ces aspects particuliers sont en effet difficiles à intégrer dans un algorithme général d'optimisation.

ROSO possède également certaines fonctionnalités qui n'existent pas dans les autres logiciels. Ainsi l'utilisateur a la possibilité de définir un intervalle de localisation des sondes sur les gènes, en proposant les positions de la première et de la dernière base à partir de l'une ou l'autre des deux extrémités. Cette caractéristique permet d'éviter de choisir des sondes trop proches de l'extrémité 3' lorsque des amorces spécifiques sont choisies pour le marquage, ou encore d'éviter les positions les plus proches de l'extrémité 5' qui peuvent être des régions mal annotées (mauvais positionnement du codon *start*). ROSO est également le seul logiciel à intégrer l'étude de la formation potentielle d'homodimères pour les sondes. Enfin, le choix final des sondes dans ROSO est basé sur des critères de stabilité, contrairement à la plupart des logiciels présentés qui utilisent un critère de localisation. Les algorithmes commencent en effet la recherche des sondes à partir de l'une des deux extrémités et la première sonde valide est simplement retenue.

Tableau B.1.6 Présentation des principaux logiciels de détermination de sondes oligonucléotidiques. Les règles de Lockhart *et al.* (1996) qui sont utilisées par certains logiciels sont un ensemble de règles de décision qui permettent d'exclure les sondes qui vérifient au moins une des conditions suivantes : (1) le nombre de bases seules (A,T, C ou G) dépasse la moitié de la taille de la sonde, (2) la longueur totale de séquences de bases identiques dépasse un quart de la séquence, (3) le taux de GC est au-dessous de 40 % ou au-dessus de 60 % et (4) il n'existe pas de séquences palindromiques dans la séquence. L'abréviation PVV, qui est utilisée dans le tableau, signifie que le logiciel utilise le modèle thermodynamique du plus proche voisin pour le calcul du Tm.

N°	Nom	Pays	Publication	Taille des sondes	Spécificité	Complexité	Épingle à cheveux	Homodimères	Tm	Homogénéisation des Tm	Choix final des sondes	Bonus	Validation	Disponibilité
1	ArrayOligoSelector	USA	(Bozdech *et al.*, 2003)	70 mers	*BLASTN* (défaut) ΔG d'hybridation croisée	Algorithme de Smith et Waterman (1981)	NON	NON (calcul du taux de GC)	NON	Première sonde à partir de l'extrémité 3'	Possibilité d'éliminer certaines séquences	*Plasmodium Falciparum* 6272 gènes	Exécutable Linux http://arrayoligosel.souceforge.net
2	OligoArray	USA	(Rouillard *et al.*, 2002)	50 mers	*BLASTN* (-S1)	NON	*MFOLD* (T=50 °C) Seuil de l'utilisateur	NON	PPV $C=10^{-6}$M	Intervalle imposé par l'utilisateur	Première sonde à partir de l'extrémité 3' (avec un pas de 10 mers)	Possibilité d'éliminer certaines séquences	*Saccharomyces cerevisae* 6343 gènes	Exécutable et sources Linux Utilisable en ligne : http://berry.engi.umimch.edu/oligoarray2
3	OligoArray 2.0	USA	(Rouillard *et al.*, 2003)	Jusqu'à 50 mers	*BLASTN* (-W7 –e variable -S1) Calcul duTm d'hybridation croisée	Masque les séquences répétées et les quintuplés de bases identiques	*MFOLD* (T=65 °C) ΔG<0	NON	PPV $C=10^{-6}$ M	Intervalle imposé par l'utilisateur mais adaptation de la taille des sondes	Première sonde à partir de l'extrémité 3' (avec un pas de 5 mers)	Détail des hybridations aspécifiques potentielles Fonction reliant le Tm, la taille et le taux de GC pour aider l'utilisateur à choisir l'intervalle de Tm	*Arabidopsis thaliana* 75764 sondes pour 26140 gènes	Exécutable Linux Utilisable en ligne : http://berry.engi.umimch.edu/oligoarray2
4	Oligodb	Allemagne	(Mrowka *et al.*, 2002)	Libre	*BLASTN* (défaut)	*DUSTN*	*MFOLD* (T=65°C) ΔG<0	NON	*MELTING* (Le Novere, 2001)	NON	Laissé à l'utilisateur	Localisation au choix proche de l'extrémité 3' ou 5'	Dédié uniquement aux séquences humaines (base ENSEMBL)	Utilisable en ligne : http://oligodbs.charite.de
5	OligoDesign	Danemark	(Tolstrup *et al.*, 2003)	Libre	*BLASTN* (-W 9 –e 100) paramètres modifiables par l'utilisateur	NON	Sondes : (Smith et Waterman, 1981) Cibles : (Nussinov *et al.*, 1990)	NON	PPV	NON	Pondération des des paramètres (*score* de logique floue)	dédié uniquement aux sondes d'ALN pour des cibles d'ADN ou d'ARN	*Caenorhabditis elegans* 120 gènes (50 mers)	Utilisable en ligne : http://lmatools.com
6	OligoPicker	USA	(Wang et Seed, 2003)	20 à 100 mers	*BLASTN* (-e 1000) et table de hachage	*DUSTN*	Uniquement sur les cibles avec *BLASTN*	NON	NON (calcul du taux de GC)	Intervalle fixe de 10°C sur la valeur médiane des sondes	Première sonde à partir de l'extrémité 3' ou 5'	Base de données de sondes pour la Souris et l'Homme	Souris 14 554 gènes Humain 17558 gènes	Exécutable et sources Linux http://pga.mgh.harvard.edu/oligopicker/index.html
7	OligoWiz	Danemark	(Nielsen *et al.*, 2003)	Libre	*BLASTN* (défaut) Élimination des introns	Faible complexité gère les bases N	NON	NON	PPV $C = 2,5\ 10^{-10}$ M	OUI (ajustement de la taille des sondes)	Visualisation graphique permettant l'ajout de paramètres personnels Pondération des différents paramètres	Position en 5' ou 3' Interface graphique permettant une interactivité Sondes multi-transcriptome	*Saccharomyces cerevisae* 6600 gènes (45-55 mers)	Exécutables Linux, MacOSX et Windows Utilisable en ligne : http://www.cbs.dtu.dk/services/OligoWiz
8	Oliz	USA	(Chen et Sharp, 2002)	50 mers	*BLASTN* (défaut)	Recherche dans les régions 3'UTR seulement	NON	NON	*EMBOSS-PRIMA*	Intervalle fixe de 10 °C centré sur 76 °C	NON Sondes ordonnées en fonction de la localisation	Taux de GC compris entre 45 et 50 %	Rat 1814 gènes	Exécutables Linux http://www.utmen.edu/pharmacology/otherlinks/oliz.html
9	ProbeSel	Allemagne	(Kaderali et Schliep, 2002)	Libre	Arbres de décision	NON	NON	NON	PPV	OUI (ajustement de la taille des sondes)	NON Sondes ordonnées en fonction du Tm	Détermination de la température d'hybridation optimale	HIV 58 gènes (18-21 mers) ARN 28 S 1230 gènes	Exécutables Windows et Solaris

1.1.4 Conclusion

Le développement du logiciel ROSO a bien sûr été l'occasion d'un véritable apprentissage de la programmation informatique avec l'utilisation de différents langages (C, PHP et Javascript), mais également l'occasion de comprendre l'architecture d'un serveur comme le PBIL. Et même si l'installation de ROSO en ligne n'a pas été aussi simple que prévue, la conception de l'interface Web a été l'occasion de moments créatifs avec, par exemple, la conception des logos. Actuellement le logiciel est utilisé par de nombreux utilisateurs, aussi bien via l'interface Web (plus de 700 utilisations) qu'avec l'utilisation des exécutables en ligne de commandes (18 exécutables distribués).

1.2 Recherche d'un jeu de sondes optimales pour l'étude du transcriptome de *Buchnera*

Une démarche d'optimisation en plusieurs étapes a été développée pour permettre une utilisation multicritère de ROSO. L'explication de cette démarche est intégrée dans cette partie à la présentation du choix des sondes dédiées à *Buchnera*.

1.2.1 Les séquences d'entrée

Les fichiers de séquences ont été obtenus à partir du génome de *Buchnera aphidicola* Ap, disponible dans la base *Genbank*[22] (version du 15/08/02). Les deux plasmides sont annotés AP001070 et AP001071 et le chromosome est divisé en deux parties : AP001118 et AP001119. Sur ces quatre parties du génome, l'outil *Query-win,* disponible sur le serveur du PBIL, a permis d'extraire 610 CDS au format FASTA. Les pseudogènes ne sont pas disponibles dans *Genbank*. La base de données du *Riken Institute*[23] contient toutes les séquences obtenues lors du séquençage {Shigenobu, 2000 #381}, mais ne fournit pas non plus d'annotation spécifique pour les pseudogènes. L'utilisation de *BLAST* pour comparer les deux banques a donc permis d'identifier les neuf pseudogènes (cf. **Tableau B.1.7**). Ces neuf séquences ont ensuite été ajoutées aux 610 séquences obtenues à partir de la base *Genbank* pour créer le fichier de séquences d'intérêt.

[22] http://www.psc.edu/general/software/packages/genbank/genbank.html.
[23] http://*Buchnera*.gsc.riken.go.jp/, integrated *Buchnera* Genome Database.

Le fichier de séquences externes est constitué des quatre parties de *Genbank* qui représentent le génome complet de *Buchnera,* avec l'ensemble des régions intergéniques. Les séquences de trois cibles de contrôle ont été ajoutées. Il s'agit de trois gènes bactériens, qui sont disponibles dans des constructions plasmidiques au laboratoire, et dont les ARNm peuvent donc être produits par expression *in vitro*. Ces gènes sont les gènes *pcp* (Awadé *et al.*, 1992) et *pelK* (Nasser *et al.*, 1993) de *Bacillus subtilis* et le gène *pelL* (Lojkowska *et al.*, 1995) d'*Erwinia chysanthemi*.

Tableau B.1.7 Liste des neuf pseudogènes de *Buchnera aphidicola* (d'après les données obtenues sur le site de l'institut RIKEN.

Nom	BU	Taille (mers)	Protéine
cvpA	BU168	482	Colicine V
ddlB	BU214	923	D-alanylalanine synthétase (ligase B)
apbE	BU227	1076	Lipoprotéine précurseur de la biosynthèse de la thiamine
rnhA	BU247	470	Ribonucléase H
ycfW	BU297	1238	Composant hypothétique du transporteur membranaire ABC
cmk	BU310	656	Cytidilate kinase
fabD	BU350	959	Transacylase
fis	BU400	297	*Factor-for-inversion stimulation protein*
hemD	BU592	758	Uroporphyrinogène-III synthétase

1.2.2 Les spécifications des sondes

Parmi les 619 gènes de *Buchnera*, 3 codent pour des ARN ribosomiaux (*rrl* pour l'ARNr 23S, *rrf* pour l'ARNr 5S et *rrs* l'ARNr 16S), 32 pour des ARN de transfert et 1 pour un ARN stable (*rnpB* qui est le composant ARN de la ribonuclease P). Compte tenu de ce nombre relativement faible de gènes, il a été décidé d'utiliser deux sondes pour représenter chacun des gènes, à l'exception des gènes codant pour les 32 ARN de transfert dont la petite taille (de l'ordre de 70 mers) ne permet de choisir qu'une seule sonde. De plus il a semblé important de choisir une sonde supplémentaire pour 102 gènes d'intérêt par rapport à notre problématique. Il s'agit des 55 gènes impliqués dans le métabolisme des acides aminés, des 18 gènes intervenant dans les phénomènes de transport {Shigenobu, 2000 #381}, auxquels ont été ajoutés les 26 gènes qui codent pour le flagelle et les 2 gènes qui codent pour les porines (*ompA* et *ompF*), en raison de leur rôle potentiel dans les phénomènes de transport. Enfin, les gènes situés sur les deux plasmides et qui n'appartiennent pas aux catégories précédentes ont également été ajoutés, en raison de l'intérêt particulier des plasmides dans la problématique du métabolisme des acides aminés.

Le premier paramètre essentiel qui guide le choix de la taille des sondes est le rendement d'hybridation. Ce rendement augmente avec le nombre (N) de nucléotides de la sonde. En effet, plus N est élevé, plus la probabilité de réaction inverse est faible et plus la vitesse de formation de l'hybride est rapide (la probabilité d'hybridation augmente en fonction de $N^{0,5}$) (Maskos et Southern, 1992). Certaines expériences sur puces à ADN ont donc été réalisées avec des sondes de 70 mers qui assurent une sensibilité équivalente à celle qui est obtenue avec des sondes d'ADNc (Wang *et al.*, 2003). Malheureusement, les sondes de 70 mers sont difficiles à synthétiser. Certains auteurs préconisent donc l'utilisation de sondes de 60 mers (Hughes *et al.*, 2001). Des études expérimentales montrent cependant qu'elles sont beaucoup moins spécifiques que des sondes plus courtes, et que le meilleur compromis entre sensibilité et spécificité est obtenu avec des sondes de 30 (Relogio *et al.*, 2002) ou de 40 mers (Maskos et Southern, 1992). Un autre paramètre, qui intervient dans le choix de la taille des sondes, est le nombre de sondes potentielles qui existent dans un génome. Ce nombre décroît de façon linéaire avec la taille des sondes (Wang et Seed, 2003), ce qui implique de choisir des sondes d'autant plus petites que le génome étudié contient peu de gènes (Li et Stormo, 2001). *Buchnera* possède un génome de taille réduite et les risques de contamination par le matériel génétique du puceron ne peuvent pas être négligés. Il semble que des sondes de 30 à 40 mers offrent une bonne spécificité sans perte importante de la sensibilité et sont beaucoup moins coûteuses. Les variations des valeurs de Tm entre les différentes sondes sont évidemment plus importantes pour des sondes courtes (Nielsen *et al.*, 2003), mais elles sont gérées par le logiciel ROSO. Une taille de 35 mers a donc finalement été retenue, d'autant plus qu'une étude préliminaire sur une mini-puce (cf. 2.1) a permis de vérifier l'intérêt de ce choix expérimentalement.

En ce qui concerne la localisation des sondes sur les gènes, elle est soumise à deux contraintes. La première concerne la méthode de marquage qui est utilisée. Il s'agit d'une méthode indirecte, basée sur l'utilisation d'un mélange d'amorces semi-spécifiques choisies parmi les cent dernières paires de bases situées à l'extrémité 3' des séquences d'intérêt, et d'amorces aléatoires réparties sur l'ensemble du génome (cf. partie C). Ce type de marquage impose donc une sélection des sondes en dehors de cet intervalle. La seconde contrainte est liée au fait que l'assignation des ORF n'a pas été vérifiée expérimentalement chez *Buchnera*. Compte tenu de la richesse en AT du génome, le codon *start* peut ne pas être localisé de façon correcte chez certains gènes. Pour cette raison les cinquante premières bases situées à l'extrémité 5' des gènes ont donc également été éliminées de l'analyse (cf **Figure B.1.13**).

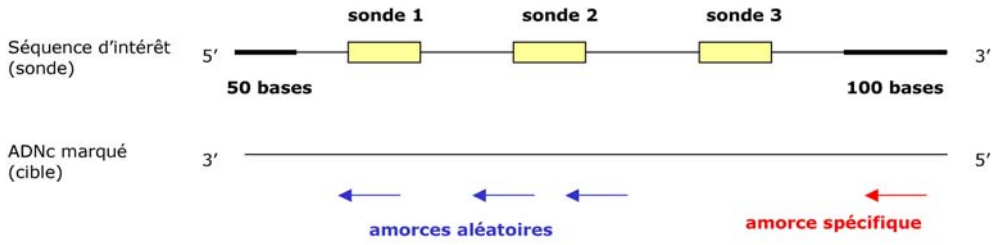


Figure B.1.13 Localisation des différentes sondes sur les gènes de *Buchnera*.

1.2.3 La démarche d'optimisation

Une étape préliminaire est réalisée pour rechercher l'intervalle de Tm optimal. Pour cela une recherche de trois sondes par gène est réalisée avec les paramètres par défaut de ROSO. Elle permet d'obtenir 1154 sondes dont les valeurs de Tm sont comprises entre 64,4 °C et 83,4 °C, pour un Tm moyen de 75 °C. Sur cet ensemble de sondes, un intervalle réduit entre 69 et 80 °C ne conduit à l'exclusion que de 77 sondes (6,7 %). Ces valeurs de Tm sont donc retenues pour réaliser une véritable recherche de sondes en cinq étapes (cf. **Tableau B.1.8**). Après chaque étape, les séquences des gènes pour lesquels aucune sonde n'est disponible, ou les gènes ayant des sondes avec une mauvaise spécificité (taux de similitude supérieur à 80 %), sont automatiquement conservées dans un nouveau fichier qui est utilisé pour l'étape de recherche suivante.

Tableau B.1.8 Présentation des cinq étapes de la démarche de recherche des sondes pour *Buchnera*. Les différents critères qui sont utilisés sont : (1) N4 : absence de séquences de 4 nucléotides identiques, (2) localisation : deux zones sont exclues 50 bases en 5' et 100 bases en 3', (3) seuil de rejet des énergies libres de formation des épingles à cheveux et des homodimères (en kcal.mol^{-1}).

N°	N4	Localisation	Épingle à cheveux	Homodimère	Nombre de sondes obtenues	
1	Non	Oui	2	-6	836	
2	Oui	Oui	2	-6	218	85,6 %
3	Oui	Non	2	-6	34	
4	Oui	Non	0	-8	166	13,1 %
5	Oui	Non	-2	-10	17	1,3 %

La localisation des sondes obtenues a été vérifiée. Pour les gènes de grande taille, et surtout pour ceux qui possèdent un taux élevé de GC, une attention particulière a été accordée pour retenir de préférence des sondes dans un intervalle compris entre les positions 100 et 900 à partir de l'extrémité 5'. En effet, les amorces spécifiques étant localisées à proximité de l'extrémité 3', la réaction de synthèse des ADNc risque d'être interrompue de façon prématurée

au-delà de neuf cents paires de bases. De plus, pour certains gènes les seules sondes optimales sont localisées sur l'intervalle normalement exclu des cent premières bases situées à l'extrémité 3'. Pour l'ensemble de ces sondes, il a été vérifié que les amorces spécifiques étaient toujours localisées en amont des sondes. Enfin, les sondes ont été choisies de préférence de façon à ne pas se chevaucher sur le gène et même à être les plus éloignées possible. Les cinq étapes de recherche, suivies de ce travail de filtration des sondes surnuméraires, ont finalement permis de sélectionner 1271 sondes.

En raison des caractéristiques particulières des 32 gènes codant pour les ARN de transfert, leur étude a fait l'objet d'une analyse séparée. En effet, les ARN de transfert sont tous de petite taille (de l'ordre de 70 mers), ce qui rend difficile la recherche d'une région spécifique. De plus, compte tenu de leur fonction, certaines régions forment des structures secondaires très stables. Enfin, leur richesse en GC conduit à des valeurs élevées de Tm. Pour obtenir des sondes en accord avec l'intervalle de Tm initialement retenu il a été nécessaire de réduire la taille des sondes. Des étapes de recherche ont donc été réalisées successivement pour des tailles de sondes de 35, 32, 30 et 28 mers. Une seule sonde a ensuite été retenue, avec le taux d'identité le plus faible et dans la mesure du possible la taille la plus importante, pour limiter au maximum les différences de taille entre les sondes. Quant à l'absence de structures secondaires, elle n'a pu être prise en compte qu'en dernier lieu.

1.2.4 Résultats

1.2.41 Caractéristiques du jeu de sondes pour Buchnera

Le jeu complet de sondes spécifiques de *Buchnera* est constitué de 1303 oligonucléotides répartis sur 617 gènes. En effet, le plasmide Tryptophane porte deux répétitions en tandem des gènes *trpG* (identique à *trpG*2) *trpE* (identique à *trpE2*). Les 32 gènes codant pour les ARN de transfert possèdent une sonde, les 102 gènes d'intérêt possèdent trois sondes et les gènes restant possèdent deux sondes. L'ensemble de ces sondes est réparti sur un intervalle de Tm compris entre 69 et 80 °C (cf. **Figure B.1.14**) et 95,5 % des sondes sont parfaitement spécifiques de leur cible. Seul 0,7 % des sondes présentent des risques d'hybridation aspécifique et les dix sondes concernées représentent toutes des ARN de transfert (cf. **Figure B.1.15**).

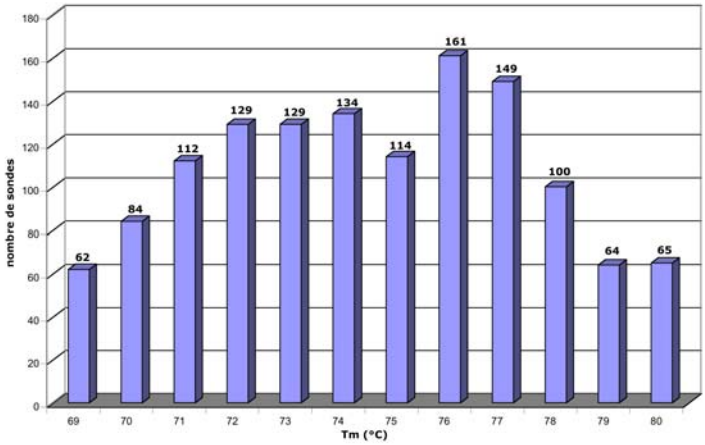


Figure B.1.14 Répartition des sondes de *Buchnera* en fonction des valeurs de Tm.

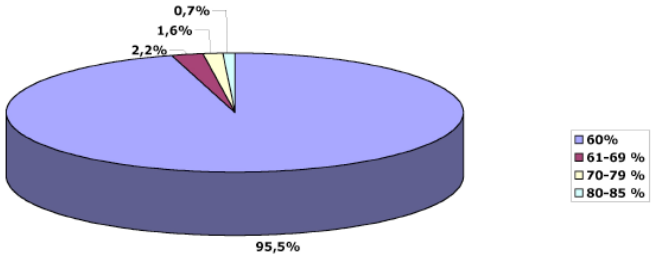


Figure B.1.15 Répartition des sondes de *Buchnera* en fonction des taux de similitude. Seules les sondes qui présentent des taux de similitude supérieurs à 80 % possèdent des risques réels d'hybridation aspécifique.

1.2.42 Caractéristiques des sondes de contrôle

ROSO a également été utilisé pour choisir onze sondes de contrôle correspondant à trois types de témoins à intégrer sur la puce *Buchnera*. Toutes ces sondes ont été déterminées avec les paramètres de la première étape du processus d'optimisation (cf. **Tableau B.1.8**) et dans l'intervalle de valeur de Tm défini pour les sondes *Buchnera* (paramètres optimaux).

Pour le choix des trois témoins positifs, un fichier d'intérêt contenant les deux séquences des gènes *pcp* et *pelK* de *Bacillus subtilis* et celle du gène *pelL d'Erwinia chrysanthemi* a été utilisé contre le fichier de séquences externes pour l'étude de la spécificité. Pour chaque gène, il a été vérifié que les sondes retenues étaient localisées en amont des amorces spécifiques.

Pour les témoins d'hybridation aspécifique, les quatre gènes de *Buchnera* utilisés pour l'expérience préliminaire de mini-puce (cf. 2.1) ont été retenus (*aroH*, *eno*, *ilvH*, et *pheA*). Sur la première sonde (la sonde la plus proche de l'extrémité 5') de chacun de ces gènes, cinq mutations aléatoires ont été introduites de façon régulière (cf. **Figure B.1.16**).

Sonde normale	CAAGAACTACTGAAAGTCAAATTCATAGAGAAATG
Sonde témoin	CAAGACCTACTTAAAGTGAAATTAATAGACAAATG

Figure B.1.16 Exemple de la sonde normale et de la sonde témoin pour le gène *aroH*.

Pour les témoins négatifs, c'est-à-dire les sondes sur lesquelles il n'existe théoriquement aucune possibilité d'hybridation, quatre sondes ont été définies (TN1 à TN4) à partir d'un fichier de séquences aléatoires dont la spécificité a été testée contre le fichier de séquences externes.

1.2.5 Discussion

Compte tenu du mode de vie symbiotique de la bactérie *Buchnera*, il existe un risque d'hybridation croisée avec des séquences de puceron. Cependant, seul un faible nombre de séquences de puceron sont actuellement disponibles. Ces dernières n'ont donc pas été ajoutées au fichier de séquences externes. En revanche, les sondes obtenues ont été testées avec *BLASTN* contre les séquences de *Drosophila melanogaster*. Cette analyse indique qu'aucune sonde ne présente un risque potentiel d'hybridation aspécifique. De plus, l'utilisation de sondes de petite taille assure une bonne spécificité d'hybridation. Une nouvelle vérification pourra néanmoins être réalisée lorsque le génome du puceron sera disponible. En ce qui concerne les risques d'hybridation aspécifique avec des séquences issues d'autres bactéries, ils restent limités. En effet, le clone *LL01* d'*Acyrthosiphon pisum* utilisé au laboratoire est dépourvu de symbiotes secondaires, et les bactéries risquant de contaminer les échantillons ont des génomes très riches en GC, contrairement à la bactérie *Buchnera*. Enfin, les éventuels problèmes d'hybridation, qui peuvent affecter la réponse de certaines sondes (Lockhart *et al.*, 1996 ; Wodicka *et al.*, 1997), sont minimisés car chaque gène est représenté au minimum par deux sondes sur la puce (sauf dans le cas particulier des ARN de transfert).

En ce qui concerne la démarche générale de choix des sondes, il est important de noter qu'il s'agit typiquement d'une recherche multicritère. Cette dimension est difficile à prendre en compte, d'autant plus qu'elle implique la plupart du temps des contraintes très spécifiques liées à l'organisme étudié, aux connaissances associées ou encore aux techniques utilisées. De plus, pour certains gènes (comme les gènes paralogues, les gènes issus de familles multigéniques, ou encore les transcrits alternatifs …), il peut parfois être utile de disposer de différentes sondes répondant à des contraintes plus ou moins fortes imposées aux différents critères de sélection (localisation, structures secondai-

res etc.). Il est certain que cette démarche doit être l'objet d'une réflexion personnelle propre à chaque utilisateur, comme celle présentée ici pour *Buchnera*. Afin de les aider, une démarche générale d'optimisation a été définie d'après les résultats obtenus pour *Buchnera* mais aussi pour d'autres organismes d'étude (cf. **Tableau B.1.5**). Cette démarche est disponible avec des explications d'utilisation sur le site Web de ROSO. Il s'agit d'un processus d'optimisation en cinq étapes que les utilisateurs peuvent modifier et complexifier en fonction de leurs propres problématiques. Deux types de paramètres ont été définis en fonction du taux de GC de l'organisme d'étude (cf. **Tableau B.1.9**). À l'issue de chaque étape, les utilisateurs doivent conserver les fichiers de sortie contenant les gènes pour lesquels aucune sonde n'a été définie et, s'ils le désirent, les fichiers contenant des sondes présentant un taux de similitude trop élevé (au-dessus de 80 %). Ces deux fichiers sont concaténés et réutilisés pour l'étape suivante à la place du fichier d'intérêt. Un algorithme d'automatisation de ces étapes qui utilise les paramètres présentés ici, a été développé. Il n'est disponible que pour les utilisations en ligne de commande afin de limiter les problèmes de saturation du serveur du PBIL.

Tableau B.1.9 Démarche générale d'optimisation permettant une utilisation de ROSO en plusieurs étapes successives. Les seuils de rejet des énergies libres de formation des structures secondaires (en kcal.mol^{-1}) dépendent du taux de GC de l'organisme d'étude contrairement aux modifications de la taille des sondes (en mers).

N°	Épingles à cheveux		Homodimère		Taille
	GC<55 %	GC>=55 %	GC<55 %	GC>=55 %	
1	0	-4	-6	-10	Initiale
2	-2	-8	-8	-14	Initiale
3	-4	-12	-10	-20	Initiale
4	-4	-12	-10	-20	Initiale-10
5	-4	-12	-10	-20	Initiale-20

1.2.6 Conclusion

L'utilisation de ROSO, pour déterminer les sondes nécessaires à la conception de la puce *Buchnera,* a permis l'élaboration d'une stratégie de recherche originale des sondes dédiées à l'analyse du transcriptome. Cette stratégie est bien sûr à adapter en fonction des problématiques des différents utilisateurs, mais elle permet ainsi l'intégration de contraintes qui peuvent difficilement être prises en compte par un algorithme. L'étape ultime de validation de ce travail est maintenant expérimentale. Il s'agit à présent de définir un plan de dépôt optimal à partir des 1314 sondes retenues pour finalement aboutir à la fabrication et à l'utilisation de la puce *Buchnera*.

2

2 La puce *Buchnera*

« En essayant continuellement on finit par réussir. Donc: plus ça rate, plus on a de chance que ça marche. »
Jacques Rouxel[24]

Bien que de nombreux protocoles expérimentaux soient disponibles, seul un nombre limité d'informations méthodologiques sont publiées en ce qui concerne la conception des lames, leur utilisation, l'acquisition des images et la filtration des données (Relogio *et al.*, 2002). Or, des images de bonne qualité et un rapport signal sur bruit élevé sont les pré-requis nécessaires à l'obtention de résultats valables (Yu *et al.*, 2002). Pour les obtenir, il est nécessaire d'utiliser des protocoles reproductibles et adaptés aux problématiques du modèle d'étude. Compte tenu du coût de fabrication (Holloway *et al.*, 2002) et d'utilisation des puces et de l'évolution des protocoles, une étude méthodologique a été réalisée en deux temps. Un travail préliminaire a été réalisé par Federica Calevro sur une mini-puce pour déterminer les aspects pratiques de fabrication et les grandes étapes d'utilisation de la puce *Buchnera* (Calevro *et al.*, 2004). Les résultats obtenus ont permis la conception de la puce complète qui a été utilisée pour achever la validation des protocoles d'utilisation.

2.1 Une mini-puce devenue grande

La mini-puce qui a été développée contenait quatre gènes intervenants dans le métabolisme des acides aminés chez *Buchnera* et de nombreux témoins, représentant au total une centaine de plots (cf. **Figure B.2.1**).

[24] Rouxel, J. ***Les Shadoks.***

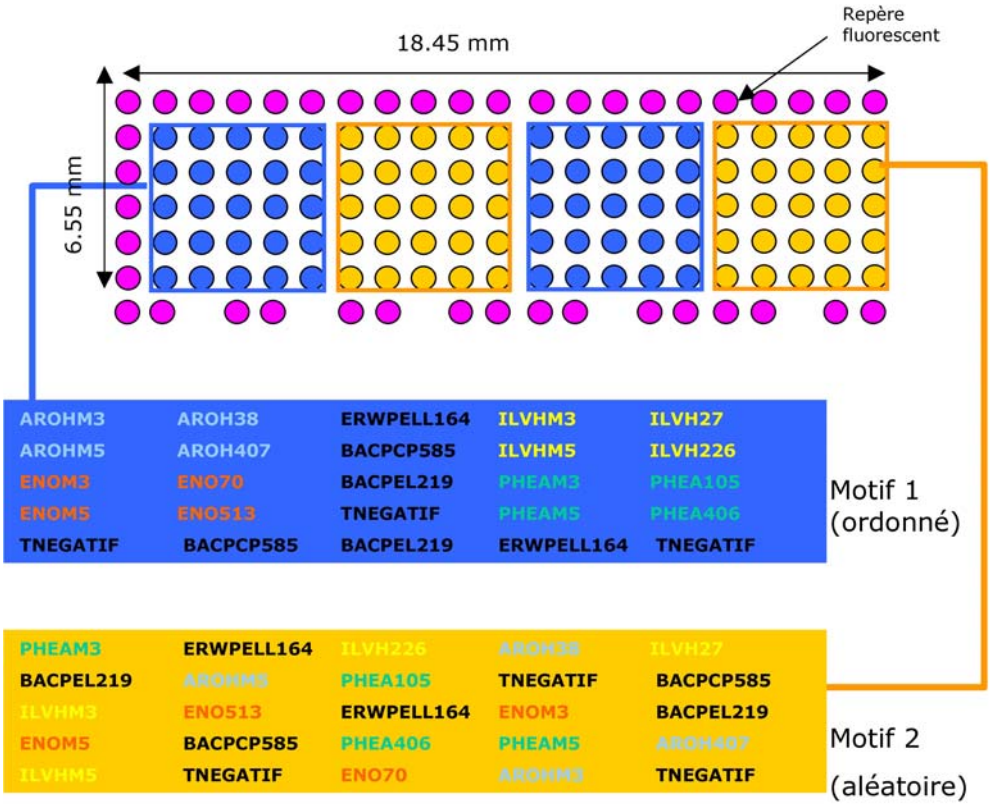


Figure B.2.1 Plan de la première mini-puce *Buchnera*.

Cette mini-puce a permis de définir les principaux critères de conception de la puce *Buchnera* complète. Les solutions de dépôt suivantes ont ainsi été testées : une solution de citrate de sodium salin SSC 3X contenant de la bétaïne (1,5 M), une solution de DMSO à 20, 30 et 40 % et un tampon de phosphate de sodium (150 mM, pH 8.5). Les meilleurs plots ont été obtenus avec la solution (SSC 3X, bétaïne 1,5 M). En effet, l'utilisation de bétaïne permet d'augmenter l'efficacité de liaison des sondes sur la lame et offre une meilleure homogénéité des plots en raison d'une évaporation limitée (Diehl *et al.*, 2001). Par ailleurs, la bétaïne réduit la différence de stabilité d'hybridation qui existe entre les bases AT et les bases CG (Rees *et al.*, 1993). De même des concentrations comprises entre 20 et 80 µM ont été testées pour la solution de dépôt. Les résultats obtenus indiquent que l'intensité d'hybridation augmente avec la concentration en sondes, ce qui laisse penser que même à 80 µM, le nombre de sondes disponibles pour l'hybridation n'est pas significativement en excès par rapport à la concentration en cibles.

La mini-puce a également permis de déterminer les méthodes d'utilisation conduisant aux meilleurs résultats pour *Buchnera*. Cette étude

montre qu'une étape de pré-hybridation permet d'augmenter de façon significative le rapport signal sur bruit de fond, et qu'une méthode de marquage indirecte aboutit à l'obtention de signaux significativement plus élevés et plus spécifiques qu'une méthode de marquage directe. Enfin, en ce qui concerne la méthode de marquage indirect, des expériences de RT-PCR réalisées sur les quatre gènes étudiés montrent que l'utilisation d'amorces spécifiques accroît de façon significative le rendement de l'étape de transcription inverse.

Cette mini-puce a donc permis de déterminer la nature (SSC 3X, bétaïne 1,5 M) et la concentration (120 µM)de la solution de dépôt à utiliser pour la conception de la puce *Buchnera*. Elle a également montré l'intérêt d'une étape de pré-hybridation et d'une méthode de marquage indirect avec des amorces spécifiques. Ces différents aspects ont donc été retenus pour la production de la puce complète.

2.2 La puce complète

Ma contribution personnelle a porté sur la seconde partie des développements méthodologiques, avec la conception de la puce *Buchnera* complète et la définition des principaux critères techniques, méthodologiques et statistiques concernant son utilisation. L'ensemble de cette étude a abouti à la mise au point du protocole complet d'utilisation de la puce qui est donc logiquement présenté dans la partie « matériel et méthodes » de la partie C. Seuls les principaux résultats méthodologiques qui ont été obtenus sont présentés succinctement dans cette partie.

La puce complète a tout d'abord permis de valider les choix techniques concernant le type de lames, d'hybridation et d'échantillon à utiliser. Pour cela deux types de lames de verre ont été comparés. Il s'agit des lames ROSA® (*RosaTech* SA, Ecully, France) et des lames QUANTIFOIL® aldéhyde (*Quantifoil Micro Tools*, Jena, Allemagne). Les premières portent des acides activés (Dugas *et al.*, 2004), et les secondes des groupements aldéhydes[25] qui peuvent réagir avec un groupe amine porté par les sondes pour former une liaison covalente très stable chimiquement (formation d'une base de Schiff). Ces lames ont été utilisées pour réaliser trois types d'hybridations. Les deux premières sont des hybridations manuelles qui ont été réalisées avec deux lames QUANTIFOIL® (hybridation 1), puis avec deux lames ROSA® (hybridation 2). Pour chacune de ces hybridations, une lame a été hybridée avec des ARN issus de

[25] http://www.quantifoil.com.

bactéries obtenues par filtration et la seconde avec des ARN issus directement de pucerons. L'hybridation sur lames QUANTIFOIL® (hybridation 1) a été réalisée en l'absence des cibles de contrôle de façon à tester les risques d'hybridation aspécifique. Une hybridation automatique a finalement été réalisée avec deux lames QUANTIFOIL® (hybridation 3) sur lesquelles les échantillons d'ARN sont issus de bactéries obtenues par filtration. Les hybridations ont été réalisées à 45 °C sur la première lame et à 50 °C sur la seconde.

La comparaison des hybridations réalisées sur les lames ROSA® et les lames QUANTIFOIL® montre que les lames QUANTIFOIL® offrent à la fois une réduction significative du bruit de fond (visible sur l'ensemble de la lame) et une meilleure sensibilité (notamment pour les marquages réalisés avec le Cy3). De plus, avec une méthode d'hybridation manuelle, les lames QUANTIFOIL® permettent d'obtenir en moyenne 39 % de bons plots contre seulement 23 % pour les lames ROSA®.

En ce qui concerne la spécificité, les résultats obtenus sur les deux types de lames indiquent qu'il est préférable d'utiliser des échantillons obtenus à partir de filtration de pucerons.

La quantité de cibles de contrôle à ajouter aux échantillons a également été déterminée. Pour l'hybridation 1 les cibles de contrôle n'ont pas été ajoutées de façon à tester la spécificité des sondes. Les résultats montrent qu'il n'existe pas d'hybridation aspécifique sur les sondes de contrôle (contamination par du matériel génétique du puceron par exemple). L'hybridation 2 a permis de tester l'ajout de cibles en quantité égale dans les deux marquages et l'hybridation 3 des ajouts en rapport 1 : 3 et 1 : 5 (cf. **Tableau B.2.1**). Compte tenu des intensités obtenues par rapport aux signaux des autres sondes réparties sur la lame, il est nécessaire de diviser par dix les quantités les plus faibles utilisées pour chacun des témoins au cours des différentes expériences.

Enfin, la méthode d'hybridation automatique qui utilise un système original permet de maintenir la solution de cibles en mouvement continuel sur la lame. Elle assure ainsi une hybridation parfaitement homogène sur l'ensemble de la lame, contrairement à l'hybridation manuelle qui génère des zones d'hybridation variables, avec notamment une zone circulaire fortement hybridée au centre de la lame par rapport aux bords très faiblement hybridés. L'hybridation automatique a donc été retenue, d'autant plus qu'elle permet d'obtenir en moyenne 63 % de bons plots, 30 % de plots non visibles et seulement 7 % de plots indexés mauvais. Pour les deux températures d'hybridation testées, aucun signal d'hybridation aspécifique n'est observé sur les témoins négatifs et sur les témoins d'hybridation aspécifique. La température de 45 °C a été retenue car elle permet d'assurer (en présence de formamide) une bonne spécificité d'hybridation.

Tableau B.2.1 Quantités de cibles de contrôle ajoutées pour les trois types d'hybridation testées : lame QUANTIFOIL® avec hybridation manuelle (1) ou automatique (3) et lames ROSA® avec hybridation manuelle (2).

	Hybridation	Cy3	Cy5
PelL	1	0	0
	2	50 ng (1)	50 ng (1)
	3	500 ng (5)	100 ng (1)
PelK	1	0	0
	2	150 ng (1)	150 ng (1)
	3	120 ng (3)	40 ng (1)

D'un point de vue méthodologique, la puce complète a permis de définir et rédiger un protocole d'acquisition des images pour le scanner *GeneTAC™ L SIV* (*Genomic Solutions*, Huntingdon, UK) ainsi qu'un protocole de filtration et d'acquisition des données pour le logiciel *GenePix Pro® 6.0* (*Axon Instruments*, Union City, CA, USA). Ces protocoles sont à présent disponibles pour les utilisateurs de la plateforme transcriptome Rhône-Alpes.

Enfin d'un point de vue statistique, une étude a été réalisée sur les coefficients de variation (CV) calculés en divisant la moyenne des intensités des plots (μ) par l'erreur standard (σ) pour chaque sonde (quatre répétitions), puis pour chaque gène (deux ou trois sondes par gène). Elle a permis de définir une procédure destinée à conserver la meilleure valeur d'intensité possible pour chaque gène. Pour cela, les intensités des plots disponibles pour chaque sonde sont moyennées en fonction d'un indice qualité attribué lors de l'étape de filtration (les valeurs présentant l'indice qualité le plus élevé sont moyennées et les autres ne sont pas utilisées). Ce calcul permet de conserver une valeur d'intensité pour chaque sonde. Les valeurs obtenues pour les différentes sondes correspondant à un même gène montrent que le comportement des sondes ne dépend pas significativement de leur localisation sur le gène. La procédure de moyennage par indice de qualité est donc appliquée pour les sondes représentant un même gène. Elle permet d'obtenir une valeur unique de signal d'intensité pour chaque gène en augmentant considérablement la quantité des données. Pour les lames QUANTIFOIL® en hybridation automatique, cette procédure permet en effet d'obtenir en moyenne 93 % de bons signaux, 5,5 % de signaux non visibles et 1,5 % de signaux indexés mauvais.

2.3 Conclusion

La mini-puce a permis de définir les principaux critères de conception de la puce *Buchnera* complète, qui a elle-même été utilisée pour achever les développements méthodologiques. Les résultats ont montré que l'utilisation de

Buchnera obtenues par filtration est préférable à celle de pucerons entiers et ont permis de déterminer les quantités et les rapports de cibles de contrôle à ajouter aux échantillons. Pour l'hybridation à proprement parler, la puce a permis de sélectionner les lames QUANTFOIL® hybridées de façon automatique et d'optimiser les conditions de lavages. Par ailleurs, bien qu'il soit impossible de décider *a priori* d'une méthode d'acquisition ou de filtration universelle, la puce a été utilisée pour mettre au point un protocole permettant de systématiser ces étapes quel que soit l'utilisateur. L'ensemble des protocoles obtenus offre à présent un cadre permettant de réaliser des expérimentations de qualité sur la puce *Buchnera* et d'analyser les données obtenues de façon correcte. La première utilisation de cette puce pour répondre à une question biologique est présentée dans la dernière partie de cette thèse et concerne l'étude de l'expression des gènes chez *Buchnera* en réponse à des modifications de la composition des milieux nutritionnels des pucerons.

Partie C
Analyse du transcriptome de *Buchnera aphidicola*

1

1 Introduction

« Il y a bien moins de difficulté à résoudre un problème qu'à le poser. »
Joseph de Maistre[26]

1.1 Contexte

Deux approches complémentaires ont été menées au laboratoire pour étudier le métabolisme des acides aminés et sa régulation chez *Buchnera*.

Une première expérience a été réalisée pour comparer les profils d'expression de bactéries issues de pucerons élevés soit sur un milieu artificiel complet (milieu AP3), soit sur un milieu dépourvu en deux acides aminés aromatiques essentiels : la tyrosine et la phénylalanine (milieu YF^0). Pour cela, huit lames ont été réalisées au sein d'un plan expérimental en flip-flop. Il s'agit d'une expérimentation qui permet de tester la réponse des bactéries dans des conditions de carence très ciblées et drastiques, avec pour but la détection éventuelle d'une régulation transcriptionnelle de certaines enzymes impliquées dans la biosynthèse des deux acides aminés absents du milieu nutritionnel. En revanche, modifier un paramètre de façon aussi drastique dans un système complexe implique le risque d'une perturbation importante d'un grand nombre de paramètres (par exemple l'induction d'une réponse de stress), qui risque de masquer la réponse d'intérêt. Les résultats de cette première expérience, réalisée en partie par Federica Calevro, sont présentés dans une publication en préparation et ne sont pas abordés ici.

Pour répondre de façon plus fine à la question biologique de l'adaptation de la production des acides aminés chez *Buchnera* en fonction des conditions environnementales du puceron, cette première étude a donc été suivie d'une seconde expérience, dans laquelle des modifications plus fines de composition du milieu nutritionnel du puceron ont remplacé les perturbations drastiques initiales. Cette seconde expérience fait l'objet de ce chapitre.

[26] de Maistre, J. et Peyrefitte, A. *Considérations sur la France.* Imprimerie Nationale, Paris (1993).

1.2 Problématique

L'expérimentation qui est présentée ici a été réalisée en collaboration avec le laboratoire de biologie de l'Université d'York (RU). Le matériel biologique a été préparé à York et les hybridations sur puces ont été réalisées sur la plateforme transcriptome à Lyon. Les conditions utilisées dans cette expérience sont de véritables conditions physiologiques. Il s'agit en effet de tester la réponse de la bactérie à une variation modérée du milieu environnemental du puceron. Pour cela, l'influence des deux composants principaux du phloème est étudiée, à savoir, la concentration en saccharose, source essentielle de carbone et responsable de la pression osmotique de la sève, et le rapport entre acides aminés essentiels et non essentiels. Ces deux facteurs permettent d'étudier la relation symbiotique dans la mesure où, la composition en acides aminés permet de tester l'influence de la modification du milieu extérieur sur l'expression des gènes chez *Buchnera,* alors que la concentration en saccharose permet de prendre en compte des variations de la demande métabolique de l'hôte. Le but de l'étude simultanée de ces deux composants est d'observer à la fois les différences d'expression causées par chacun des deux facteurs pris séparément, et les effets simultanés des deux facteurs, c'est-à-dire lorsque ces effets sont en interaction.

Deux compositions en acides aminés et deux concentrations en saccharose ont été utilisées pour réaliser quatre types de milieux nutritionnels pour l'élevage des pucerons (cf. **Tableau C.1.1**). Les compositions en acides aminés essentiels sont respectivement de 50 % (rapport de 1 : 1 entre acides aminés essentiels et non essentiels) et 25 % (rapport de 1 : 3 entre acides aminés essentiels et non essentiels) pour une concentration totale dans les deux cas de 150 mM. Les concentrations en saccharose utilisées sont de 0,5 et 1 M. Parmi ces quatre milieux, le milieu « de référence » est le milieu 1 qui correspond à une concentration totale en acides aminés de 150 mM (dont 50 % d'acides aminés essentiels) et une concentration de saccharose de 0,5 M. Il s'agit du milieu permettant un développement optimal des pucerons.

Tableau C.1.1 Présentation des quatre milieux nutritionnels utilisés pour l'élevage des pucerons.

		Composition en acides aminés essentiels	
		50 %	25 %
Concentration	0,5 M	1	2
en saccharose	1M	3	4

1.3 Plan expérimental

Le choix du plan expérimental représente sans doute l'une des étapes les plus importantes de l'expérimentation, car elle conditionne l'ensemble des résultats obtenus (Fisher, 1951). Pour cette expérience, le plan retenu comporte seize lames (huit lames toutes réalisées en double au cours de deux étapes distinctes d'hybridation). Pour chaque lot de huit lames, quatre répétitions techniques de chaque échantillon ont été réalisées. L'ensemble de l'expérience contient donc deux répétitions biologiques pour chaque échantillon. Il s'agit d'un plan équilibré qui permet la comparaison des différentes modalités (cf. **Figure C.1.1**). Au sein de ce plan, les modalités sont équilibrées non seulement en nombre mais également pour chacun des deux fluorochromes (cf. **Tableau C.1.2**).

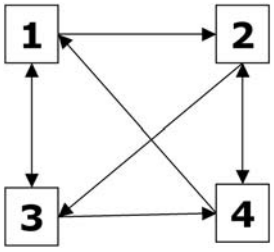


Figure C.1.1 Schéma du plan expérimental associant les huit lames (répété deux fois). Chaque flèche représente une hybridation sur une même lame et les chiffres correspondent aux milieux nutritionnels sur lesquels ont été élevés les pucerons.

Tableau C.1.2 Présentation des hybridations réalisées sur les huit lames de l'étude avec le fluorochrome associé à chaque échantillon.

Lames	1	2	3	4	5	6	7	8
Cy3	1	2	3	4	1	2	4	3
Cy5	2	3	4	1	3	4	2	1

2

2 Matériel et méthodes

« Se donner du mal pour les petites choses, c'est parvenir aux grandes, avec le temps. »
Samuel Beckett[27]

2.1 Préparation des lames

2.1.1 Dépôt des sondes

Les sondes sont des oligonucléotides, synthétisés par la société *Eurogentec* (Hampshire, RU). Ces oligonucléotides sont simplement purifiés par la méthode *SePop* (*Selective Precipitation Optimized Process*). Cette méthode, dite de désalage, permet d'éliminer les impuretés inorganiques, et notamment les produits nécessaires à la synthèse. Elle offre une pureté comprise entre 75 et 85 % pour des sondes de 35 mers, ce qui est suffisant pour une utilisation sur puce à ADN, d'autant plus que des études montrent que les performances de ces oligonucléotides sont tout à fait comparables à celles des oligonucléotides purifiés par une étape supplémentaire de HPLC (Relogio *et al.*, 2002). Leur extrémité 5' est couplée à un bras carboné (C_6) qui les éloigne de la lame afin de faciliter leur accessibilité au cours de l'hybridation. Ce bras porte une amine terminale qui permet la formation d'une liaison covalente avec la lame. Ces sondes sont stockées dans 14 plaques « 96 puits ». Une redistribution en 7 plaques « 384 puits » a été réalisée par *Eurogentec* pour permettre leur utilisation par le robot de dépôt. La solution de dépôt est un tampon salin de citrate de sodium (SSC) 3X (*Euromedex*, Mundolsheim, France) contenant de la bétaïne (N,N,N –trimethylglycine) 1,5 M (*Sigma-Aldrich*, Saint Louis, MO, USA). Des emplacements ont été conservés vides dans chacune des plaques « 384 puits » de façon à permettre l'ajout manuel de sondes de contrôle ou simplement de solution tampon (estimation du bruit de fond).

[27] Beckett, S. *Molloy*. Éditions de Minuit, Paris (1982).

Les sondes sont déposées sur des lames QUANTIFOIL® (*Quantifoil*) par le robot *MicroGrid II Pro* (*BioRobotics,* Cambridge, RU) au moyen de quatre aiguilles (*MicroSpot 2500 pins, BioRobotics*) et à raison de 0,6 nl de solution à 120 µM par plot. Au cours du dépôt, la température et le niveau d'humidité sont maintenus respectivement à 20 °C ± 1 °C et à 50 % ± 3 %. Le diamètre moyen des plots obtenus dans ces conditions est de 180 µm.

Toutes les sondes sont déposées par paire et par deux aiguilles différentes, de façon à permettre l'analyse d'un effet aiguille. Le dépôt s'effectue du bas vers le haut de la lame. L'ensemble des sondes n°1 (localisées à l'extrémité 5'des gènes) est déposé en bas de la lame, les sondes n°2 (localisées en position médiane sur les gènes) sont au centre de la lame et les sondes n°3 (les sondes supplémentaires pour les gènes d'intérêt, plus proche de l'extrémité 3') sont situées au-dessus. Au sein de ce plan de dépôt, des plots réservés pour les témoins (quatre témoins d'hybridation aspécifique, trois témoins positifs, quatre témoins négatifs et de nombreux plots contenant du tampon) sont régulièrement répartis dans chacun de ces trois ensembles. Enfin, le haut de la lame est réservé aux témoins positifs (*pelL*, *pelK*) et à de nouveaux témoins de normalisation, qui n'ont pas été utilisés dans cette expérience. Il s'agit du témoin *pcp* et des témoins commerciaux *ArrayControl™ (Amersham,* Piscataway, NJ, USA). Des emplacements vides ont également été réservés pour permettre l'ajout de nouveaux témoins. Au total, la puce complète contient 5430 plots. Ces plots sont répartis sur quatre colonnes (correspondant aux quatre aiguilles de dépôt) et sept lignes (correspondant aux sept plaques « 384 puits ») (cf. **Figure C.2.1**). La huitième ligne actuellement incomplète contient les témoins de normalisation (stockés dans une plaque commune à l'ensemble des utilisateurs de la plateforme trancriptome Rhône-Alpes).

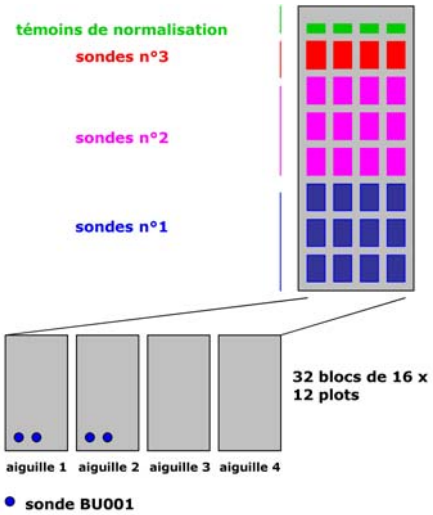


Figure C.2.1 Plan de dépôt de la puce *Buchnera*.

2.1.2 Traitement des lames

Une fois le dépôt achevé, les lames sont incubées 15 minutes à 37 °C, en atmosphère humide puis 90 minutes à 120 °C pour fixer les sondes. Une étape de rinçage permet ensuite d'éliminer les sondes non liées aux lames. Pour cela, les lames sont réhydratées quelques secondes au-dessus d'un bain de vapeur puis lavées 5 minutes dans du sodium dodecyl sulfate (SDS) 0,2 % à température ambiante et dans de l'eau ultra-pure 3 fois 2 minutes à température ambiante puis 2 minutes à 95 °C. Finalement, les sites d'hybridation aspécifiques sont bloqués. Pour cela, les lames sont incubées 15 minutes à température ambiante dans un tampon de borhydrure de sodium 0,25 % [$NaBH_4$ (*Sigma-Aldrich*), 0,5 g ; PBS 1X, 150 ml ; ETOH 99 %, 50 ml]. Elles sont ensuite rincées à température ambiante, 2 fois 2 minutes dans du SDS 0,2 %, et 3 fois 2 minutes dans de l'eau ultra-pure.

2.2 Préparation du matériel biologique

2.2.1 Élevage des pucerons

Pour l'ensemble des expérimentations, les pucerons utilisés sont issus du même clone parthénogénique d'*Acyrthosiphon pisum* (Harris) *LL01* (obtenu lors d'une infestation de luzerne à Lusignan en 1986). Le clone *LL01* a été utilisé à la place du clone séquencé {Shigenobu, 2000 #381} car il est dépourvu de symbiotes secondaires. En effet, il est peu probable qu'il existe un polymorphisme important entre ce clone et le clone séquencé. En revanche, ce choix permet d'éviter tout risque de contamination par du matériel génétique issu de symbiotes secondaires.

Les pucerons sont maintenus sur de jeunes plants de pois (*Vicia faba L.* variété *Aquadulce*) placés dans des cages en plexiglas. Ces cages sont maintenues en conditions contrôlées (température constante de 21 °C, humidité relative de 70 %, photopériode de 16 heures), de façon à obtenir uniquement des pucerons se reproduisant par parthénogenèse.

Les pucerons sont choisis à un même stade physiologique de développement. Pour cela, 1000 adultes sont isolés et sont « mis à pondre » durant 24 heures sur leur plante hôte. Les adultes sont alors retirés des plantes et les larves sont maintenues 24 heures supplémentaires sur les plantes. Elles sont ensuite transférées sur milieu artificiel. Le protocole de préparation des différents milieux est donné en annexe 3.2. Les milieux sont filtrés à travers une membrane de pores de 0,45 µm de diamètre (*Millipore* ,Billerica, MA, USA) et distribués entre deux feuilles de Parafilm M® (*Structure Probe*, West Chester, PA,

USA) simulant une feuille (cf. **Figure C.2.2**). Ces feuilles sont tendues sur des bagues en chlorure de polyvinyle (PVC, 40 mm de haut et 100 mm de diamètre). Les larves sont déposées à l'intérieur de ces bagues (100 larves par bague et 6 bagues pour chaque milieu nutritionnel) et sont placées durant 6 jours dans les mêmes conditions de photopériode, d'humidité et de température que les adultes. À l'issue de cette période, les individus ont atteint le stade L4 (dernier stade larvaire). L'ensemble des pucerons vivants obtenus sur chaque bague est pesé puis utilisé pour les expérimentations.

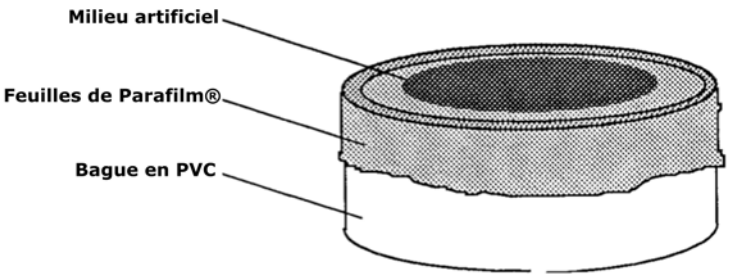


Figure C.2.2 Dispositif expérimental de préparation des milieux artificiels.

2.2.2 Purification des bactéries par filtration

Les bactéries *Buchnera* sont purifiées à partir des pucerons selon le protocole de Charles et Ishikawa (1999). Les pucerons (1 g) sont broyés dans 10 ml de tampon de lyse [KCl, 0,93 g ; $MgCl_{2,}$, 1,01 g ; saccharose, 42,8 g ; Tris, 2,11g ; qsp H_20 DEPC (eau contenant 0,1 % de diéthyl pyrocarbonate), 500 ml, pH 7,05 ; solution incubée 2 heure à température ambiante]. Le broyat est filtré sur une toile de nylon de 100 µm et successivement sur des membranes *isopore*™ de 30, 5, et 3 µm (*Millipore*). Le filtrat est ensuite centrifugé 5 minutes à 4000 g, à 4 °C. Le culot obtenu est repris dans 1 ml de tampon puis centrifugé 5 minutes à 3500 g, à température ambiante. Il est ensuite pesé et conservé à -80 °C.

2.2.3 Extraction des ARN

L'ARN total est extrait au TRIzol (*Invitrogen*, Paisley, RU) à partir des culots de *Buchnera* purifiées. Les contaminants d'ADN génomique sont éliminés par un traitement de 30 minutes à 37 °C avec la DNase *RQ1 RNAse-free* (*Promega*, Madison, WI, USA). Les ARN sont ensuite purifiés sur colonne avec le kit *Rneasy kit* (*Qiagen*, Hilden, Allemagne) puis re-suspendus dans de l'eau DEPC. Cette étape de purification des échantillons est essentielle pour éliminer les ARNases (Holloway *et al.*, 2002). La pureté et la qualité des ARN

sont finalement contrôlées par mesures au spectrophotomètre à 260 nm et 280 nm, puis par électrophorèse sur gel d'agarose en conditions dénaturantes.

2.2.4 Préparation des cibles de contrôle

Les cibles de contrôle sont produites par transcription *in vitro* à partir de deux gènes (*pelL* et *pelK*) clonés dans des vecteurs d'expression. Pour cela, le plasmide est linéarisé par des enzymes de restriction (cf. **Tableau C.2.1**) et le résultat de cette digestion est contrôlé sur gel d'agarose.

Tableau C.2.1 Présentation des vecteurs et des enzymes de restriction utilisés pour la préparation des cibles de contrôle.

Gènes	Vecteurs	Enzymes
PelK (1,6 kpb)	PBluescript II KS (-) sites Hind III/Hind III (2,9 kbp)	Xho I (Promega)
PelL (2,2kpb)	pBC KS (+) sites Pst I/Bam HI (3,4 kpb)	Hind III (Boringher)

La solution, contenant le plasmide digéré, est purifiée par l'ajout de 3 µl de protéinase K suivi d'une incubation 30 minutes à 37 °C. Le plasmide est extrait avec 2 fois 200 µl d'une solution de phénol/chloroforme/alcool isoamilyque (25/24/1). Le plasmide obtenu est précipité 2 heures à -80 °C dans 20 µl d'acétate de sodium 3 M et 400 µl d'éthanol absolu puis centrifugé à 12000 g 30 minutes à 4 °C. Le culot est finalement repris à la concentration de 1 µg/µl dans de l'eau DEPC. Le résultat est contrôlé par mesures au spectrophotomètre et migration sur gel d'agarose. La réaction de transcription *in vitro* est ensuite réalisée avec le kit *RNAMaxx™ High Yield Transcription* (*Stratagene*, Cedar Creek, TX, USA) et les transcrits obtenus sont conservés à -80 °C.

2.3 Marquage

Pour la technique de marquage indirect qui a été retenue, les séquences des amorces nécessaires à la synthèse des ADNc ont été choisies de façon à accroître la spécificité du marquage. Pour cela, le logiciel *GDP* (*Genome Directed Primers*) développé par Talaat *et al.* (2000) a été utilisé. Il a permis de définir un nombre minimal d'amorces semi-spécifiques localisées parmi les cent premières paires de bases à partir de l'extrémité 3'. Il s'agit de 22 octamères et de 16 hexamères, qui permettent la synthèse de 493 et 108 ARNm respectivement. Pour les 15 gènes restants, il a été nécessaire de déterminer 15 octamères spécifiques. Ces amorces sont toujours utilisées en association avec un ensemble d'amorces aléatoires de 9 mers fournies avec le kit de marquage.

Les échantillons d'étude sont préparés en mélangeant 15 µg d'ARN total de *Buchnera* avec les ARNm des cibles de contrôle *pelL* et *pelK,* en quantité connue et en rapport respectivement 1 : 1 et 1 : 3 pour les deux fluorochromes (cf. **Tableau C.2.2**).

Tableau C.2.2 Quantités de cibles de contrôle ajoutées pour l'ensemble des échantillons.

	Cy3	Cy5
PelL	5 ng (1)	5 ng (1)
PelK	7 ng (1)	21 ng (3)

La réaction de transcription inverse est réalisée en deux étapes avec le kit *Cy-Scribe post-labelling (Amersham,* Piscataway, NJ, USA). Les échantillons sont d'abord mélangés à 1,5 µl d'une solution d'amorces spécifiques et aléatoires (chacune à 50 nM) pour atteindre un volume total de 11 µl. Le mélange obtenu est incubé 5 minutes à 70 °C pour permettre l'hybridation des amorces sur les ARN. La réaction de transcription est ensuite réalisée en ajoutant à la solution précédente 4 µl de tampon *CyScript*^{TM} 5X, 1 µl du mélange commercial (dATP, dCTP, dGTP, et dTTP) à 25 µM, 2 µl de DTT à 10 mM, 1 µl d'aminoallyl-dUTP (aa-dUTP) à 0,2 mM et 1 µl d'enzyme *CyScribe RT* (200 U), puis en incubant le tout 3 heures à 42 °C.

Une fois l'ADNc synthétisé, l'ARN matrice est hydrolysé par traitement alcalin, en ajoutant 2 µl de NaOH 2,5 M puis en incubant 15 minutes à 37 °C. Le pH est finalement rétabli à 7,4 par ajout de 10 µl de Tris-HCl 1 M.

L'ADNc amino-allylé est purifié par précipitation durant 60 minutes à -80 °C dans 105 µl d'éthanol absolu avec 3 µl d'acétate de sodium 3M (pH 5,2). La solution est ensuite centrifugée 30 minutes à 12000 g, à température ambiante. Le culot obtenu est lavé avec 1 ml d'éthanol 75 % glacé puis centrifugé à nouveau 15 minutes à 12000 g et finalement déshydraté 10 minutes au *speed vacuum*.

Le culot d'ADNc amino-allylé est re-suspendu dans 40 µl de bicarbonate de sodium 0,1 M (*Merck*, Darmstadt, Allemagne). La solution obtenue est ajoutée aux cyanines Cy3 ou Cy5 lyophilisées *(Amersham*) et incubée 90 minutes à température ambiante, dans l'obscurité. Les cyanines qui n'ont pas été incorporées sont éliminées par passage sur colonnes AutoSeq™ G-50 (*Amersham*). Les ADNc marqués en Cy3 et en Cy5 sont finalement mélangés et déshydratés au *speed vacuum*.

2.4 Hybridation et lavages

L'hybridation automatique est réalisée sur la machine *Discovery®XT System*[28] (*Ventana*, Tucson, AZ, USA). Les lames étiquetées sont placées dans la machine (jusqu'à 20 lames) pour subir un pré-traitement automatisé de 1 heure 20 qui intègre trois cycles de lavages dans du *ChipSpread*™ (30 minutes à 42 °C), du *ChipPrep1*™ (10 minutes à 70 °C) et du *ChipPrep2*™ (30 minutes à température ambiante). Durant ce pré-traitement, les lames subissent une pré-hybridation et une fine couche d'huile est déposée à la surface de chaque lame.

Entre-temps, les cibles marquées sont re-suspendues dans 10 µl d'eau ultra-pure, dénaturées par incubation 5 minutes à 95 °C et centrifugées 5 minutes à 12000 g. Elles sont ensuite mélangées à 190 µl de tampon *ChipHybe*™. Une fois le pré-traitement achevé, les cibles sont déposées manuellement sur les lames et l'hybridation automatisée est réalisée durant 8 heures 20. La machine assure une circulation continue de l'huile déposée à la surface des lames et donc de la solution de cibles. À la fin du cycle, les lames sont automatiquement conservées dans une solution de *Ribowash*™ à température ambiante.

Le cycle final de lavage est ensuite déclenché par l'utilisateur et permet d'éliminer l'huile par un rinçage automatique dans du *ChipClean*™ (12 min, 37 °C). Les lames sont alors sorties de la machine et lavées à température ambiante successivement dans le bain n°1 [*Ribowash*™ , 2 fois 5 minutes], le bain n°2 [SSC 2X, 2 fois 5 minutes] et le bain n°3 [SSC 0,2X, 2 fois 2 minutes]. Elles sont finalement rincées rapidement dans du SSC 0,05X et de l'eau ultra-pure puis séchées par centrifugation 1 minutes à 2000 g.

2.5 Acquisition des images

Les images sont obtenues avec le scanner *GeneTAC*™ *L SIV* (*Genomic Solutions*, Huntingdon, RU). Les lames sont observées successivement avec les longueurs d'onde 635 nm (associée au fluorophore Cy5) et 532 nm (associée au fluorophore Cy3), car le Cy5 est plus sensible à la dégradation que le Cy3. Pour chaque canal, le gain du photomultiplicateur (PMT) est ajusté manuellement à une valeur telle que seuls 30 à 50 plots (environ 1 % des plots) présentent des pixels saturés. Pour affiner la différence de gain entre les deux canaux, les profils des signaux des témoins positifs sont ajustés pour que les rapports de leurs intensités soient en accord avec les rapports des quantités de cibles de contrôle ajoutées à l'échantillon (Leung et Cavalieri, 2003). Le meilleur compromis entre vitesse d'acquisition et qualité de l'image est obtenu avec le calcul des in-

[28] http://www.ventanadiscovery.com/.

tensités de chaque pixel à partir d'une moyenne de quatre mesures ,et un paramètre de résolution de 10 µm. Les images obtenues pour chaque longueur d'onde sont enregistrées dans deux images séparées au format 16-bit TIFF.

2.6 Acquisition et de filtration des données

Les données sont obtenues en analysant les images au format TIFF avec le logiciel *GenePix Pro® 6.0* (*Axon Instruments*, Union City, CA, USA). Une étape de filtration est réalisée au préalable. Son but est d'attribuer, à chaque plot, le nom du gène ou du témoin correspondant et un indice de qualité.

Pour l'attribution des noms, le fichier contenant les sondes et leurs caractéristiques (fichier de sortie de ROSO) a été utilisé en association avec le plan de dépôt pour créer un fichier intermédiaire, associant les noms des gènes et leur emplacement sur la puce. Ce second fichier a permis d'obtenir le fichier au format GAL accepté par le logiciel *Genepix,* et qui est utilisé pour ajouter une grille sur l'image, associant un nom à chaque plot.

L'attribution des indices de qualité est réalisée en plusieurs étapes. La première étape est l'élimination manuelle des plots situés sous des taches ou des rayures. Tous ces plots, dont le signal est visiblement très affecté, sont indexés « mauvais » (index de qualité de -100). Les plots non détectés par le logiciel ne nécessitent aucun traitement, car ils sont automatiquement indexés comme « non visibles » (index de qualité de -50). Enfin, il est très fréquent d'avoir sur la grille un grand nombre de plots vides, par exemple sur le haut de la lame pour la puce *Buchnera,* car les grilles définies sous *GenePix* ne peuvent contenir que des blocs de dépôt de taille identique. Il est important d'indexer l'ensemble de ces plots vides comme « absents » (indice de qualité de -75). La seconde étape est l'analyse de la distribution de la taille des plots. Pour la réaliser, il est nécessaire de représenter le diamètre en fonction du numéro de la sonde. Le diagramme obtenu permet d'éliminer les plots dont le diamètre est très éloigné de la distribution globale (diamètre trop important ou au contraire trop faible). La troisième étape est l'indexation des plots présentant un signal saturé comme « mauvais ». Le seuil de rejet qui a été retenu est un minimum de 50 % des pixels du signal saturés, pour l'un ou l'autre des fluorophores. La quatrième étape est l'analyse de la corrélation entre la moyenne et la médiane du signal de chaque plot, car une mauvaise corrélation indique une distribution atypique de la fluorescence des pixels du plot. Pour visualiser ces mauvaises corrélations, la moyenne de chaque plot est représentée en fonction de la médiane pour chacun des fluorochromes, et les points très distants de la bissectrice sont éliminés. La cinquième étape est l'étude de la variabilité du signal et du

bruit de fond. Pour cela, il est nécessaire de représenter l'erreur standard en fonction de la médiane pour le signal puis pour le bruit de fond, et cela pour chacun des deux fluorochromes. Cette représentation permet de visualiser rapidement les points les plus éloignés du nuage, en particulier lors de l'étude du bruit de fond, et de les indexer comme « mauvais ». Finalement une dernière étape d'analyse du rapport signal sur bruit de fond permet de retenir les meilleurs plots et de les indexer comme « bons » (index de qualité de 100) lorsqu'ils possèdent une valeur élevée de ce rapport. Pour cela, au moins 55 % des pixels associés au signal doivent posséder une intensité supérieure à la somme des pixels associés au bruit de fond (en pourcent) et de l'erreur standard de mesure, et cela pour les deux fluorochromes.

Pour l'acquisition finale du signal et du bruit de fond associés aux différents plots, il est possible d'utiliser le volume, la moyenne ou la médiane des pixels du plot. La moyenne et la médiane sont généralement utilisées car elles présentent moins de variabilité que le volume. Pour cette étude, la médiane a été retenue car elle permet une estimation plus robuste du signal pour des distributions anormales des pixels.

2.7 Analyse statistique

L'analyse statistique est réalisée sur les valeurs des intensités corrigées, et après application d'une transformation logarithmique en base 2, avec le logiciel R[29] (version 1.9.1) et l'ensemble d'outils *Bioconductor*[30] (version 1.4) (Gentleman *et al.*, 2004). Les bibliothèques *Bioconductor* utilisées pour instancier, visualiser et normaliser les données sont *marrayNorm* et *marrayPlots*. Les analyses de variance ont été réalisées avec la bibliothèque *maanova*[31] (version 0.97-7).

2.7.1 Étude du bruit de fond

La valeur du bruit de fond est calculée automatiquement par le logiciel *Genepix,* puis retranchée à l'intensité du signal. L'étude des intensités résultantes révèle une quantité importante de plots possédant des valeurs négatives. Ces valeurs sont inutilisables pour l'analyse, à cause de la transformation logarithmique, et génèrent donc des trous dans les tableaux de données. De nombreuses méthodes plus ou moins complexes ont donc été développées pour limiter leur

[29] http://cran.univ-lyon1.fr.
[30] http://www.bioconductor.org.
[31] http://www.jax.org/staff/churchill/labsite/software/anova/rmaanova/.

nombre (cf. partie A). Pour les études réalisées dans cette thèse, la méthode empirique d'Edwards (2003) a été retenue pour sa simplicité de mise en œuvre. Lorsque l'intensité obtenue après soustraction du bruit de fond est inférieure à une valeur δ, la transformation d'Edwards (2003) remplace la simple soustraction du bruit de fond.

Cette correction a été intégrée dans une fonction R[32] qui associe le calcul des valeurs corrigées à une analyse qualitative des plots. Cette fonction permet le calcul des intensités corrigées pour tous les plots d'une lame qui ne sont ni « absents », ni « mauvais ». Une valeur de δ égale à 1 a été retenue et un indice de qualité de « -40 » est attribué aux plots corrigés par la méthode d'Edwards (2003). Cette valeur a été choisie car elle permet de situer les plots corrigés sur une échelle de qualité entre les plots « non visibles » (-50) et les plots « corrects » (0). Un nombre important de plots subit cette correction. Il s'agit essentiellement des plots dont les intensités sont très faibles dans les deux conditions, c'est-à-dire des plots peu intéressants pour la recherche de gènes différentiels. Une analyse qualitative a donc été ajoutée pour ne conserver que les plots corrigés qui présentent un intérêt, c'est-à-dire les plots pour lesquels le signal est corrigé pour une seule des deux conditions, et dont l'intensité du signal dans l'autre condition est supérieure à une valeur définie empiriquement. Cette valeur est le quatre-vingt-quinzième quantile de la distribution des intensités des témoins négatifs qui ne sont pas indexés « mauvais », ou en d'autres termes, la plus petite valeur d'intensité parmi les 5 % de témoins négatifs présentant les intensités les plus importantes. Les signaux qui ne remplissent pas ces deux conditions sont indexés « mauvais », et ne sont donc pas retenus pour l'analyse ultérieure.

2.7.2 Normalisation des données

Contrairement à la majorité des études réalisées sur puce à ADN, la puce *Buchnera* ne contient qu'un faible nombre de gènes. L'hypothèse classiquement utilisée de stabilité de l'expression de la plupart des gènes (ou au moins de leur répartition symétrique entre les gènes sur- et les gènes sous-exprimés) est donc difficile à envisager. Quant aux témoins positifs déposés sur chaque lame, leur nombre n'est pas suffisant pour permettre leur seule utilisation pour normaliser l'ensemble des données. Pour chaque lame, une fonction développée sous R a donc été utilisée pour définir un ensemble de plots dits invariants entre les deux conditions d'étude. Cet ensemble est recherché *a posteriori* avec la méthode de somme des rangs décrite par Schadt *et al.* (2001). La

[32] Fonction dévelopée par Nicolas Morin (DEA ASMB, 2004).

fonction utilisée permet de rechercher un ensemble de plots invariants pour chaque aiguille de dépôt parmi les plots qui ne sont ni indexés comme « mauvais », ni corrigés avec la méthode d'Edwards (2003). Une fois que les plots invariants ont été identifiés, une méthode de normalisation intensité-dépendante par aiguille a été réalisée par régression quadratique locale ou *loess* (Yang *et al.*, 2002b). Pour cela la fonction *maNorm* de la librairie *marrayNorm* a été utilisée.

2.7.3 Détection des gène exprimés de façon différentielle

2.7.31 Obtention des données moyennées

Les données d'intensité obtenues sont moyennées en deux temps selon la procédure décrite dans la partie B. Des fonctions développées sous R permettent ainsi de moyenner les valeurs dont l'index qualité est le plus élevé pour les quatre plots représentant une même sonde sur la puce (moyenne intra-sondes), puis pour les deux ou trois sondes représentant un même gène (moyenne inter-sondes). Cette procédure permet d'obtenir deux valeurs d'intensité (R et G) pour chacun des 617 gènes sur les 16 lames du plan expérimental, soit un total de 32 valeurs d'intensité pour chaque gène.

2.7.32 Analyse de variance

Une analyse de variance a été réalisée sur les intensités après application d'une transformation logarithmique en base 2.

Un premier modèle de normalisation est appliqué à l'ensemble des données pour prendre en compte les effets des facteurs fixes « lame » et « fluorochrome » : $y_{ij} = \mu_0 + A_i + D_j + AD_{ij} + \varepsilon_{ij}$ où :

- y_{ij} représente le logarithme des intensités d'expression,
- μ_0 est l'effet global,
- A_i est l'effet de la $i^{ième}$ lame (16 lames),
- D_j est le $j^{ième}$ niveau de l'effet fluorophore D (2 couleurs),
- ε_{ij} représente les erreurs stochastiques.

Les résidus r_{ijgr} obtenus avec ce modèle (c'est-à-dire pour chaque gène les différences obtenues entre valeur observée et valeur prédite par le modèle) sont utilisés pour réaliser 617 analyses de variance gène spécifique. Pour chaque gène *g*, ces modèles à effets fixes sont de la forme :

$r_{ijkl} = \mu + A_i + D_j + aa_k + saccharose_l + (aa : saccharose)_{kl} + \varepsilon_{ijkl}$ où :

- μ traduit l'effet global du gène *g* étudié,
- A_i est l'effet de la $i^{ième}$ lame (16 lames),
- D_j est le $j^{ième}$ niveau de l'effet fluorophore D (2 couleurs),
- aa_k représente l'effet « taux d'acides aminés essentiels » (2 niveaux),

- $saccharose_l$ représente l'effet « concentration en saccharose » (2 niveaux),
- $(aa : saccharose)_{kl}$ est l'interaction entre les facteurs « taux d'acides aminés essentiels » et « concentration en saccharose » (4 niveaux),
- ε_{ijlm} représente les erreurs stochastiques.

Ce premier modèle, dit complet, permet de tester l'effet du terme d'interaction. Un second modèle gène-spécifique est ensuite ajusté sans terme d'interaction : $r_{ijkl} = \mu + A_i + D_j + aa_k + saccharose_l + \varepsilon_{ijkl}$. Il est utilisé pour tester les effets des facteurs *aa* et *saccharose*.

2.7.33 Test de F

Le test utilisé est le test *Fs* (test de Fischer basé sur une estimation, dite réduite, de la variance entre gènes, qui est déterminée selon la méthode de *James-Stein* (Cui *et al.*, 2004). Il s'agit d'une amélioration du test F_2, qui utilise comme estimation de la variance la moyenne entre variance globale et variance individuelle des gènes (cf. partie A). Dans le test *Fs*, une fonction de pondération permet d'accorder une importance accrue aux contributions individuelles des gènes présentant des variances très hétérogènes. Ce test implémenté dans la bibliothèque *maanova* permet de déterminer les gènes pour lesquels le terme d'interaction (*aa :saccharose*) et les deux facteurs (*aa* et *saccharose*) ont un effet, c'est-à-dire les gènes exprimés de façon différentielle. Il a été couplé pour certaines analyses aux résultats obtenus avec le test F_3, qui utilise une variance globale calculée à partir de l'ensemble des variance des gènes.

Ces deux tests diffèrent du test classique, appelé aussi F_1, leur distribution nulle n'est donc pas disponible dans des abaques. Une approche non paramétrique de re-échantillonnage est donc utilisée pour établir ces distributions nulles (absence d'effet du facteur étudié). Pour cela, 1000 permutations des résidus sont effectuées sans remise pour tester l'interaction puis pour tester chacun des deux facteurs. Elles permettent d'obtenir trois listes de gènes différentiellement exprimés classés dans l'ordre des probabilités du test (chacune d'entre elles correspond à la probabilité que la valeur du test statistique soit supérieure ou égale à la valeur observée lorsque l'hypothèse nulle est vraie).

2.7.34 Classification K-means

La méthode de classification dite des *K-means* a été utilisée pour séparer en deux groupes les gènes présentant un terme d'interaction significatif. Pour cela, une technique de *bootstrap* a été utilisée (Kerr et Churchill, 2001a) et cent itérations ont été réalisées à partir des matrices d'expressions relatives. Chaque groupe dit consensus a ensuite été défini à partir des résultats obtenus lorsque les gènes étudiés ont été retrouvés pour au moins 60 % des itérations dans l'un des groupes.

3

3 Résultats

« Rien ne vaut la recherche lorsqu'on veut trouver quelque chose. »
Tolkien[33]

3.1 Étude physiologique de l'impact de la composition des milieux nutritionnels sur les pucerons

L'impact des différents milieux nutritionnels sur les performances des pucerons a fait l'objet d'une étude préliminaire réalisée à York par Angela Douglas. Pour cela trente larves âgées de deux jours ont été placées individuellement sur chaque milieu pour une durée de cinq jours. Pour les quatre milieux nutritionnels (cf. **Tableau C.3.1**), trois concentrations totales en acides aminés différentes (100, 150 et 200 mM) ont été testées sur trois lots de dix pucerons. Les larves survivantes ont ensuite été comptées et pesées individuellement pour calculer leur taux de croissance relatif ou GR (*Growth Rate*) de la façon suivante :

$$GR = \frac{G}{TA}$$

G représente le gain de poids des larves durant la période de prise alimentaire (T en jours) et A leur poids moyen.

Tableau C.3.1 Récapitulatifs des résultats obtenus sur trente larves élevées durant cinq jours sur les quatre milieux nutritionnels de l'expérience (deux concentrations en saccharose et deux taux d'acides aminés essentiels).

Milieu	Saccharose (en M)	Taux d'acides aminés essentiels (en %)	Survivants	GR moyen (erreur type)
1	0,5	50	21	0,21 (0,068)
2	0,5	25	24	0,23 (0,033)
3	1,0	50	12	0,13 (0,031)
4	1,0	25	15	0,14 (0,035)

[33] Tolkien, J. *Bilbo le Hobbit*. Livre de Poche, Paris (1989).

L'analyse de variance qui a été réalisée sur les taux de croissance relatifs obtenus montrent une diminution du taux de croissance des pucerons lorsque la concentration en saccharose du milieu nutritionnel augmente. En revanche, la variation du taux d'acides aminés essentiels ne semble pas avoir d'influence significative sur leur croissance.

Pour l'obtention du matériel biologique destiné aux expérimentations sur puces à ADN, six lots de 600 pucerons ont été élevés, durant six semaines successives, sur les quatre milieux artificiels pour une concentration totale en acides aminés de 150 mM. Dans cette expérience la mortalité observée sur les quatre milieux a été comparable. La même observation a été réalisée dans l'expérience préliminaire pour les lots de dix pucerons élevés sur un milieu de concentration totale en acides aminés de 150 nM (avec 7, 4, 5 et 6 survivants pour les milieux numérotés de 1 à 4). La mortalité a donc été négligée et chaque semaine, le « poids moyen » des pucerons vivants a été calculé en divisant le poids total de pucerons obtenus par le nombre de pucerons déposés initialement sur chaque milieu. Une analyse de variance a de nouveau été réalisée sur ces mesures de façon à tester l'effet des facteurs « taux d'acides aminés essentiels » (facteur *aa* à 2 niveaux) et « concentration en saccharose » (facteur *saccharose* à 2 niveaux) sur le poids moyen des pucerons. L'interaction entre les deux facteurs n'étant pas significative (p=0,5061), l'ajustement d'un modèle d'ANOVA sans terme d'interaction permet de montrer que les deux facteurs ont un effet significatif sur le poids moyen des pucerons, avec des probabilités respectivement de 0,0009 pour le facteur *aa* et de 0,036 pour le facteur *saccharose*. Par ailleurs, une étude de la répartition des poids moyens sur les différents milieux (cf. **Figure C.3.1**) semble indiquer, tout comme l'expérience préliminaire, une croissance ralentie des pucerons en présence d'une forte concentration en saccharose dans le milieu, et ce, quel que soit le taux d'acides aminés essentiels. Une concentration en saccharose de 1 M est donc suffisante pour induire un stress osmotique sur les pucerons. Enfin, le poids des pucerons semble être plus élevé pour une teneur en acides aminés essentiels réduite dans le milieu. Ce résultat, bien que surprenant, ne semble pourtant pas contradictoire avec l'étude préliminaire des taux de croissance relatifs des pucerons sur des milieux contenant respectivement 25 et 50 % d'acides aminés essentiels, bien que dans l'expérience préliminaire l'effet ne soit pas significatif.

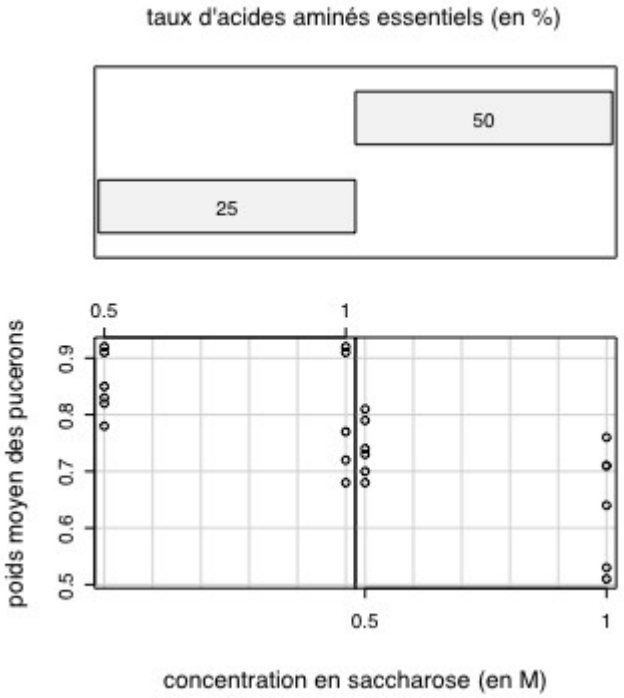


Figure C.3.1 Représentation du poids moyen des pucerons en fonction du taux d'acides aminés essentiels et de la concentration en saccharose dans le milieu nutritionnel.

3.2 Acquisition et normalisation des données d'expression

Les images TIFF acquises pour chaque lame de l'expérience (cf. **Figure C.3.2**) sont utilisées pour obtenir les seize tableaux de données de l'expérience avec le logiciel *Genepix*. Pour chaque lame, ces données sont normalisées puis moyennées selon la procédure décrite dans la partie B. Le tableau final des intensités de fluorescence transformées en logarithme base 2 est finalement instancié sous R pour réaliser les différentes analyses.

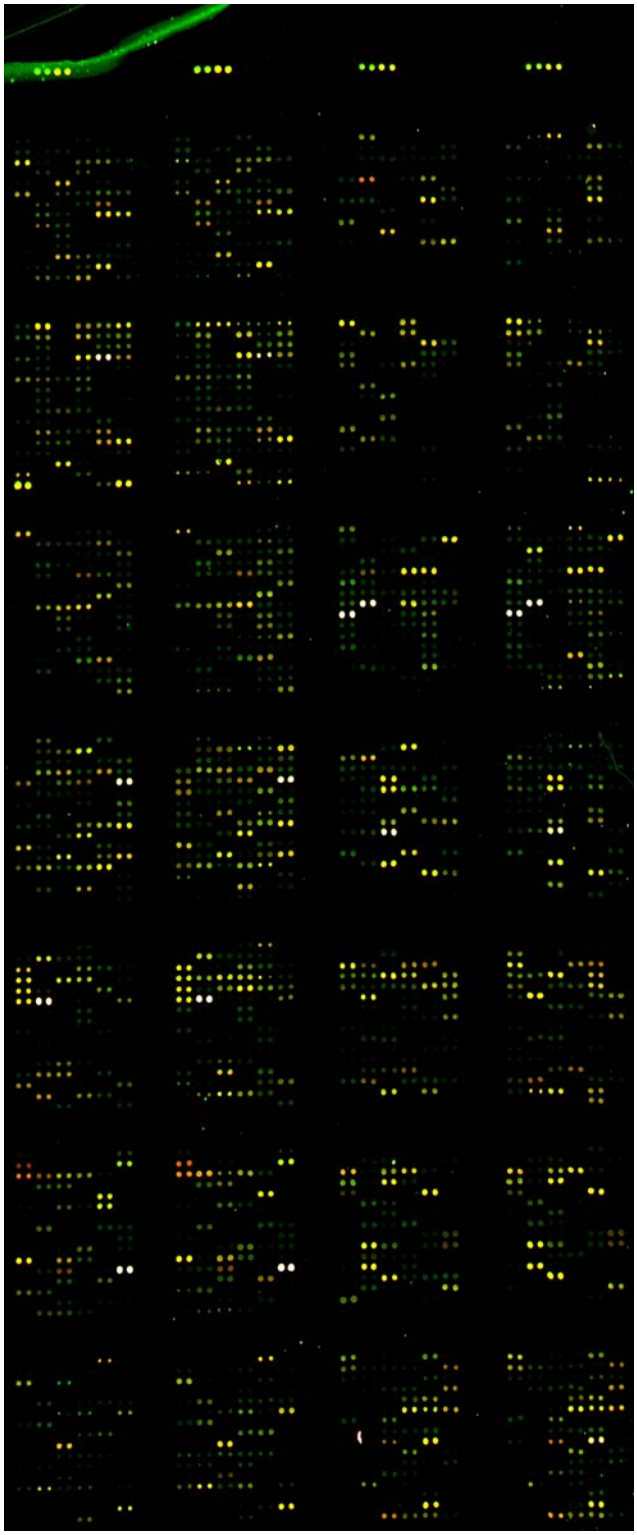


Figure C.3.2 Exemple de résultat pour une hybridation effectuée sur la puce *Buchnera* (lame 16) et après superposition des images obtenues pour chacun des deux fluorochormes avec le logiciel *Genepix*.

Pour l'étape de normalisation, la comparaison des résultats obtenus avec une seule régression pour l'ensemble des plots, et avec une régression pour chacune des quatre aiguilles de dépôt, a permis de retenir une normalisation aguille dépendante qui offre la possibilité d'éliminer les biais techniques liés au dépôt. Cette normalisation est appelée *printTip loess* dans *Bioconductor*. L'expérience montre qu'un minimum de cent plots invariants par aiguille de dépôt est nécessaire pour une normalisation correcte des données (cf. **Figure C.3.3**).

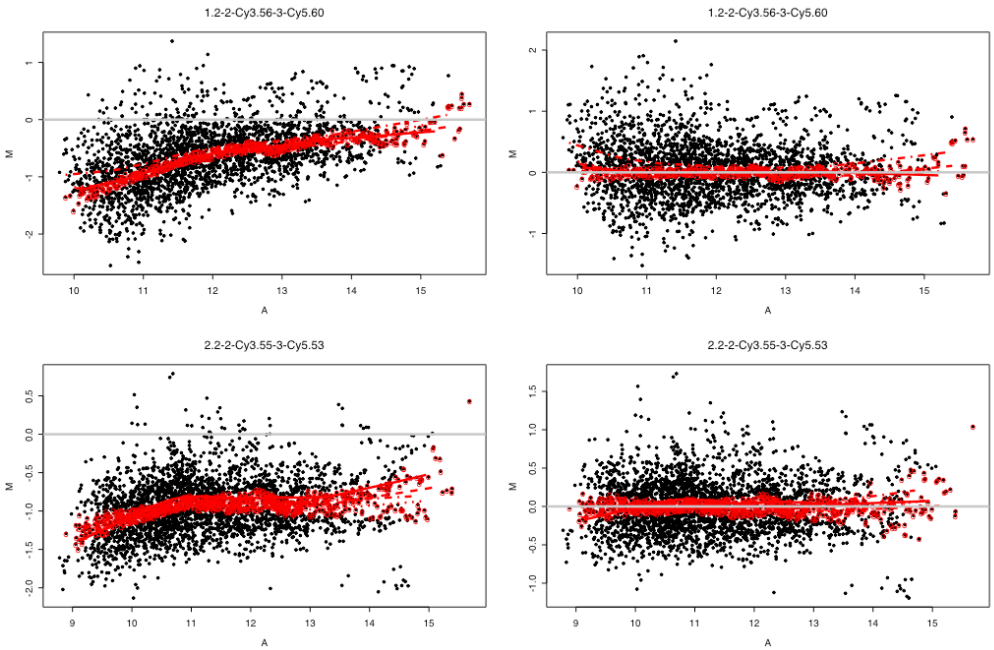


Figure C.3.3 Représentation MA pour la lame 2 (en haut) et la lame 10 (en bas) avant et après normalisation avec $M=\log_2(R/G)$ et $A=\log_2(RG)^{0,5}$. Les points en rouge représentent les plots invariants qui sont utilisés pour la normalisation.

La visualisation des intensités obtenues pour les différentes sondes de contrôle a permis de valider la qualité des données et la méthode de normalisation utilisée (cf. **Figure C.3.4**). Les intensités d'expression des plots contenant du tampon sont réparties aléatoirement dans le nuage des plots de très faibles intensités, ce qui montre une absence de structuration du bruit de fond sur les lames. Les plots représentant les témoins positifs *pelL* et *pelK,* dont les cibles ont été déposées respectivement en rapport 1 : 1 et 1 : 3, montrent que ces rapports sont très conservés après normalisation. Pour certaines lames, il existe cependant un décalage d'échelle (les points représentant *pelL* ne sont plus centrés sur 0), car la normalisation est réalisée sur un ensemble de gènes invariants défini *a posteriori,* et non pas sur ces témoins positifs, en raison de leur nombre insuffisant. Pour limiter cet inconvénient, il serait possible d'utiliser une mé-

thode de normalisation mixte incluant à la fois les témoins positifs et la recherche d'un ensemble de gènes invariants, d'autant plus que les futures puces contiendront d'avantage de témoins positifs. Compte tenu du nombre limité de témoins positifs disponibles actuellement, ceci n'a cependant pas été réalisé pour cette expérience.

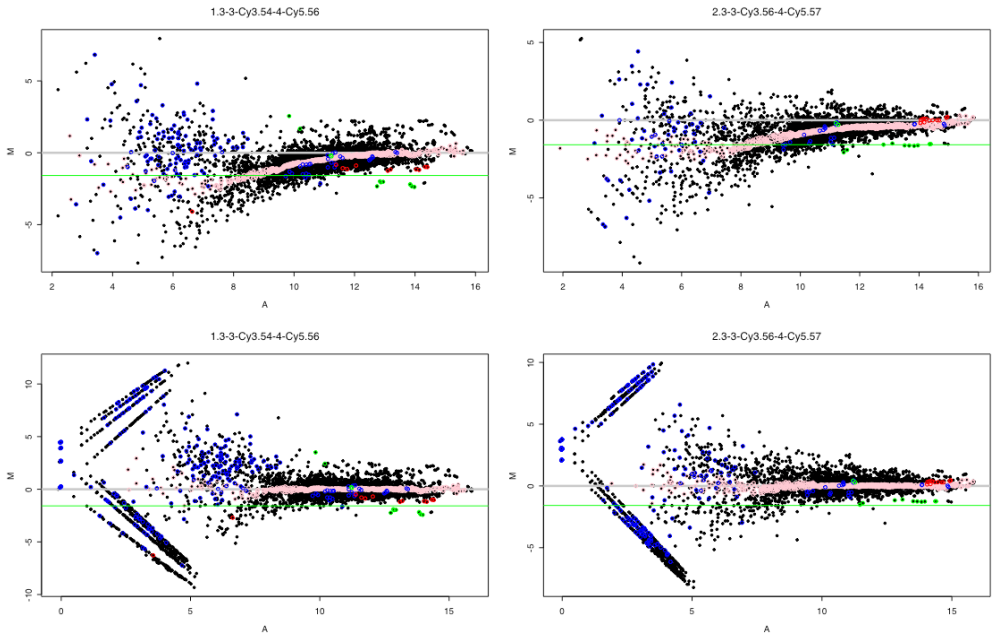


Figure C.3.4 Représentation MA pour la lame 3 (à gauche) et la lame 11 (à droite) avant et après normalisation. Les points en rose représentent les réponses des plots invariants utilisés pour la normalisation et les points bleus les réponses des plots contenant du tampon. Les points en rouge représentent les intensités des cibles de contrôle *pelL* introduites en rapport 1:1 (ce rapport est représenté par la ligne grise M=0) et les points en vert celles des cibles *pelK* introduites en rapport 1:3 (ce rapport est représenté par la ligne horizontale verte). Les structures particulières situées à gauche des graphes après normalisation représentent les plots corrigés par la méthode d'Edwards (2003). Pour cette expérience, ils seront tous éliminés au cours de l'étape de moyennage qualité dépendante.

Une fois la normalisation effectuée, le moyennage des intensités en fonction des indices de qualité attribués lors de l'étape de filtration est réalisé (cf. partie B). Les intensités obtenues pour les répétitions d'une même sonde sur la lame sont d'abord moyennées (moyennes intra-sondes), puis la même démarche est appliquée aux différentes sondes représentant un même gène (moyennes inter-sondes). Cette méthode originale de moyennage permet d'obtenir en moyenne 92,7 % de « bons » signaux (contre 61,5 % pour les données initiales) et de réduire le nombre de « mauvais signaux » de 7,8 à 1,8 % (cf. **Tableau C.3.2** et **Tableau C.3.3**).

Tableau C.3.2 Répartition en pourcentage des données en fonction des indices qualité (-100 : mauvais, -50 :non détecté, 0 : correct, 100 : bon) après filtration avec le logiciel *Genepix* selon la démarche présentée dans la partie « matériel et méthodes ».

Lames		Mauvais (-100)		Non détectés (-50)		Corrects (0)		Bons (100)	
1	9	6	7	27	43	1	6	66	44
2	10	10	7	34	32	5	2	51	60
3	11	5	6	28	37	3	2	64	55
4	12	9	10	28	31	5	3	58	56
5	13	5	5	25	34	10	2	61	59
6	14	9	8	22	34	3	1	66	57
7	15	6	7	21	38	1	2	72	53
8	16	20	6	36	29	8	1	36	65
Moyenne		7,8 %		30,7 %		3,2 %		58,3 %	

Tableau C.3.3 Répartition des données pour les 617 gènes en fonction des indices qualité après moyennage intra et inter-sondes.

Lames		Mauvais (-100)		Non détectés (-50)		Corrects (0)		Bons (100)	
1	9	7	8	21	79	0	43	589	487
2	10	11	2	43	49	30	8	533	558
3	11	7	16	24	56	7	3	579	542
4	12	17	5	17	41	27	10	556	561
5	13	17	3	6	41	32	6	562	567
6	14	2	10	10	48	5	4	600	555
7	15	7	5	5	57	3	4	602	551
8	16	66	2	31	32	58	6	462	577
Total		189 (1,8 %)		576 (5,5 %)		247 (2,3 %)		9477 (90,4 %)	

Avant l'étape de moyennage des différentes sondes représentant un même gène, il a été nécessaire de vérifier que leurs réponses étaient relativement similaires. Ainsi, les différentes visualisations des réponses des sondes en fonction de leur position sur le gène ne montrent pas de biais particulier, contrairement à ce que certains auteurs observent parfois dans le cas d'un marquage réalisé avec des amorces spécifiques (Wang et Seed, 2003). Cette absence de corrélation est sans doute liée à l'utilisation d'amorces aléatoires en complément des amorces semi-spécifiques dans les expériences réalisées pour *Buchnera*.

En revanche, une surexpression systématique des ARN de transfert est observée dans l'échantillon marqué en Cy5, et ce, quelle que soit la condition d'étude. Cet effet n'était pas visible sur les lames QUANTIFOIL® hybridées de façon manuelle. L'hypothèse a donc été avancée que la structure particulière des ARN de transfert interagit de façon préférentielle avec l'huile utilisée pour les hybridations automatiques, huile qui produit une fluorescence de même lon-

gueur d'onde que celle du Cy5. La mesure de l'expression des gènes codant pour les ARN de transfert semble donc biaisée dans les expériences. Néanmoins une étude des rapports d'intensités des 32 gènes des ARN de transfert pour lesquels des marquages réciproques de deux conditions ont été réalisés (le premier flip-flop concerne les lames 5/8 et 6/7 et le second les lames 13/16 et 14/15) montre une différence significative des valeurs de M (cf. **Figure C.3.5**). Il semble donc possible de détecter une expression différentielle pour ces gènes malgré le biais technique, et ce d'autant plus que l'effet fluorochrome spécifique de chaque gène est intégré dans le modèle d'ANOVA (cf. 2.7.32).

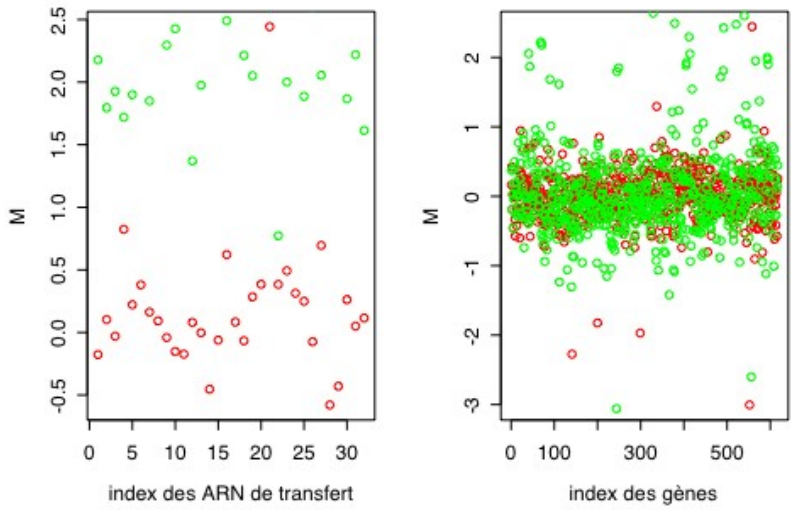


Figure C.3.5 Représentation des rapports d'intensités pour la lame 16 (en rouge) et la lame 13 (en vert) pour les ARN de transfert (à gauche) et pour l'ensemble des gènes (à droite).

3.3 Analyse statistique

Un premier modèle d'analyse de variance, contenant le terme d'interaction des deux facteurs *aa* et *saccharose,* a été utilisé pour réaliser un test de Fischer sur le terme d'interaction. Il a permis d'obtenir la liste de l'ensemble des gènes classés dans l'ordre de significativité des probabilités d'expression différentielle du test *Fs*. Dans cette liste, 66 des gènes présentent une probabilité inférieure au seuil 0,05 (sans correction liée aux tests multiples). De la même façon, une nouvelle analyse de variance sans le terme d'interaction a permis de tester l'influence des deux facteurs séparément et d'obtenir deux listes de gènes classés dans l'ordre de la statistique. Pour le facteur *aa*, seul 11 gènes possèdent une probabilité inférieure au seuil de 0,05. En revanche pour le facteur *saccharose*, 90 gènes possèdent une probabilité signi-

ficative. Les résultats de ces tests peuvent être visualisés sur des graphes de type « Volcano » représentant pour chacun des facteurs étudiés les probabilités du test de F associées à chaque gène (après transformation logarithmique en base 10) en fonction des rapports d'expression estimés entre les deux conditions du facteur étudié (cf. **Figure C.3.6**). L'étude de ces graphes suggère de compléter les listes obtenues par le test *Fs* avec celles qui sont obtenues par le test F_3, plus sensible à la valeur absolue des rapports M. Ces listes contiennent 92 gènes présentant un terme d'interaction significatif, 25 pour le facteur *aa* et 104 pour le facteur *saccharose* (cf. annexe 4).

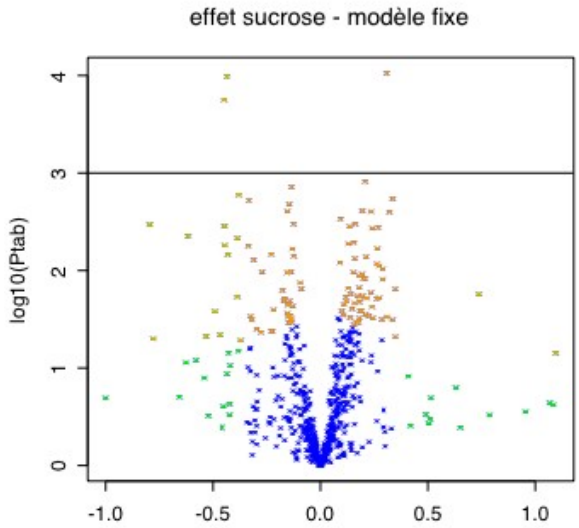


Figure C.3.6 Graphe « Volcano » représentant les gènes différentiellement exprimés pour l'effet *saccharose*. Les points en bleu représentent les gènes pour lesquels aucun effet n'est observé, les points en orange les gènes détectés par le test *Fs* et les points en vert sont les gènes détectés par le test F_3.

Pour chaque gène, des effets sont estimés lors de l'ajustement des modèles d'analyse de variance pour chacun des niveaux des facteurs *aa* et *saccharose*. Les données brutes étant transformées en logarithme base 2, la différence de ces valeurs a été utilisée pour déterminer les valeurs estimées de M correspondant aux rapports d'expression de chaque gène pour les deux niveaux des facteurs *aa* (rapports des taux d'acides aminés essentiels 50 % : 25 %) et *saccharose* (rapports des concentrations en saccharose 0,5 M : 1 M). La possibilité de calculer ces M estimés montre tout l'intérêt d'utiliser un modèle d'analyse de variance pour l'étude des données d'expression car ces valeurs de M ne peuvent jamais être observées réellement sur une même lame.

De la même façon, pour les gènes ayant une interaction significative, les rapports des deux conditions du facteur *aa* et du facteur *saccharose* ont été

calculés respectivement pour chacune des deux concentrations en saccharose et chacun des deux taux d'acides aminés essentiels.

Les valeurs d'expression relative obtenues lors de l'ajustement des modèles d'analyse de variance peuvent également être utilisées pour observer les profils d'expression des gènes dont l'interaction est significative. Ces valeurs ont été utilisées pour réaliser un classement des gènes présentant des profils similaires par la méthode dite des *K-means*. Cette analyse a permis de distinguer deux classes de gènes en fonction de leur profil d'expression, pour les quatre types d'interaction existant, correspondant aux quatre milieux nutritionnels (cf. **Figure C.3.7**). Le premier groupe correspond aux milieux nutritionnels 1 (acides aminés essentiels 50 % et saccharose 0,5 M) et 4 (acides aminés essentiels 25 % et saccharose 1 M) et le second groupe aux milieux 2 et 3.

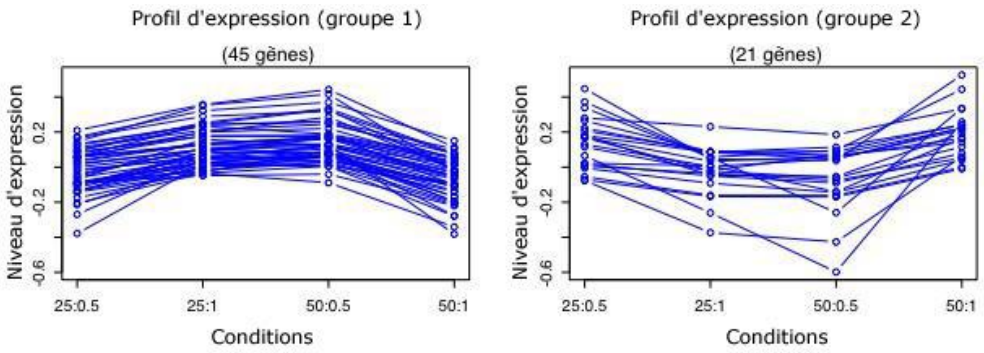


Figure C.3.7 Représentations des gènes en groupes de profils d'expression pour la liste de gènes présentant une interaction significative.

Les trois listes de gènes exprimés de façon différentielle associent donc M estimés et probabilités. Les probabilités obtenues résument la significativité des tests statistiques d'expression différentielle en combinant la variation de l'expression des gènes et les erreurs standard de mesure. Cependant, les tests qui ont été appliqués sont très conservatifs, en raison de l'absence de procédure de contrôle de risque. Les listes obtenues comptent donc un grand nombre de faux positifs. De très faibles mesures d'expression peuvent en effet être très significatives si les erreurs standard sont faibles par le simple fait du hasard (Golfier *et al.*, 2004). Les tests F_3 et *Fs* qui ont été utilisés sont construits pour limiter ces effets. Néanmoins, pour intégrer une notion de significativité biologique dans les résultats, seuls les gènes présentant une variation de leur taux d'expression supérieur à 1,2 ont été conservés pour les interprétations.

Parmi les gènes exprimés de façon différentielle pour le facteur *aa*, 5 gènes codant pour des ARN de transfert et un gène de synthèse des ARN

transfert sont surexprimés lorsque le taux d'acides aminés essentiels diminue de 50 à 25 %. Le gène *rnpA* codant pour la composante protéique de la ribonucléase P, enzyme responsable de la maturation des ARN de transfert, est également surexprimé. Il semble qu'il existe donc une production accrue d'ARN de transfert chez la bactérie lorsque le puceron subit un stress nutritionnel En revanche, les gènes intervenant dans la traduction (les gènes codant pour certaines protéines ribosomales et pour le facteur d'élongation *efp*) sont réprimés. Pour le facteur *saccharose,* 15 gènes codant pour des ARN de transfert sont surexprimés pour la concentration en saccharose la plus élevée (1 M). Des gènes impliqués dans la réplication de l'ADN (*dnaA* et *dnaG*) et dans la division cellulaire (*ftsL* et *ftsI*) sont également réprimés, indiquant vraisemblablement que le stress osmotique subit par le puceron est responsable d'une division cellulaire moins importante chez *Buchnera*.

3.4 Interprétation biologique

3.4.1 Relation entre niveaux d'expression et organisation du génome

Différentes analyses ont été réalisées essentiellement pour valider la cohérence des données d'expression obtenues. Ces études, visant à relier niveaux d'expression et organisation du génome, ont permis, d'une part, de valider en partie les résultats obtenus et d'autre part de tester certaines hypothèses concernant la régulation de l'expression chez *Buchnera*. Pour cela, la mesure absolue du niveau d'expression des gènes est obtenue en utilisant le terme μ du modèle d'analyse de variance (cf. 2.7.32). Ces niveaux d'expression sont très fortement corrélés avec les niveaux d'expression moyens des gènes calculés sur l'ensemble des intensités de fluorescence transformées en logarithme base 2 (32 mesures par gènes), mais présentent l'avantage d'être corrigés des effets lames et fluorochromes (cf. **Figure C.3.8**).

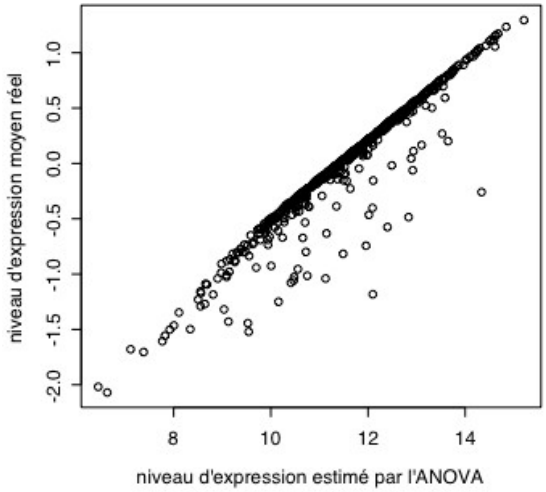


Figure C.3.8 Représentation des niveau d'expression moyens calculés à partir des mesures d'intensités réelles en fonction des niveaux d'expression estimés par l'analyse de variance.

L'étude des niveaux moyens d'expression montre pour différents opérons connus (*mopA*/*mopB*, *arg*, *his*) un niveau d'expression similaire au sein d'un même opéron (données non présentées). Par ailleurs, des études de classification, réalisées sur les gènes présentant un terme d'interaction significatif (cf. 3.3) et sur les gènes présentant un effet *saccharose* montrent une cohérence de profils d'expression pour différents petits groupes de gènes contigus sur le chromosome (données non présentées). Ces groupes pourraient représenter des unités de transcription et indiquer une régulation potentielle au niveau transcriptionnel. Ils pourraient donc être utilisés pour la recherche d'éléments de régulation, en amont de la position sur le chromosome des groupes de gènes présentant des profils d'expression similaires. Cette étude préliminaire nécessite naturellement de nombreux approfondissements qui n'ont pas été réalisés dans le cadre de cette thèse. Cette analyse présente néanmoins l'intérêt de mettre en évidence une cohérence des niveaux d'expression au sein de certains opérons connus.

Les valeurs moyennes d'expression ont ensuite été utilisées pour étudier le niveau d'expression des gènes, en fonction de leur localisation sur le chromosome par rapport à l'origine de réplication. Pour cela, le site de terminaison de la réplication a été choisi à la moitié du chromosome (en amont du gène indexés BU293). Le test de chi2 réalisé sur quatre classes de niveaux d'expression des gènes (découpage en fonction des quantiles de la distribution) et quatre classes de distances à l'origine (découpage du chromosome en quatre

parties) présente une probabilité fortement significative (p=0,0035), indiquant que les gènes les plus fortement exprimés sont préférentiellement localisés à proximité de l'origine de réplication (cf. **Tableau C.3.4**).

Tableau C.3.4 Répartition des effectifs pour la réalisation du test de chi2 comparant niveaux d'expression et distances à l'origine de réplication des gènes chez *Buchnera*.

Distance à l'origine de réplication	Niveau d'expression			
	Élevé	Moyen	Faible	Très faible
Proche	25	19	19	21
Moyen	44	19	19	15
Éloigné	16	16	19	23
Très éloigné	11	12	25	20

Une étude a également été réalisée pour comparer le niveau d'expression des gènes en fonction de leurs positions sur le brin direct ou indirect du chromosome. L'étude de la distribution des niveaux d'expression sur chacun des deux brins semble indiquer un niveau d'expression légèrement plus élevé sur le brin direct que sur le brin indirect (cf **Figure C.3.9**). Un test de chi2 réalisé pour les quatre classes de niveaux d'expression montre qu'il existe une différence significative d'expression entre les deux brins (p=0,0065).

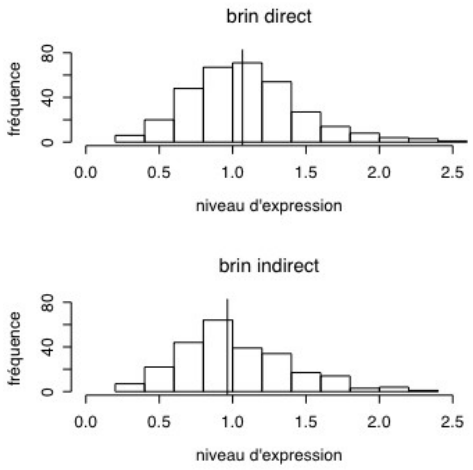


Figure C.3.9 Représentation des fréquences des niveaux d'expression (sans transformation logarithmique) sur le brin direct (en haut) et indirect (en bas).

Parallèlement à l'étude de la localisation, l'étude de la composition des séquences en fonction des niveaux d'expression a été réalisée. Une première comparaison a été effectuée entre le taux de GC des gènes et leur niveau d'expression (cf. **Figure C.3.10**). Un test de chi2, réalisé avec deux classes de gènes et deux classes de taux de GC (déterminées en fonction des médianes des répartitions), indique que les gènes les plus fortement exprimés sont également

les plus riches en GC ($p=2,326*10^{-11}$). Ce résultat est très cohérent car les gènes les plus riches en bases G et C sont également les gènes qui sont le moins soumis au biais mutationnel vers les bases A et T, donc les gènes qui sont les plus conservés.

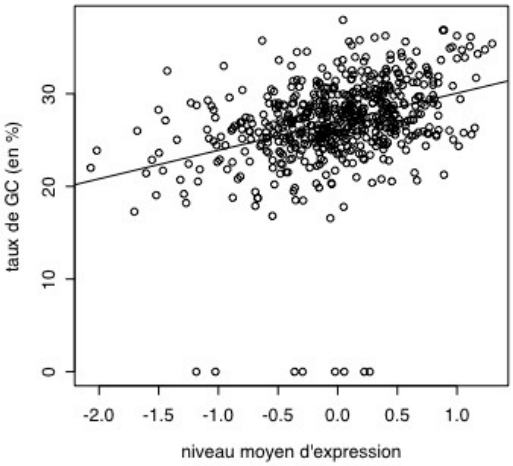


Figure C.3.10 Représentation du taux de GC des gènes chez *Buchnera* en fonction de leur niveau moyen d'expression.

Pour réaliser une étude plus précise de l'usage du code en fonction des niveaux d'expression, le *CAI* (*Codon Adaptation Index*) de chaque gène codant pour une protéine chez *Buchnera* a été calculé. Cet index de biais d'usage des codons synonymes est représentatif de l'impact de la sélection sur l'évolution d'une séquence (Sharp et Li, 1987). Chez *Escherichia coli*, il existe une forte corrélation entre *CAI* et niveau d'expression (Gouy et Gautier, 1982). Elle reflète une sélection des codons optimaux au sein des gènes les plus fortement exprimés, pour assurer une transcription et une traduction rapide, et l'absence d'une telle sélection sur les gènes faiblement exprimés, pour lesquels les codons moins optimaux sont conservés (Shields, 1990). Pour obtenir les *CAI*, une table de référence d'usage relatif de codons synonymes (*RSCU*) a tout d'abord été construite à partir d'un sous-ensemble de cinquante gènes fortement exprimés (issus des données expérimentales). Pour cela, la fréquence observée de chaque codon d'un gène de ce sous-ensemble a simplement été divisée par la fréquence attendue de ce codon sous l'hypothèse d'un usage identique des codons synonymes pour un acide aminé donné. Cette première étape a été réalisée avec le logiciel *EMBOSS CUSP*[34]. Le *CAI* a ensuite été calculé pour chaque

[34] http://bioweb.pasteur.fr/seqanal/interfaces/cusp.html.

gène codant pour des protéines de l'organisme comme la moyenne géométrique des valeurs de *RSCU* de chaque codon, divisée par le *CAI* maximum qui peut être calculé pour un gène de même composition en acides aminés. Cette seconde étape a été effectuée avec le logiciel *EMBOSS CAI*[35]. Un test de corrélation de Spearman, réalisé sur les rangs des niveaux d'expression et des valeurs des *CAI* obtenus pour les 572 gènes codant pour des protéines chez *Buchnera,* montre finalement une absence de corrélation significative entre les deux valeurs. Il n'existe donc pas de sélection de codons optimaux chez *Buchnera*.

Parmi les cinquante gènes les plus fortement exprimés, quinze gènes codent pour des protéines ribosomales, et cinq pour des protéines de choc thermique (*mopA*, *mopB*, mais aussi *dnaK*, *hscA* et *htpX*), qui sont toutes des protéines connues pour être fortement exprimées chez *Buchnera* (Wernegreen et Moran, 1999). Les gènes homologues chez *Escherichia coli* sont également connus pour être surexprimés. Pour étudier les niveaux d'expression de l'ensemble des gènes des deux bactéries, une comparaison a été réalisée entre les *CAI* des gènes d'*Escherichia coli* homologues des gènes de *Buchnera* (d'après Rispe *et al.*, 2004) et le niveau d'expression des gènes de *Buchnera*. Ces *CAI* ont été calculés avec le logiciel*CODONW*[36] à partir d'un ensemble de 27 gènes codant pour des protéines ribosomales qui sont fortement surexprimés chez *Escherichia coli* (Sharp et Li, 1986). L'étude du graphe correspondant révèle qu'une multiplication par 1,5 de la valeur des *CAI* calculé chez *Escherichia coli* se traduit par une multiplication du niveau moyen d'expression par 7 chez *Buchnera* (cf. **Figure C.3.11**). Un test de corrélation de Spearman réalisé sur les rangs révèle une corrélation forte entre le *CAI* des gènes d'*Escherichia coli* et le niveau d'expression des gènes homologues chez *Buchnera* ($p=9,12*10^{-7}$).

[35] http://bioweb.pasteur.fr/seqanal/interfaces/*CAI*.html.
[36] http://www.molbiol.ox.ac.uk/cu/.

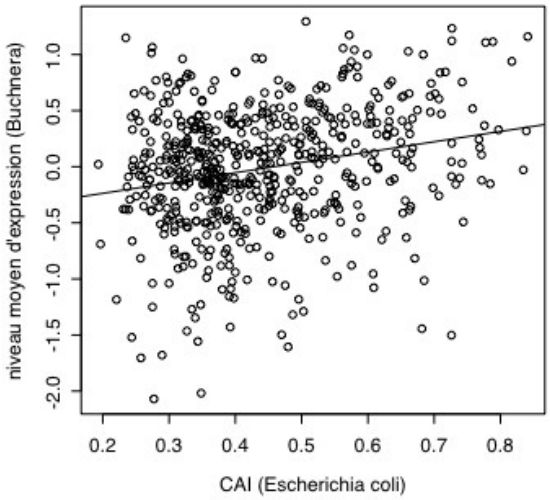


Figure C.3.11 Représentation du niveau moyen d'expression des gènes de *Buchnera* en fonction du *CAI* des gènes homologues chez *Escherichia* coli.

Une analyse des résidus les plus grands d'un modèle d'analyse de variance, testant l'influence de la valeur du *CAI* des gènes d'*Escherichia coli* sur le niveaux d'expression des gènes homologues de *Buchnera,* a ensuite permis de mettre en évidence des gènes fortement exprimés chez *Buchnera* qui ne le sont pas chez *Escherichia coli* et réciproquement. Cette étude révèle une expression forte chez *Buchnera* de plusieurs gènes du métabolisme des acides aminés essentiels, avec notamment *ilvI* impliqué dans le métabolisme de l'isoleucine ou encore *aroH* et *aroK* impliqués dans la biosynthèse des acides aminés aromatiques. Cette différence s'explique par le mode de vie symbiotique de la bactérie qui fournit des acides aminés à son hôte. Par ailleurs, huit gènes codant pour le flagelle se trouvent fortement exprimés chez *Buchnera*. Ce résultat s'avère d'autant plus surprenant que *Buchnera* ne possède pas de flagelle. Les gènes normalement impliqués dans la synthèse du flagelle sont donc sans doute impliqués chez *Buchnera* dans une autre fonction, probablement une fonction symbiotique importante comme le transport des acides aminés vers l'hôte. Aucune étude expérimentale n'a été réalisée chez *Buchnera*, mais il a été montré chez la bactérie pathogène *Yersinia enterocolitica,* que le flagelle assure une fonction d'export des protéines (Young *et al.*, 1999). Les auteurs de cette étude concluent d'ailleurs que ce système d'export pourrait être un mécanisme général de transport des protéines jouant un rôle dans les interactions entre hôte et bactéries. Enfin, les chaperonnes *mopA* et *mopB* (correspondant respectivement à *GroEL* et *GroES* chez *Escherichia coli)* et les chaperonnes *hsca* et *htpx*, qui sont relativement bien exprimées chez *Escherichia coli*, se démarquent d'avantage chez *Buchnera.* Certains auteurs ont avancé l'hypothèse que la su-

rexpression importante des chaperonnes pouvait permettre à *Buchnera* de lutter contre le biais mutationnel vers les bases A et T auquel est soumis son génome. Cependant, des études réalisées sur les mitochondries de levure laissent envisager un autre rôle à ces chaperonnes. En effet, dans les mitochondries *Hsp60* forme avec *Hsp10* un hétérodimère qui intervient dans la renaturation des protéines, et principalement pour les protéines régulatrices provenant de la cellule qui ont été dénaturées pour traverser les membranes mitochondriales (Dubaquie *et al.*, 1998). De la même façon, ces chaperonnes pourraient favoriser le repliement de protéines régulatrices synthétisées par l'hôte. Réciproquement, les gènes faiblement exprimés chez *Buchnera* et fortement exprimés chez *Escherichia coli,* sont des gènes d'adaptation à des variations environnementales (*ksgA*) et des gènes impliqués dans les phénomènes de recombinaison (*recB*), qui sont devenus très peu utiles à *Buchnera* en raison de son mode de vie intracellulaire.

L'ensemble des analyses présentées dans ce paragraphe a été initié essentiellement pour retrouver des éléments de cohérence dans les données obtenues plutôt que pour tester de véritables hypothèses sur la régulation de *Buchnera*. Cette étude des données d'expression au niveau génomique sera poursuivie de façon plus complète au laboratoire. Elle a toutefois permis de montrer que les observations réalisées sur d'autres génomes bactériens, en ce qui concerne niveaux d'expression et localisation des gènes, sont également valables chez *Buchnera*. Elle confirme par ailleurs des études théoriques réalisées sur le génome de la bactérie, qui proposent un positionnement préférentiel des gènes les plus exprimés sur le brin direct et une résistance de ces derniers au biais mutationnel vers les bases A et T (Rispe *et al.*, 2004). L'évolution particulière de la composition du génome, en raison du mode de vie symbiotique de la bactérie (cf. partie A), semble de plus avoir affecté fortement le biais d'usage de code qui est observé chez d'autres bactéries. Elle révèle enfin une modification du profil d'expression des gènes de la bactérie (par rapport au profil d'expression des gènes d'une bactérie libre comme *Escherichia coli*) en adaptation à son mode de vie symbiotique et intracellulaire.

3.4.2 Relation entre niveaux d'expression et fonction des gènes

L'ensemble des gènes classés dans l'ordre des probabilités obtenues pour les tests de *Fs* réalisés pour les effets *aa* et *saccharose* ont été associés aux différentes catégories fonctionnelles proposées par Riley (1993). Le test de *Fs* a été retenu car il ne nécessite aucune hypothèse concernant la distribution des variances entre gènes, contrairement au test F_3 qui impose des variances homo-

gènes pour l'ensemble des gènes. Cette dernière hypothèse se révèle en effet plus difficilement acceptable en raison de la taille du jeu de données.

Un premier classement contenant 6 catégories a d'abord été utilisé, puis un classement plus fin a été défini avec 17 catégories. Les classes contenant un faible nombre de gènes (entre un et trois) ont été rassemblées dans la catégorie « autres » (cf. **Tableau C.3.5**). Ce regroupement des gènes en différentes catégories a été utilisé pour tenter de mettre en évidence une catégorie fonctionnelle dans laquelle la plupart des gènes sont exprimés de façon différentielle pour l'un des deux facteurs. Pour cela, la fréquence des différentes catégories fonctionnelles a été représentée en fonction de la classification des gènes (d'après les probabilités du test *Fs*). Le graphe a ensuite été obtenu à partir du calcul des fréquences de représentation des catégories, dans des intervalles de dix gènes dans la classification. Si la réponse de *Buchnera* à la variation de l'un des facteurs environnementaux est responsable de l'activation ou de la répression d'un processus biologique particulier, alors les gènes associés à ce processus devraient être exprimés en majorité de façon différentielle et apparaître dans les premiers rangs du classement. L'observation des courbes de fréquence des différentes catégories, en fonction des classifications obtenues pour ce facteur environnemental, devrait alors montrer une pente importante pour les premiers rangs de la classification et apparaître au-dessus des courbes moyennes (Mercier *et al.*, 2004).

Tableau C.3.5 Présentation des catégories fonctionnelles utilisées pour le classement des gènes chez *Buchnera* (d'après Riley, 1993).

Classes	Catégories fonctionnelles
1	Biosynthèse des acides aminés
2	ARN de transfert
3	Biosynthèse de cofacteurs, groupements prosthétiques et transporteurs
4	Fonctions de régulation
5	Division cellulaire
6	Enveloppe cellulaire
7	Intermédiaires du métabolisme central
8	Chaperonnes
9	Dégradation de macromolécules
10	Métabolisme énergétique
11	Biosynthèse des acides gras
12	Sécrétion de peptides et de protéines
13	Synthèse de purine, pyrimidine, nucléosides et de nucléotides
14	Synthèse et modification de macromolécules
15	Transport et protéines de liaison
16	Divers
17	Autres (gènes non classés et protéines hypothétiques)

Sur les deux types de graphiques obtenus (6 catégories : **Figure C.3.12** ou 17 catégories : **Figure C.3.13**), la catégorie des ARN de transfert compte le plus grand nombre de gènes exprimés de façon différentielle pour les deux facteurs. Pour le facteur *aa*, la catégorie des « fonctions de régulation » compte également plusieurs gènes dans les premiers rangs de la classification, et notamment le gène *rpoD* codant pour un facteur σ de l'ARN polymérase qui est surexprimé pour un taux élevé d'acides aminés essentiels (50 % par rapport à 25 %). Cette sur-représentation des fonctions de régulation parmi les premiers rangs de la classification semble indiquer, chez *Buchnera,* une adaptation de la transcription à une variation du taux d'acides aminés essentiels présents dans le milieu nutritionnel du puceron.

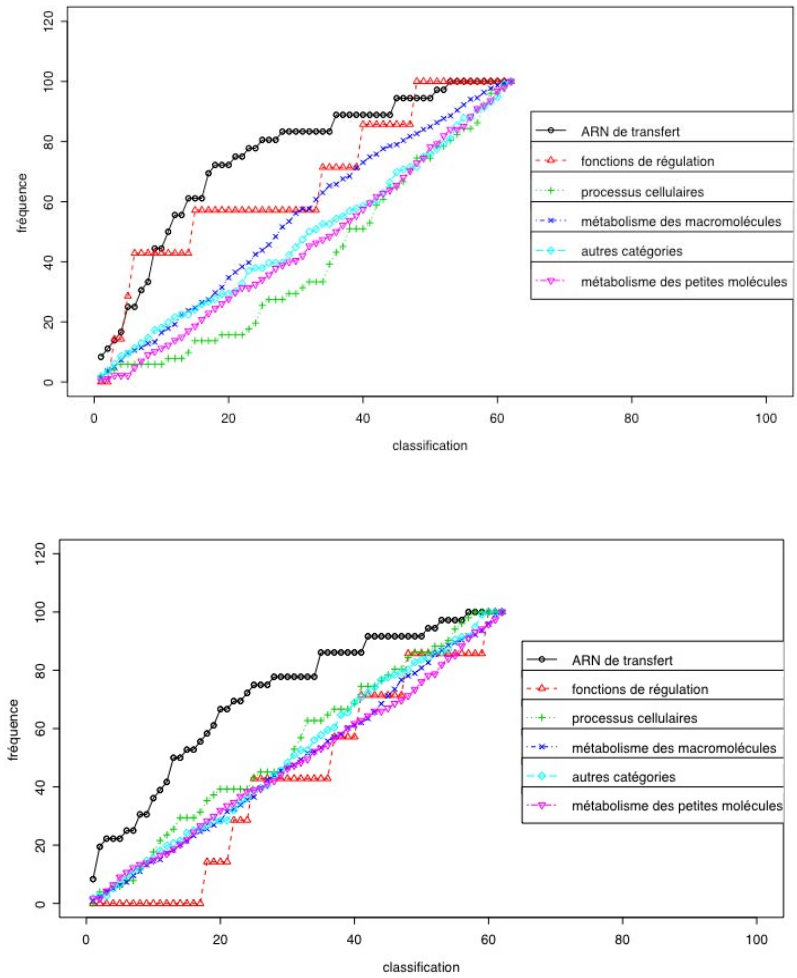
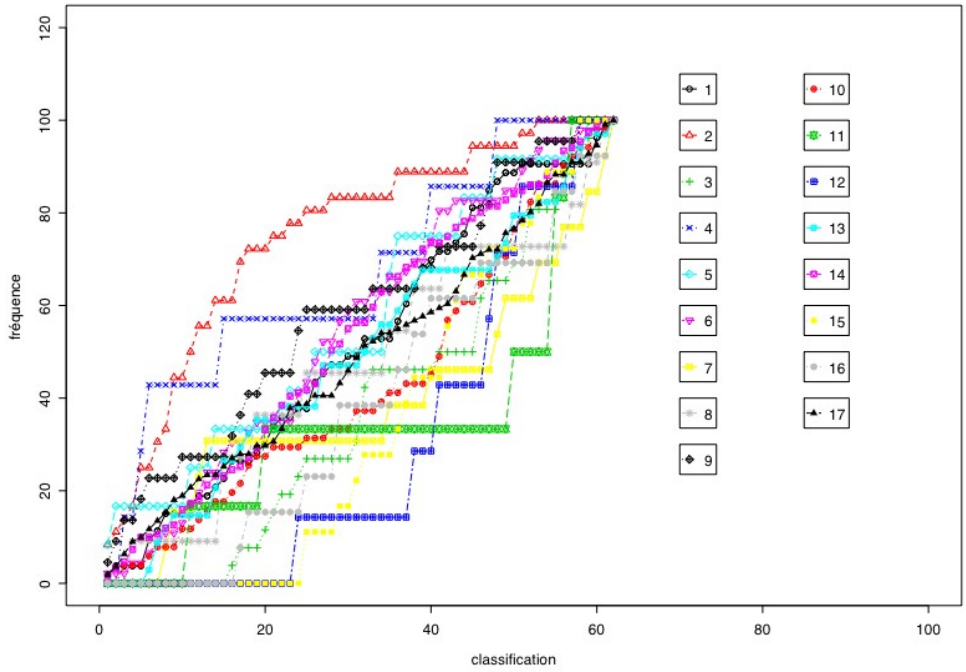


Figure C.3.12 Représentation des fréquences d'apparition de 6 grandes catégories fonctionnelles au sein de la classification des gènes obtenus après analyse de variance (probabilités du test *Fs*) pour l'effet *aa* (en haut) et l'effet *saccharose* (en bas).



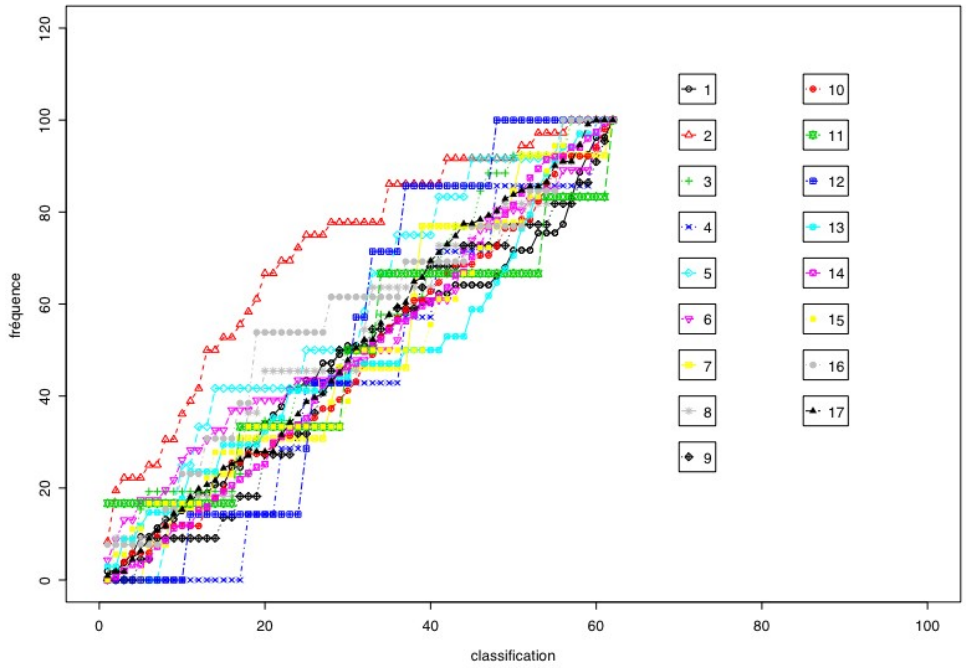


Figure C.3.13 Représentation des fréquences d'apparition de 17 catégories fonctionnelles au sein de la classification des gènes obtenus après analyse de variance (probabilités du test *Fs*) pour l'effet *aa* (en haut) et l'effet *saccharose* (en bas).

Pour déterminer une ou des catégories fonctionnelles de gènes qui sont activées ou réprimées, les gènes désignés comme exprimés de façon différentielle par les tests *Fs* et F_3 pour chacun des deux facteurs et pour le terme d'interaction ont été classés au sein des différentes catégories fonctionnelles en groupe d'activation et de répression, à partir des valeurs de M estimées (cf. **Figure C.3.14**). L'étude de la répartition des effectifs révèle que certaines catégories présentent un nombre significatif de gènes surexprimés ou sous-exprimés (loi binomiale).

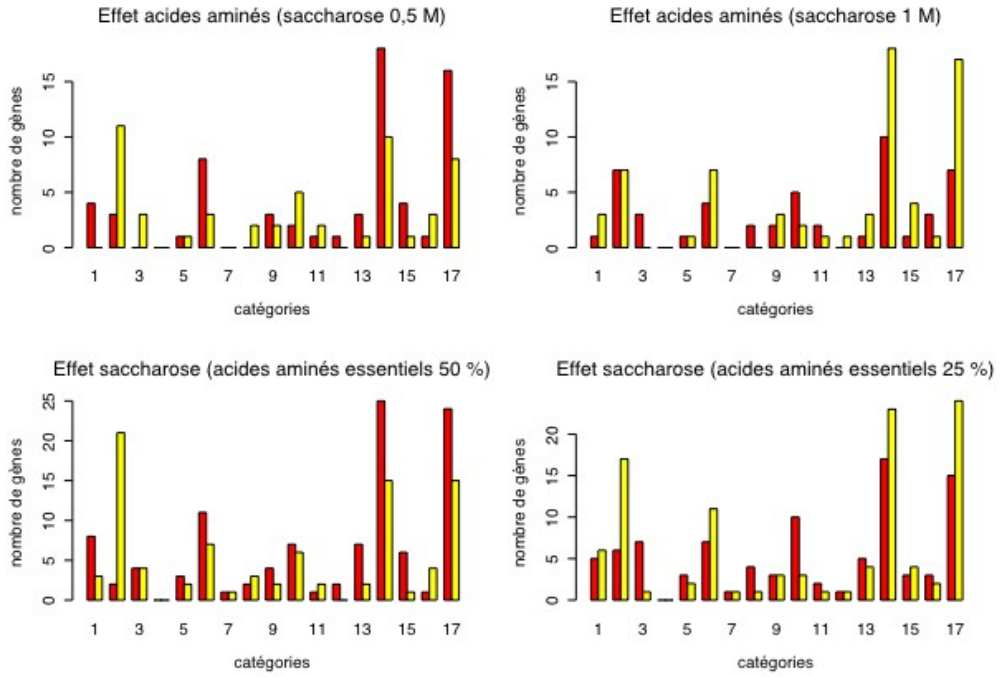


Figure C.3.14 Répartition des gènes sur-exprimés (en rouge) et sous-exprimés (en jaune) pour les différentes conditions d'étude (à partir des résultats des tests *Fs* et F_3). Pour l'effet « acides aminés essentiels » les rapports utilisés sont 50 % : 25 % et pour l'effet « concentration en saccharose » les rapports sont 0,5 M : 1 M.

L'étude des rapports d'expression entre les deux taux d'acides aminés essentiels (50 % : 25 %) en condition normale de saccharose (0,5 M) révèle, pour un taux élevé d'acides aminés essentiels, une surexpression des gènes impliqués dans la biosynthèse des macromolécules (14) et des gènes codant pour la production de protéines hypothétiques (17). Il est à noter que la catégorie de gènes appelée « biosynthèse des macromolécules » regroupe essentiellement des gènes impliqués dans la synthèse des protéines intervenant dans les mécanismes de transcription et de traduction. Les gènes codant pour certains ARN de transfert (classe 2) sont en revanche sous-exprimés.

L'étude des rapports des concentrations en saccharose (0,5 M : 1 M) pour un taux faible d'acides aminés essentiels (25 %) montre également une surexpression des gènes des voies de synthèse des macromolécules et des protéines hypothétiques en condition de stress osmotique (1 M). Par ailleurs, les gènes intervenant dans la synthèse de cofacteurs et dans le métabolisme énergétique sont sous-exprimés en condition de stress osmotique (1 M). Enfin, il existe une forte surexpression des gènes codant pour les ARN de transfert en condition de stress osmotique, et ce indépendamment du taux d'acides aminés essentiels dans le milieu.

Il semble donc qu'il existe chez *Buchnera* une surexpression des gènes codant les protéines hypothétiques et des gènes impliqués dans la biosynthèse des macromolécules, lorsque les pucerons sont élevés sur les milieux nutritionnels 1 (50 % d'acides aminés essentiels et 0,5 M de saccharose) et 4 (25 % d'acides aminés essentiels et 1 M de saccharose), en comparaison respectivement des milieux 3 (25 % d'acides aminés essentiels et 0,5 M de saccharose) et 2 (50 % d'acides aminés essentiels et 1 M de saccharose).

3.4.3 Relation entre niveaux d'expression et métabolisme

L'étude des niveaux d'expression des gènes au sein des catégories fonctionnelles ne permet pas de relier les modifications d'expression au métabolisme de la bactérie. Pour interpréter les résultats au niveau métabolique, les niveaux d'expression qualitatifs des gènes exprimés de façon différentielle ont donc été visualisés sur les cartes métaboliques de référence disponibles dans la base de données KEGG[37]. Bien que ces cartes représentent des voies métaboliques générales et théoriques, elles ont permis d'étudier l'impact des modifications d'expression au sein de différents réseaux.

Le stress osmotique subit par le puceron est responsable d'un ralentissement de sa croissance qui se traduit par une demande métabolique moins importante. Une première étude a permis d'observer les effets de ce stress osmotique infligé au puceron sur le métabolisme de la bactérie, indépendamment du taux d'acides aminés essentiels du milieu. L'analyse de l'effet principal du facteur *saccharose* montre que le stress osmotique subi par le puceron se traduit par une surexpression des gènes *murG* et *murD*, impliqués dans la voie de biosynthèse du peptidoglycane, et ce malgré une sous-expression importante des gènes impliqués dans la division cellulaire (*ftsI* et *ftsL*) et dans la réplication de l'ADN (*dnaA* et *dnaG*). Par ailleurs, le gène codant pour le transporteur de

[37] http://www.genome.jp/kegg/.

mannitol extracellulaire (*mtlA*) est sous-exprimé, indiquant une diminution de l'import de mannitol dans la bactérie. De la même façon, le gène *ompA* codant une porine impliquée dans les phénomènes de diffusion et le gène *ptsH*, codant une protéine du système de transport phosphoenolpyruvate-carbohydrate phosphotransférase (PTS), sont également sous-exprimés. En ce qui concerne le métabolisme des acides aminés essentiels, il existe une répression, en condition de stress osmotique, des trois gènes (*thrA*, *dapA* et *lysA*), impliqués la voie de biosynthèse de la lysine.

Pour étudier les effets d'une variation du taux d'acides aminés essentiels essentiels dans le milieu nutritionnel du puceron, une seconde étude de l'effet principal du facteur *saccharose* a été réalisée, en intégrant les gènes présentant un profil d'expression différent selon le taux d'acides aminés essentiels, c'est-à-dire les gènes présentant une interaction significative entre les effets *aa* et *saccharose*. L'étude du métabolisme des acides aminés essentiels montre que les gènes *aroB* et *aroE,* impliqués dans la voies de biosynthèse des acides aminés aromatiques, sont surexprimés en condition de stress osmotique (cf. **Figure C.3.15**). En revanche, le gène *trpB* impliqué dans la biosynthèse du tryptophane est sous-exprimé, et ce indépendamment du taux d'acides aminés essentiels du milieu. Quant aux gènes *aroA* et *trpC,* ils sont surexprimés pour un taux d'acides aminés essentiels de 25 % et sous-exprimés pour un taux de 50 %. Le même comportement est observé pour le gène *argD*, impliqué dans la biosynthèse de l'arginine. Pour ces trois gènes, il existe donc une activation de l'expression dans les bactéries issues de pucerons élevés sur les milieux 1 et 4 (en comparaison respectivement des milieux 3 et 2). Il s'agit des deux couples qui ont été rassemblés lors de l'analyse des profils d'expression (cf. **Figure C.3.7**).

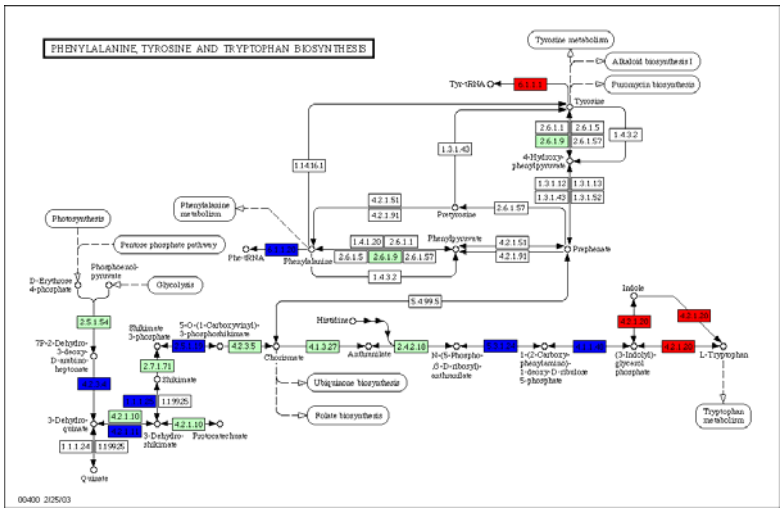


Figure C.3.15 Représentation théorique de la voie de biosynthèse de la phénylalanine, de la tyrosine et du tryptophane (carte métabolique disponible dans KEGG). Les enzymes en vert correspondent aux gènes présents chez *Buchnera* dont l'expression ne varie pas. Les enzymes en bleu correspondent aux gènes qui sont surexprimés en condition de stress osmotique, quel que soit le taux d'acides aminés essentiels ou pour 25 % d'acides aminés essentiels. Les enzymes en rouge correspondent en revanche aux gènes sous-exprimés.

L'étude du métabolisme des sucres révèle, pour la voie des pentoses phosphates une sous-expression de trois gènes (*gnd*, *pfkA* et *zwf*) en condition de stress osmotique, quel que soit le taux d'acides aminés essentiels. Les deux gènes impliqués dans la synthèse de l'érythrose phosphate (*pgi* et *tktB*) sont également sous-exprimés en condition de stress osmotique lorsque le taux d'acides aminés essentiels est de 25 % (cf. **Figure C.3.16**). En revanche, ces gènes sont sur-exprimés pour un taux de 50 %. En ce qui concerne la glycolyse, les gènes impliqués dans la synthèse du phosphoénol pyruvate (PEP) sont activés lorsque la partie de la voie des pentoses qui aboutit à la production d'érythrose phosphate est réprimée, et réciproquement. Il existe donc une sous-expression des gènes de biosynthèse de l'érythrose phosphate, qui est associée à une activation des gènes de biosynthèse du PEP pour les conditions correspondant aux milieux 1 et 4 (en comparaison respectivement des milieux 3 et 2).

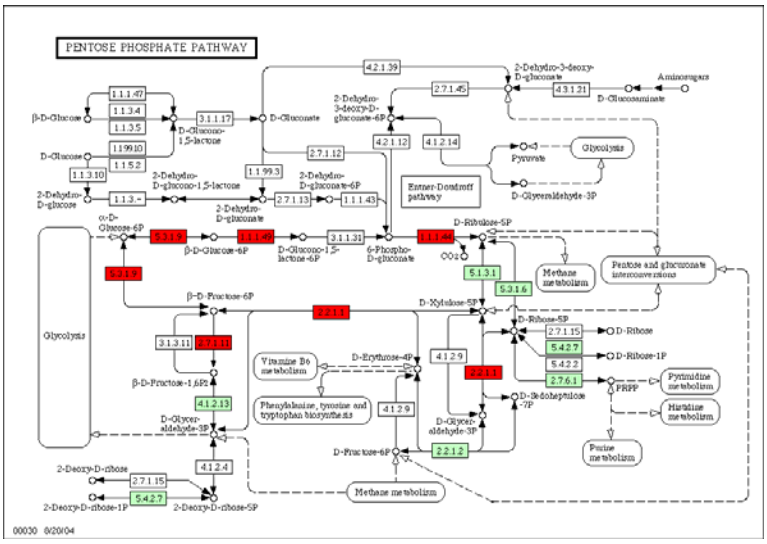


Figure C.3.16 Représentation théorique de la voie des pentoses phosphates (carte métabolique disponible dans KEGG). Les enzymes en vert correspondent aux gènes présents chez *Buchnera* dont l'expression ne varie pas et les enzymes en rouge correspondent aux gènes qui sont sous-exprimés en condition de stress osmotique, quel que soit le taux d'acides aminés essentiels ou pour 25 % d'acides aminés essentiels. Aucun des gènes ne présente de surexpression.

En ce qui concerne les protéines potentiellement impliquées dans des phénomènes de transport, les variations d'expression des trois gènes impliqués dans le système de sécrétion de protéines de type II ou système *sec* (*ftsY*, *ffh* et *rpls* codant pour la protéine ribosomale *L19*) (Agarraberes et Dice, 2001) sont toutes activées pour les conditions correspondant aux milieux 1 et 4 (en comparaison respectivement des milieux 3 et 2). Le même résultat est observé pour les gènes codant les protéines hypothétiques *ycfV* et *ynfM* et certains gènes impliqués dans la biosynthèse de protéines du complexe flagellaire (cf. **Figure C.3.17**), qui sont toutes des protéines potentiellement impliquées dans des phénomènes de transport

Enfin, il est à noter que le gène (*spE*) codant pour la transformation de la putrescine, fournie par l'hôte à *Buchnera* (Nakabachi et Ishikawa, 2000), en spermidine, est sous-exprimé pour les conditions correspondant aux milieux 1 et 4 (en comparaison respectivement des milieux 3 et 2).

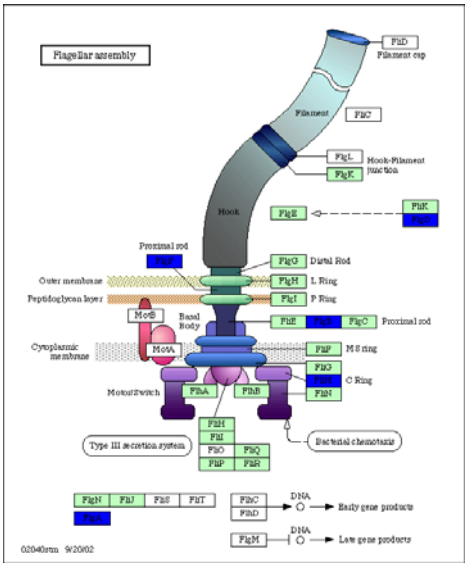


Figure C.3.17 Représentation théorique de la formation du flagelle (carte métabolique disponible dans KEGG). Les protéines en vert correspondent aux gènes présents chez *Buchnera* dont l'expression ne varie pas. Les protéines en bleu correspondent aux gènes qui sont surexprimés en condition de stress osmotique, quel que soit le taux d'acides aminés essentiels ou pour 25 % d'acides aminés essentiels. Aucun des gènes ne présente de sous-expression.

4

4 Discussion

« C'est une erreur de croire nécessairement faux ce qu'on ne comprend pas. »
Gandhi[38]

4.1 Aspects statistiques

Le plan d'expérience utilisé et les analyses statistiques réalisées dans les expérimentations sur puces à ADN jouent un rôle essentiel pour l'analyse des résultats obtenus.

Dans la stratégie expérimentale qui a été mise en œuvre, il est essentiel de prendre en compte les limitations liées au matériel biologique utilisé et celles qui sont associées à la conception de la puce.

Les pucerons utilisés sont tous des individus femelles vivipares à reproduction parthénogénétique. La population de bactéries qui leur est associée est donc très hétérogène et inclut à la fois des *Buchnera* issues de bactériocytes maternels et des *Buchnera* issues de bactériocytes embryonnaires. Cette variabilité ne peut être contrôlée que par une dissection des bactériocytes maternels. Une telle manipulation ne peut pas être mise en œuvre à l'heure actuelle, compte tenu de la quantité de matériel biologique nécessaire pour les expérimentations sur puces à ADN. En revanche, une attention particulière a été accordée au stade de développement de l'hôte. En effet, tous les pucerons utilisés sont des larves de stade quatre (L4) issues d'élevages synchronisés et elles ont été prélevées de façon aléatoire dans la population disponible.

En ce qui concerne la puce, les contraintes liées au stockage des oligonucléotides ne permettent pas une répartition aléatoire des plots sur la puce. Les sondes ont été déposées dans l'ordre de leur numéro d'accession (BU), c'est-à-dire dans l'ordre d'apparition des gènes qu'elles représentent sur le chromosome. Il existe donc une confusion entre un effet d'hybridation dit géographique (dans le cas par exemple de disparités d'hybridation dans différentes zones

[38] Gandhi. *Lettres à l'Ashram*. Albin Michel, Paris (1990).

de la lame ou encore de problèmes survenant lors des lavages), et les effets d'intérêt (par exemple l'effet de la localisation des gènes sur le chromosome). Cet inconvénient est toutefois atténué par la présence de plots répétés pour une même sonde, au sein de deux blocs de dépôt différents (dépôt par deux aiguilles différentes), et l'existence de plusieurs sondes pour un même gène (deux ou trois sondes par gène) qui ne sont jamais disposées côte à côte. La puce qui a été réalisée contient donc deux niveaux de répétition. Le premier niveau correspond à quatre dépôts d'une même sonde, ce qui permet d'augmenter la précision des résultats, par le calcul d'une moyenne pour chaque sonde. Cependant, ces plots étant disposés sous forme de paires localisées dans deux blocs adjacents, ils restent soumis aux mêmes effets éventuels d'hybridation et de lavages. Le second niveau de répétition est la représentation de chaque gène par deux ou trois sondes. Ce niveau inclut donc une variabilité encore plus importante que le précédent. Une étude graphique révèle cependant qu'il n'existe aucun biais systématique entre les valeurs des intensités de fluorescence observées et la position des sondes sur les gènes. Les intensités obtenues sur ces sondes ont donc également été moyennées pour chaque gène, ce qui permet de conserver une valeur d'intensité unique pour chaque gène. Cette méthode présente l'intérêt d'obtenir des données pertinentes et d'accroître ainsi la qualité des interprétations biologiques qui en découlent (Kerr *et al.*, 2002). Cette stratégie permet de plus de conserver un tableau de données complet, condition indispensable à la réalisation d'une analyse de variance avec le logiciel MAANOVA. Pour la réalisation d'un nombre plus important de lames, il aurait été possible de coupler cette stratégie à l'utilisation d'un algorithme de re-estimations des quelques données (1,8 %) qui présentent une qualité moins importante (Bo *et al.*, 2004).

Une fois les données obtenues, l'une des étapes les plus cruciales de l'analyse est l'étape de normalisation. Avec le modèle d'analyse de variance qui est utilisé, la normalisation n'est en théorie pas nécessaire car les différentes sources de variations techniques (lame et fluorophores) sont représentées dans le modèle de la première étape d'analyse (cf. 2.7.2). Cependant, les effets calculés dans ce modèle sont linéaires, ce qui sous-entend que les sources de variations sont elles-mêmes linéaires. Pour s'affranchir de cette hypothèse de linéarité, les données ont donc été normalisées au préalable par une méthode non linéaire (*loess* intensité dépendante).

Ces deux étapes de normalisation ne permettent cependant pas la calibration des intensités de fluorescence qui est nécessaire pour permettre la comparaison des résultats avec ceux qui ont été obtenus au cours d'autres expériences. Cet aspect reste un des problèmes majeurs posés par la technologie des puces à ADN. Il est essentiellement lié à la nature relative de la mesure de fluo-

rescence réalisée et à l'absence de référence universelle dans les plans expérimentaux. Pour une bactérie comme *Buchnera*, il serait possible d'utiliser son ADN génomique comme référence (Talaat *et al.*, 2002). En effet, lors d'une hybridation avec de l'ADN génomique le nombre de copies de chaque gène est connu et pour les gènes représentés par un même nombre de copies, toutes les sondes répondent en théorie de la même façon. Si des différences sont observées dans la réponse de certaines sondes, il serait alors possible de corréler les niveaux d'expression aux propriétés des sondes (taux de GC, Tm…) et de déterminer des termes correctifs. Ce type d'étude est en cours au laboratoire afin de déterminer des termes correctifs qui pourront être utilisés pour définir une méthode de normalisation générale des données.

Une fois les données normalisées, il est nécessaire de définir une stratégie d'analyse. Contrairement à la plupart des expérimentations sur puces à ADN, qui sont basées sur l'analyse par classification d'un grand nombre de conditions, l'étude présentée ici est une approche factorielle utilisant une analyse de variance. Cette approche offre l'avantage d'estimer, pour chacun des gènes, les effets moyens des différents facteurs qui sont des effets qui ne sont jamais réellement mesurés sur les lames. En revanche, elle nécessite des hypothèses fortes d'indépendance des données et de distribution, et notamment une distribution normale des erreurs (Cui et Churchill, 2003). Pour s'affranchir de l'hypothèse de normalité des données, la distribution du test de F a été réalisée par permutations des résidus. Cette méthode impose toutefois une variance constante et indépendante entre les mesures. Dans l'idéal, la réalisation de permutations sur les échantillons plutôt que sur les résidus serait sans doute plus pertinente. Elle est cependant difficile à envisager en raison du faible nombre d'échantillons utilisés dans ce plan d'expérience

En ce qui concerne le modèle d'analyse de variance utilisé, les interactions d'ordre supérieur ou égal à deux sont difficiles à interpréter en termes de variations expérimentales et elles n'ont donc pas été prises en compte. De plus, l'analyse du graphe des résidus (cf. **Figure C.4.1**) montre que ces termes influent peu sur la reproductibilité des résultats, contrairement à ce qui serait observé si chaque valeur mesurée dépendait d'interaction d'ordre élevé entre le gène considéré et certains facteurs de l'expérience (Kerr *et al.*, 2000).

Le modèle utilisé est un modèle dans lequel tous les facteurs sont fixes. Il serait sans doute plus pertinent d'utiliser un modèle mixte en considérant le facteur « lame » comme un facteur aléatoire. Ce modèle permettrait également d'inclure le facteur « échantillon », qui ne peut pas être testé dans le modèle fixe en raison d'un manque de degrés de liberté. Ce modèle mixte n'a cependant pas été utilisé car le logiciel MAANOVA ne permet pas d'utiliser une méthode de permutation des résidus sur ce type de modèle. Par ailleurs, le

modèle implémenté est toujours un modèle croisé. Or toutes les expérimentations sur puces correspondent typiquement à des modèles hiérarchisés dans lesquels le facteur « lame », et éventuellement le facteur « échantillon », sont des facteurs partiellement confondus avec d'autres facteurs de l'expérience.

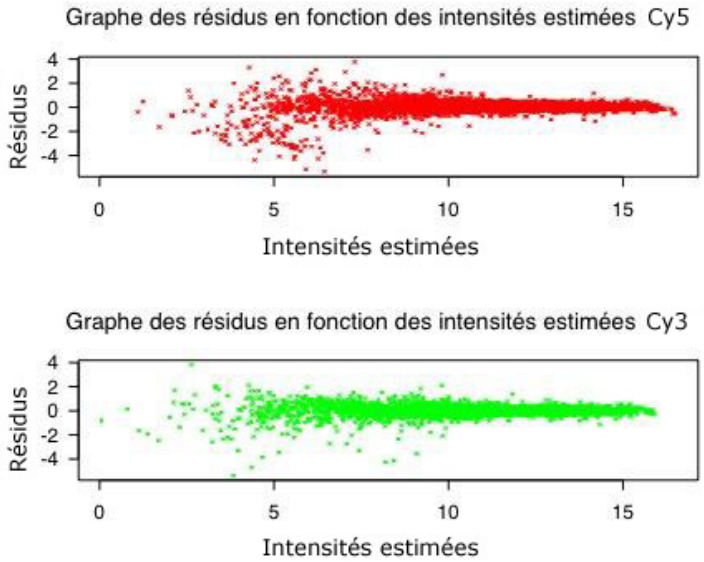


Figure C.4.1 Graphes des résidus obtenus lors de l'ajustement du modèle d'ANOVA avec interaction, pour chacun des deux fluorochromes.

L'utilisation des tests multiples pose le problème du choix du seuil de rejet des hypothèses testées car le nombre important de tests qui sont réalisés (617) augmente la proportion de faux positifs, c'est-à-dire le nombre de gènes qui apparaissent exprimés de façon différentielle alors qu'ils ne le sont pas. La plupart des méthodes classiquement utilisées pour résoudre ce problème sont extrêmement conservatives (Cui et Churchill, 2003). Elles sont essentiellement destinées à l'identification d'un très petit nombre de gènes exprimés de façon différentielle, ce qui est peu utile pour une étude telle que la réponse physiologique globale de *Buchnera* en réponse à des variations du milieu nutritionnel de son hôte. Le risque de première espèce classiquement utilisé (5 %) a donc été conservé pour l'obtention des listes de gènes exprimés de façon différentielle. Pour pallier ce problème, certaines analyses et notamment celle de la fréquence d'apparition des catégories fonctionnelles ont été réalisées sur la liste complète des gènes simplement classés en fonction de leurs probabilités. Pour de nouvelles analyses, il serait possible de calculer des probabilités corrigées en utilisant des procédures de contrôle du risque comme le *FDR* (*False Discovery Rate*). Enfin, pour de nouvelles expérimentations, il sera également envisageable de déterminer le nombre de faux positifs attendus dans l'expérience, en intégrant

au plan expérimental des lames sur lesquelles deux échantillons correspondant à une même condition sont comparés (Hung *et al.*, 2002).

4.2 Aspects biologiques

Une première étude sur les niveaux d'expression moyens a permis de montrer que les gènes les plus exprimés sont localisés préférentiellement, sur le brin direct du chromosome, et à proximité de l'origine de réplication. Cette analyse a été réalisée en l'absence de données précises concernant la position du site de terminaison chez *Buchnera*, contrairement à *Escherichia coli*, chez qui cette position a été déterminée expérimentalement. Pour poursuivre ce type d'étude de façon plus approfondie, tout en évitant l'incertitude de positionnement du site de terminaison, il serait donc possible de retirer une région comprise entre les positions 307340 et 320340 sur le chromosome, comme le proposent Palacios et Wernegreen (2002). L'étude préliminaire des niveaux d'expression en fonction de la composition des séquences a également montré que les gènes les plus GC riches, c'est-à-dire les gènes les plus conservés, sont également les plus exprimés. L'ensemble de ces résultats confirme les études théoriques réalisées par Rispe *et al.* (2004).

Par ailleurs, le calcul d'un *CAI* à partir des cinquante gènes les plus exprimés révèle qu'il n'existe vraisemblablement pas un usage préférentiel de certains codons dans les séquences des gènes les plus fortement exprimés chez *Buchnera*, contrairement à ce qui est observé chez *Escherichia coli* (Gouy et Gautier, 1982) où il existe une sélection des codons optimaux assurant une traduction rapide des gènes fortement exprimés. Réciproquement l'absence d'une telle sélection sur les gènes faiblement exprimés aboutit à la rétention de codons qui ne sont pas optimaux (Shields, 1990). Ce résultat suggère que, chez *Buchnera,* l'usage de codons riches en bases A et T a d'avantage été façonné par une pression mutationnelle et une dérive génétique plutôt que par une sélection des codons pour l'efficacité de la traduction (Wernegreen et Moran, 1999). Cependant, pour valider un tel résultat, il est essentiel d'étudier également la fréquence des ARN de transfert. En effet, il a été observé, chez *Escherichia coli,* que les gènes fortement exprimés possèdent une composition en codons plus fortement corrélée avec l'abondance des ARN de transfert que dans les gènes faiblement exprimés. Cette observation a été interprétée comme une adaptation de la composition en acides aminés des protéines avec les ARN de transfert disponibles, de façon à accroître l'efficacité de la traduction (Lobry et Gautier, 1994). La quantité d'ARN de transfert disponibles chez *Buchnera* pourrait être calculée à partir des niveaux d'expression des gènes correspondant, tout en restant vigilant sur les risques importants d'hybridation aspécifique qui peuvent

exister sur les sondes correspondant aux gènes des ARN de transfert (cf. partie B).

Une première comparaison des niveaux d'expression des gènes chez *Buchnera* et chez *Escherichia coli* a été réalisée en utilisant les *CAI* des gènes d'*Escherichia coli*. Cette étude montre une certaine similarité du profil d'expression global, mais révèle également des différences importantes parmi les gènes les plus fortement et les plus faiblement exprimés chez les deux bactéries. Ces différences sont révélatrices d'une adaptation de *Buchnera* à un mode de vie symbiotique, avec une spécialisation de la bactérie dans la fourniture des acides aminés essentiels aux pucerons et une vie dans un milieu intracellulaire très protégé. Pour confirmer ces résultats préliminaires, il est à présent essentiel de comparer les niveaux d'expression moyens avec des niveaux d'expression obtenus chez *Escherichia coli* et disponibles dans les bases de données d'expression.

Une étude des variations d'expression entre les différentes conditions expérimentales a également été initiée et sera poursuivie au laboratoire. Les variations des niveaux d'expression qui sont observées, bien que relativement faibles, sont en accord avec certaines études métaboliques qui ont été réalisées chez *Escherichia coli* (Tao *et al.*, 1999) ou chez *Bacillus subtilis* (Berka *et al.*, 2003). Cependant, contrairement à ces expérimentations réalisées avec des modifications directes de la composition du milieu de vie des bactéries, les faibles variations d'expression observées chez *Buchnera* (rapports compris entre 1,2 et 2,4) sont à mettre en relation avec le fait que la bactérie ne subit pas directement les modifications environnementales imposées aux pucerons, mais que ces variations sont atténuées par l'hôte. Par ailleurs, ces faibles variations sont observées pour l'ensemble des gènes des voies métaboliques concernées. Cette accumulation globale de faibles variations dans une cascade métabolique permet sans doute une amplification de la réponse dans l'organisme, avec un impact au moins aussi important, sinon plus, qu'une modification de la synthèse de quelques protéines isolées.

Une première analyse, basée sur un découpage des gènes en catégories fonctionnelles, révèle essentiellement une variation importante de la catégorie des ARN de transfert. Il convient cependant de rester vigilant sur l'interprétation de ces observations qui dépend fortement du choix des catégories fonctionnelles. En effet, en fonction du niveau d'étude choisi pour leur définition (niveau cellulaire ou moléculaire par exemple), la vision qui en découle peut être très différente. Ceci est d'autant plus vrai chez *Buchnera,* où de nombreux gènes assurent sans doute des fonctions qui ne sont pas celles qui ont été annotées. Enfin les catégories utilisées ici sont celles qui ont été décrites à par-

tir des gènes d'*Escherichia coli* par Riley (1993), et dans lesquelles chaque gène n'est associé qu'à un seul processus. Même si ces catégories décrivent des processus biologiques qui correspondent à l'une des classifications proposées par le consortium *Gene Ontology (GO),* l'utilisation de la classification *GO,* prenant en compte la multiplicité des fonctions biologiques pour un même gène, permettrait des analyses beaucoup plus riches. De plus, l'utilisation de *GO* pourrait être complétée par l'utilisation des deux autres catégories, proposées pour décrire les composants cellulaires et les fonctions moléculaires. Ce travail qui n'a pas été réalisé dans cette thèse, pourra être mis en œuvre lors de la poursuite des analyses.

Au sein des différentes catégories fonctionnelles, l'analyse de la répartition des gènes en groupes d'activation et de répression révèle une surexpression des gènes codant pour les ARN de transfert lorsque la concentration en saccharose augmente, c'est-à-dire lorsque le puceron est soumis à un stress osmotique. L'interprétation d'un tel résultat est cependant difficile. En effet, bien que les ARN de transfert représentent en moyenne 20 % des ARN chez les bactéries (et 5 % chez *Buchnera*), leur régulation est loin d'être élucidée. Contrairement à des études plus anciennes, une expérience réalisée chez *Bacillus subtilis* a permis de montrer récemment que la transcription et la maturation des ARN de transfert contribuent de façon plus importante à leur régulation que leur dégradation (Dittmar *et al.*, 2004). Ces résultats suggèrent que la régulation de la transcription des ARN de transfert est sans doute beaucoup plus complexe que ce qui avait été envisagé initialement.

Cette étude des catégories fonctionnelles montre également que les gènes impliqués dans les voies de synthèse des macromolécules et des protéines hypothétiques sont d'avantage activés que réprimés, dans dans les milieux nutritionnels 1 et 4 (en comparaison respectivement des milieux 3 et 2). Une expression différentielle des protéines hypothétiques, bien que difficilement interprétable à l'heure actuelle, a également été observée dans de nombreuses études dédiées à l'analyse globale des profils d'expression bactériens (Paustian *et al.*, 2002 ; Berka *et al.*, 2003).

L'étude complémentaire, qui a ensuite été réalisée au niveau des voies métaboliques à proprement parler, montre que *Buchnera* semble capable de s'adapter à une demande moins importante de son hôte, en réduisant l'expression de certains gènes impliqués dans les phénomènes de transport, dans le métabolisme de deux acides aminés essentiels (lysine et tryptophane) et dans les phénomènes de division cellulaire. Il existe en effet une différence importante dans la demande nutritionnelle des pucerons à croissance rapide (sur

milieu à 0,5 M de saccharose) et à croissance ralentie (sur milieu à 1 M de saccharose).

Par ailleurs, le gène codant pour le transporteur de mannitol est réprimé en condition de stress osmotique. Les pucerons soumis, à un stress thermique ou osmotique, produisent de grandes quantités de mannitol qui est une molécule impliquée dans les phénomènes d'osmoprotection (Hendrix et Salvucci, 1998). Ce mécanisme n'a été décrit que chez le puceron *Aphis gossypii*, mais il est probable qu'il existe également chez *Acyrthosiphon pisum*. Ce mécanisme de répression de la synthèse du transporteur de mannitol chez *Buchnera* permet peut-être, en régulant l'entrée de mannitol dans la bactérie, d'augmenter la concentration de mannitol disponible dans les bactériocytes pour lutter contre le stress osmotique. En accord avec cette observation, la bactérie semble répondre au stress osmotique subi par le puceron en activant la voie de biosynthèse du peptidoglycane, ce qui laisse supposer un épaississement de la paroi. Ce constat est néanmoins surprenant dans la mesure où des études montrent que, pour des milieux contenant entre 0,15 et 1 M de saccharose (soit une pression osmotique comprise entre 0,15 et 4 MPa), les pucerons semblent capables de maintenir la pression osmotique dans leur hémolymphe à environ 1 MPa. (Wilkinson *et al.*, 1997). L'étude révèle cependant des fluctuations individuelles de la pression de l'hémolymphe des pucerons entre 1 et 2 MPa pour un milieu contenant 1 M de saccharose. Si elle existe, la variation de pression osmotique de l'hémolymphe du puceron, en fonction de la pression du milieu nutritionnel, reste donc modérée. Des études microscopiques indiquent toutefois que la paroi de *Buchnera* est extrêmement réduite, ce qui pourrait la rendre sensible à de faibles variations de pression osmotique.

En ce qui concerne l'étude de l'effet du stress osmotique pour une variation du taux d'acides aminés essentiels, elle a permis de montrer une activation de trois gènes impliqués dans les voies de biosynthèse des acides aminés essentiels, pour un milieu « optimal » (milieu 1) contenant 50 % d'acides aminés essentiels en condition d'osmolarité normale, c'est-à-dire en condition de forte croissance et donc de demande importante des pucerons. La même activation est observée sur un milieu au contraire « minimal » (milieu 4) contenant 25 % d'acides aminés essentiels et en condition de stress osmotique. Sur un tel milieu, bien que la croissance des pucerons soit moins importante, le faible taux d'acides aminés essentiels nécessite sans doute une réponse de la bactérie pour surproduire les acides aminés en quantité insuffisante dans le milieu. Dans les deux cas, la synthèse de phosphoénol pyruvate par la glycolyse est préférée à celle d'érythrose phosphate par la voie des pentoses. Cette activation de la biosynthèse de certains acides aminés essentiels est couplée à une surexpression des gènes impliqués dans la biosynthèse du flagelle, des gènes impliqués dans

le système de sécrétion *sec* et des gènes codant les protéines de transport potentielles *ycfV* et *ynfM*. Ces systèmes sont donc de bons candidats pour le transport des acides aminés produits vers la cellule hôte. Enfin il existe dans les mêmes conditions, une répression de la synthèse de spermidine à partir de la putrescine fournie par l'hôte (Nakabachi et Ishikawa, 2000). Cette polyamine pourrait être impliquée dans un remaniement de la structure du chromosome de façon à modifier la transcription des gènes, tout comme cela a été observé dans les cellules eucaryotes (Matthews, 1993). Il s'agirait alors d'un mécanisme de régulation de la transcription non spécifique et ne nécessitant pas la présence de protéines régulatrices. Cependant à l'heure actuelle, cette hypothèse reste au stade de la spéculation.

Enfin, l'étude de la voie de biosynthèse de la phénylalanine, de la tyrosine et du tryptophane montre que les gènes *aroA* et *trpC* subissent des variations identiques de leur profil d'expression en fonction des facteurs *aa* et *saccharose*, alors que l'expression du gène *trpB* ne dépend pas de la variation du taux d'acides aminés essentiels. Des études menées chez *Bacillus subtilis* (Panina *et al.*, 2003) indiquent néanmoins que le gène *trpB* code pour une enzyme qui est régulée au niveau post-transcriptionnel, contrairement à *aroA* présente dans la même voie, qui est régulée au niveau transcriptionnel. Aucune étude ne permet actuellement d'affirmer que les mêmes mécanismes existent chez *Buchnera*, mais il est possible d'envisager une régulation post-traductionnelle de l'enzyme *trpB*, en fonction du taux d'acides aminés essentiels du milieu. Cette différence de niveau de régulation entre ces deux enzymes de la même voie de biosynthèse, pourrait ainsi expliquer leurs différences d'expression.

5

5 Conclusion

« Ne crains pas d'avancer lentement, crains seulement de rester immobile. »
Sagesse chinoise

Les études préliminaires décrites dans ce chapitre ne représentent que le début d'un long travail d'analyse et d'interprétation à venir. L'étude des niveaux d'expression en fonction de l'organisation du génome, initiée essentiellement pour valider les données, permettra sans doute, après approfondissement, d'apporter de nouvelles réponses concernant les aspects évolutifs de la symbiose. Les premiers résultats obtenus sur le métabolisme révèlent, quant à eux, que *Buchnera* est capable d'adapter l'intensité de son métabolisme global en fonction de la demande de son hôte. En ce qui concerne l'étude de la biosynthèse des acides aminés, il est certain que sa régulation potentielle intervient à différents niveaux. Appréhender son étude de façon complète nécessitera donc l'utilisation d'approches combinées impliquant les niveaux génomique, protéomique et métabolique. Il est également essentiel de disposer d'une connaissance fine des exigences physiologiques de l'hôte. Pour cela, une étude nutritionnelle complémentaire est actuellement réalisée dans le laboratoire de l'Université d'York. Elle repose sur l'utilisation successive de chacun des acides aminés essentiels marqués avec du ^{14}C dans le milieu nutritionnel des pucerons. Les pucerons utilisent donc le composé radioactif, et l'analyse de l'ensemble des produits marqués retrouvés dans les carcasses ou éliminés (CO_2 et miellat), permet de quantifier les transformations réalisées à partir de l'acide aminé étudié. L'ingestion alimentaire du puceron est calculée à l'aide de l'inuline marquée au ^{14}C, qui est un composé non métabolisable. Cette méthode, mise au point au cours de mon séjour à York, permettra de déterminer pour les différents milieux nutritionnels l'accroissement net et l'efficacité d'assimilation de chaque acide aminé essentiel dans les pucerons et par conséquent la contribution de chaque acide aminé à la synthèse protéique. La différence entre l'accroissement net des acides aminés et l'apport du milieu permettra finalement de déterminer la contribution de *Buchnera*. Il est probable que les résultats de cette étude métabolique éclaireront d'un jour nouveau les résultats obtenus au niveau transcriptomique. L'intégration de ces données de nature différente, puis la comparaison des résultats obtenus avec les études réalisées chez *Escherichia coli,* pourront alors

permettre d'initier la modélisation des réseaux de régulation chez *Buchnera*. À cette fin, les informations statiques obtenues à partir des expériences réalisées ici devront certainement être complétées par la réalisation d'expériences cinétiques, pour obtenir des indications sur la connectivité des voies des réseaux.

6

6 Vers une biologie des systèmes

« Je tiens impossible de connaître les parties sans connaître le tout, non plus que de connaître le tout sans connaître particulièrement les parties ».
Blaise Pascal[39]

Il n'existe pas réellement de définition pour la biologie des systèmes car il ne s'agit pas d'une discipline, mais plutôt d'une nouvelle façon d'appréhender l'étude du vivant (Kitano, 2002a). Une ancienne légende indienne illustre bien cette vision récente. Elle raconte l'histoire d'un groupe d'aveugles qui tenta un jour de décrire un éléphant en touchant chacun une seule de ses parties. La personne qui palpa la trompe eut une idée très différente de celle qui toucha les oreilles. Finalement le groupe ne parvint jamais à avoir une idée globale de l'éléphant. Cette histoire est en quelque sorte une allégorie de l'approche réductionniste de la biologie qui a dominé le siècle précédent. Il ne s'agit pas de renier les avancées fantastiques qui ont été réalisées en biologie moléculaire ces trente dernières années et qui ont permis de dresser un formidable inventaire des constituants cellulaires, de leur conformation et de leurs interactions. Si l'utilité des approches réductionnistes ne fait aucun doute, il devient à présent essentiel, pour étudier les propriétés des organismes vivants, de dépasser l'observation de chacun de leurs éléments pris séparément. En ce sens, la biologie des systèmes est une approche complémentaire de la biologie réductionniste. Au niveau cellulaire, elle propose par exemple d'étudier la structure et la dynamique de la cellule et sa fonction dans un organisme plutôt que les caractéristiques de ses parties isolées (Kitano, 2002b). Waddington avait pressenti ce tournant dès 1967 en écrivant :

« La biologie moléculaire nous a donné une connaissance considérable sur la nature des unités et des processus élémentaires constituant le vivant, mais il reste aux biologistes la question de comprendre comment ces unités sont assemblées pour former des systèmes. »

[39] Pascal, B. *Pensées*. Gallimard, Paris (1670).

Le développement de la biologie des systèmes se situe dans le cadre conceptuel du paradigme de la complexité, défini par Morin en 1977. En effet, les systèmes biologiques sont par nature complexes. Ce n'est cependant pas la multiplicité des composants, ni même la diversité de leurs interrelations, qui caractérisent leur complexité. En effet, tant que les composants d'un système restent dénombrables, le système est un système compliqué, et un dénombrement combinatoire pourrait permettre la description de tous ses comportements possibles. Ce qui fait la complexité d'un système, c'est en revanche l'imprévisibilité potentielle (non calculable *a priori*) de ces comportements, liée en particulier à l'apparition de phénomènes d'émergence. L'étude de ces phénomènes se révèle fondamentale pour la compréhension du fonctionnement des systèmes vivants. En reprenant l'exemple précédent du niveau cellulaire, cette démarche nécessite donc une représentation claire de l'intérieur d'une cellule, représentation qui est étrangement absente de la littérature. Pour pallier ce manque, Goodsell (1991) propose une vision originale de la cellule. Cette vision, éloignée des représentations habituelles, nous renseigne sur la complexité des interactions à envisager pour appréhender la cellule dans sa globalité (cf. **Figure C.6.1**).

L'environnement intracellulaire est le siège d'une grande variété de réactions enzymatiques, de transports de molécules, de diffusion ou encore d'assemblages de protéines. Ce réseau de milliers de réactions chimiques, activées ou réprimées différemment en fonction du type de cellule et de l'état dans lequel la cellule se trouve, est en perpétuelle évolution (Atlan, 2002). Pour tenter de rassembler toutes les pièces de ce puzzle en un ensemble cohérent, la modélisation constitue une approche pertinente, non seulement pour prédire les comportements du système, mais aussi pour structurer et organiser les connaissances. Cette démarche semble d'ailleurs s'imposer tout naturellement comme l'illustre une des répliques du film *π [Pi]* réalisé en 1997 par Darren Aronofsky :

> *« Les maths sont le langage de la nature. Tout ce qui nous entoure peut-être compris par les nombres. En représentant graphiquement les nombres d'un système, on obtient des modèles. Conclusion : les modèles sont partout dans la nature. »*

L'élaboration de tels modèles impose une approche pluridisciplinaire et collective associant aux connaissances biologiques et au formalisme mathématique, les simulations informatiques. Certains auteurs ont d'ailleurs tenté d'intégrer les premiers modèles d'événements cellulaires dans un environnement informatique. Actuellement, aucun outil n'est capable de mimer le fonctionnement complet d'une cellule, cependant de nombreux développements sont en cours comme le logiciel *Cellware* (Dhar *et al.*, 2004).

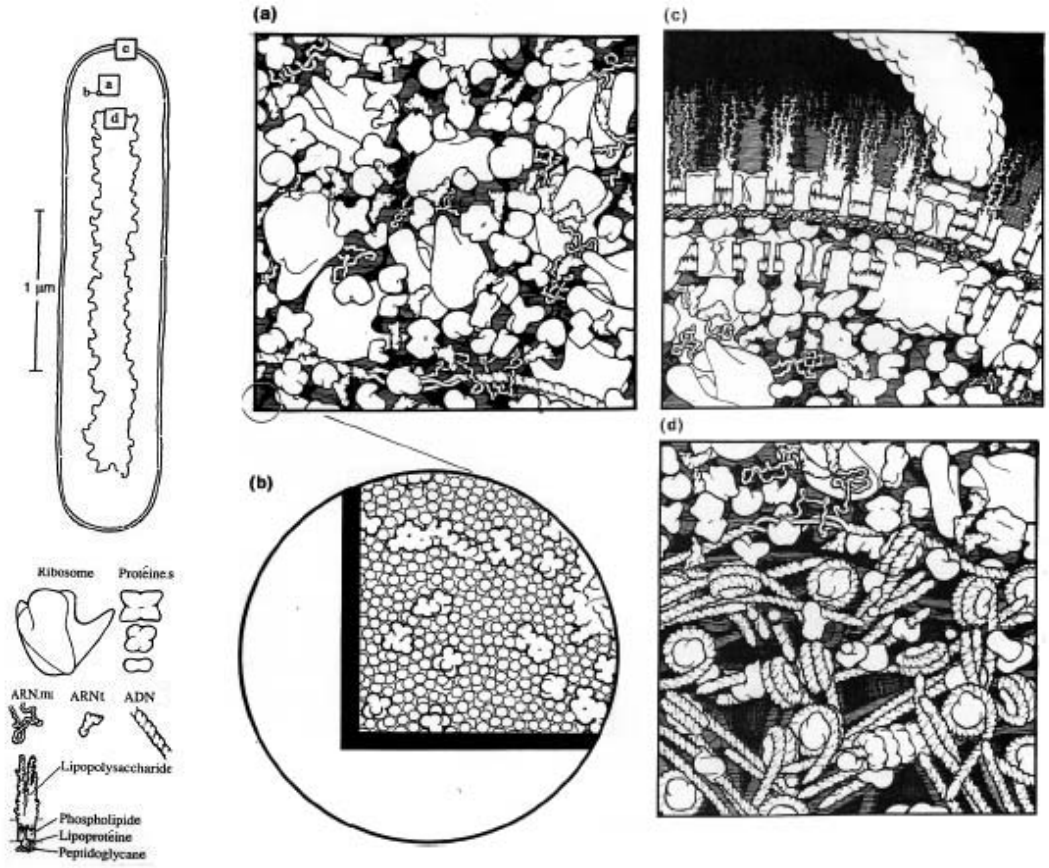


Figure C.6.1 Si vous pouviez grossir une cellule un million de fois, pour obtenir des molécules de la taille des objets quotidiens, que verriez-vous ?
Trois parties d'une cellule telle qu'*Escherichia coli* ont été grossies un million de fois : le cytoplasme (a), la membrane cellulaire (c) et la région nucléaire (d) avec l'ensemble des macromolécules. L'eau, les cofacteurs et les intermédiaires de biosynthèse ne sont pas représentés dans un souci de clarté de la distribution des molécules. L'agrandissement d'une portion de cytoplasme (b) permet cependant de visualiser toutes les molécules avec les molécules d'eau sous forme de cercles et les petites molécules soulignées en noir (d'après Goodsell, 1991).

Finalement, la biologie des systèmes étudie les systèmes biologiques en les perturbant puis en observant l'expression de leurs gènes, la production des protéines et l'activation des voies métaboliques qui en découlent. Le but est d'intégrer ces données pour formuler des modèles mathématiques décrivant la structure du système et ses réponses aux variations environnementales (Ideker *et al.*, 2001a). En ce sens, le développement des puces à ADN, en permettant de mesurer l'expression de milliers de gènes simultanément, s'inscrit donc complètement dans la biologie des systèmes et le travail présenté dans cette thèse

représente un premier pas dans cette approche du vivant, qualifiée parfois de biologie intégrative. Les analyses préliminaires, qui ont été réalisées à partir des profils d'expression obtenus, ont permis de montrer certaines corrélations entre expression des gènes et conditions biologiques. Cependant l'utilisation des seuls résultats de puces à ADN n'offre pas la possibilité d'aborder complètement les éléments de régulation chez *Buchnera*. Cette information sur les profils d'expression est nécessaire pour caractériser la structure du réseau de gènes, mais elle n'est pas suffisante pour modéliser de façon complète leur dynamique et les propriétés fonctionnelles. Pour poursuivre ce travail, il sera donc nécessaire d'intégrer d'avantage de données. Pour cela, des études devront également être initiées au niveau protéomique afin d'associer à l'étude de l'expression des gènes l'analyse de la quantité de protéines obtenues et de leurs modifications (Hecker et Engelmann, 2000). Enfin, il sera nécessaire d'intégrer des données cinétiques pour modéliser la dynamique du réseau. En effet, la plupart des voies métaboliques sont actuellement représentées dans un plan, mais il deviendra sans doute essentiel dans les années à venir de représenter également les informations spatiales et dynamiques (Shapiro et Losick, 2000). Cette biologie *in silico* est un premier pas vers la compréhension des mécanismes sous-jacents aux phénomènes d'émergence dans les systèmes vivants. Elle permettra certainement de vérifier que « le tout est plus que la somme des parties » (Gibbs, 2001).

Partie D
Conclusion générale et perspectives

« Si nous sommes dans la bonne direction, tout ce que nous avons à faire, c'est de continuer à marcher. »
Sagesse bouddhiste

L'histoire de cette thèse a débuté autour du modèle de la symbiose entre la bactérie *Buchnera aphidicola* et le puceron du pois *Acyrthosiphon pisum,* avec pour toile de fond l'adaptation du métabolisme de *Buchnera* aux besoins de son hôte. Pour tenter de répondre à cette problématique, il a été décidé au laboratoire d'appréhender de façon globale les réponses métaboliques de *Buchnera*, en étudiant son transcriptome par la technologie des puces à ADN. La réalisation de telles expérimentations a donc nécessité la conception d'une puce à ADN dédiée à *Buchnera*. Dans ce contexte à la fois biologique et méthodologique, les objectifs de cette thèse ont été à la fois de tenter de lever une partie de l'énigme de la régulation de l'expression des gènes chez *Buchnera,* et d'apporter une contribution à la bioinformatique des puces à ADN, pour les différentes étapes de leur utilisation (cf. **Figure D.1**).

D'un point de vue méthodologique, ce travail a permis le développement de ROSO, un logiciel de Recherche et d'Optimisation de Sondes Oligonucléotidiques qui est utilisé par la communauté scientifique sur le site du PBIL. De nombreuses évolutions peuvent être envisagées pour adapter ROSO à des applications dépassant l'étude du transcriptome, et notamment la détermination de sondes destinées au génotypage. L'utilisation de ROSO pour déterminer des sondes spécifiques sur le génome de *Buchnera* a abouti à la mise au point d'une démarche originale d'utilisation du logiciel. Cette démarche a été appliquée par la suite pour déterminer des sondes pour divers utilisateurs de puces. Les sondes obtenues pour *Buchnera* ont été utilisées pour concevoir une puce dédiée. De nombreux développements expérimentaux ont ensuite été réalisés sur une mini-puce puis sur la puce complète. Ils ont permis d'une part de choisir les techniques les plus adaptées à une utilisation optimale de la puce, et d'autre part, de définir un protocole complet d'acquisition des données permettant de passer de l'image des lames à des données filtrées, normalisées et moyennées. De nouveaux développements sont actuellement en cours au laboratoire pour mettre en place une technique de normalisation basée sur l'utilisation de l'ADN génomique. Elle permettra de réaliser des comparaisons entre les expériences déjà réalisées et celles à venir. Enfin, l'utilisation de méthodes originales d'analyse des gènes différentiels a été initiée, avec notamment l'utilisation du regroupement des gènes en catégories fonctionnelles, ou encore la visualisation des données génomiques sur des cartes métaboliques. Il est à présent essentiel d'y associer l'application d'ontologies au moyen d'outils comme la bibliothèque *Gene Ontology* de *Bioconductor*.

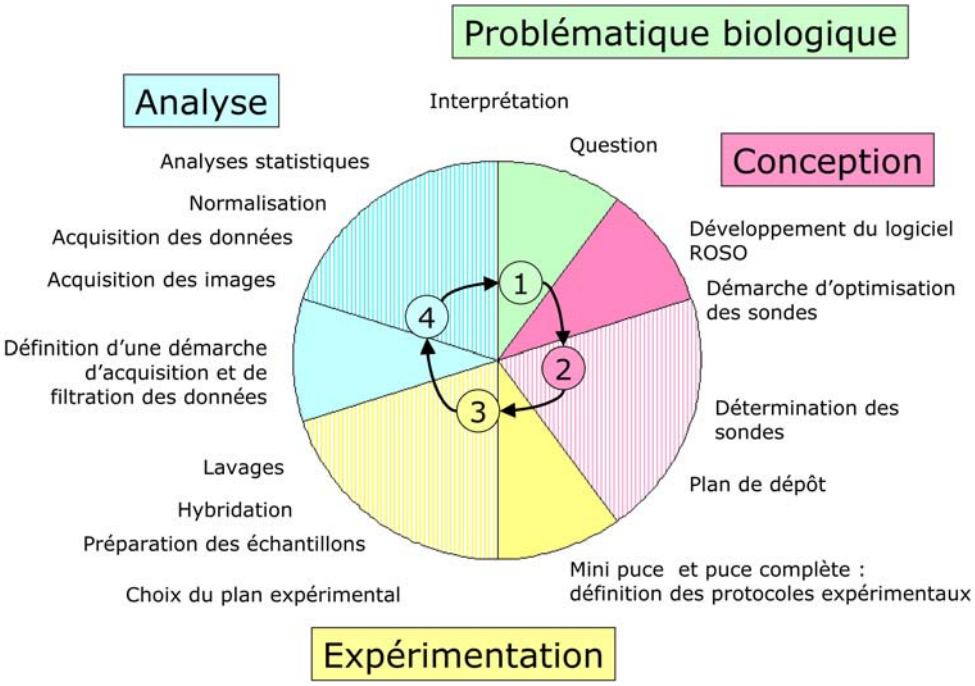


Figure D.1 Représentation schématique du travail réalisé dans cette thèse pour chacune des étapes d'utilisation d'une puce à ADN.

D'un point de vue biologique, l'utilisation de la puce dédiée à *Buchnera* a offert la possibilité d'étudier le transcriptome de bactéries issues de pucerons élevés sur des milieux contenant des quantités variables d'acides aminés essentiels et de saccharose. Des études préliminaires ont montré une corrélation entre l'organisation du génome et les niveaux d'expression. Les gènes les plus exprimés chez *Buchnera* sont en effet les plus riches en bases G et C et sont localisés préférentiellement sur le brin direct et à proximité de l'origine de réplication. De nombreuses analyses devront compléter ces premiers résultats, et notamment en ce qui concerne l'usage des codons et l'étude des quantités d'ARN de transfert disponibles. Pour de nouvelles études des niveaux d'expression des gènes en fonction de la localisation, il sera important de ne pas négliger l'organisation spatiale du génome, qui est essentielle à sa fonction. Des approches structurales et biochimiques réalisées chez *Escherichia coli* ont d'ailleurs montré que des gènes, présentant des niveaux d'expression similaires et sous le contrôle d'un même régulateur, sont disposés sur le chromosome selon une périodicité correspondant au nombre de boucles formées par le super-enroulement de l'ADN (Kepes, 2004). Une telle périodicité, si elle existe chez *Buchnera,* pourrait certainement guider la recherche d'éléments de régulation dans le génome.

En ce qui concerne les aspects physiologiques, les premiers résultats obtenus indiquent que *Buchnera* est capable de réguler son expression génique

et de réorienter son métabolisme pour s'adapter aux besoins de son hôte. En effet, une demande accrue du puceron, liée à une croissance importante ou à un milieu nutritionnel inadapté, se traduit par une activation de l'expression des gènes impliqués dans les phénomènes de division cellulaire, dans le métabolisme de certains acides aminés essentiels et une modification de l'expression de certains gènes du métabolisme des sucres. Parallèlement à ces activations, il semble qu'il existe une surexpression des gènes codant, pour certaines protéines du complexe flagellaire, pour le système de sécrétion *sec*, pour une porine et pour deux protéines de transport hypothétiques. Il est donc probable que ces différents systèmes sont utilisés par la bactérie pour exporter les acides aminés essentiels vers son hôte. Enfin, cette étude pointe un système de régulation potentiel de la fourniture des acides aminés essentiels, avec des variations de l'expression d'une enzyme permettant la biosynthèse d'une polyamine, éventuellement impliquée dans des phénomènes de régulation de la structure du chromosome. Ces résultats obtenus sur puce à ADN nécessitent à présent des validations expérimentales des niveaux d'expression observés, qui seront réalisées pour certains gènes clés par RT-PCR quantitative. Il sera également nécessaire de compléter les analyses et d'approfondir les premières interprétations qui ont été réalisées dans cette thèse.

De nombreuses interrogations demeurent pour élucider complètement les interactions symbiotiques entre *Buchnera* et le puceron. En effet, si les premiers résultats obtenus indiquent que *Buchnera* est capable d'adapter son métabolisme à la demande du puceron, il reste maintenant à comprendre comment. Seule une démarche multicritère est susceptible d'appréhender cette question dans sa globalité et sa complexité (Gibbs, 2001).

Une première partie des réponses pourra être apportée par des utilisations variées de la puce. Des études cinétiques offriront la possibilité d'étudier la dynamique des profils d'expression et l'utilisation de la puce dans des conditions très différentes permettra d'obtenir un panorama des réponses de *Buchnera*. Il sera par exemple possible d'étudier l'influence de *Buchnera* dans l'adaptation des pucerons à des plantes hôtes spécifiques (Luzerne, Pois ou Trèfle), ou l'influence de *Buchnera* au cours du développement des pucerons (avec des *Buchnera* prélevées à différents stades embryonnaires et larvaires ou encore chez des pucerons âgés). Par ailleurs, si l'analyse du transcriptome de *Buchnera* était certainement le premier niveau à considérer pour l'étude de la régulation de l'expression des gènes, il est important de conserver à l'esprit que les informations obtenues ne sont pas suffisantes. Pour les compléter, il sera nécessaire de relier quantités d'ARN et de protéines par des études du protéome de *Buchnera* (Greenbaum *et al.*, 2002).

Dans le système complexe de la symbiose, une autre partie des réponses ne pourra être obtenue qu'avec l'acquisition de connaissances sur l'hôte. Des connaissances sur le génome du puceron permettront par exemple d'étudier les éventuelles translocations de gènes de la bactérie dans le noyau de son hôte. Pour cela, un consortium international de la Génomique des Pucerons[40] a été créé à Paris en juin 2003 (*IAGC* pour *International Aphid Genomics Consortium*). Ce Consortium, qui regroupe des laboratoires de onze pays, a proposé le séquençage du génome du puceron du pois, *Acyrthosiphon pisum*, mais aussi le développement d'une puce contenant des EST de pucerons, qui a été réalisée au laboratoire de Biologie des Organismes et des Populations appliquées à la Protection des Plantes de l'INRA de Rennes (Denis Tagu). Il ne fait aucun doute que la connaissance de l'expression des gènes chez le puceron apportera un éclairage nouveau à l'étude de l'expression des gènes chez *Buchnera*.

Enfin, l'étude du couple symbiotique pourra certainement s'enrichir de toutes les recherches initiées sur le pois et la luzerne, qui en tant que plantes hôtes du puceron représentent le dernier élément d'un système d'interactions à trois partenaires.

Les réponses qui seront accumulées au fil des expériences constitueront la base nécessaire à une démarche de modélisation. Dans ce contexte de recherche résolument collectif et pluridisciplinaire, il sera également essentiel d'intégrer les réflexions sur la complexité et sur l'importance de la modélisation et de la simulation qui ont été initiées en sciences humaines (Kepes, 2001). L'enjeu pour les années à venir sera finalement de rassembler toutes les pièces du puzzle en un ensemble cohérent, de façon à aboutir à la compréhension du fonctionnement global de *Buchnera*. Les premières pièces de ce puzzle commencent tout juste à être assemblées, mais comme le disait George Pérec[41] :

> *«On peut regarder une pièce de puzzle pendant trois jours et croire tout savoir de sa configuration et de sa couleur sans avoir le moins du monde avancé : seule compte la possibilité de relier cette pièce à d'autres pièces […] ; seules les pièces rassemblées prendront un caractère lisible, prendront un sens : considérée isolément une pièce d'un puzzle ne veut rien dire ; elle est seulement question impossible, défi opaque ; mais à peine a-t-on réussi, au terme de plusieurs minutes d'essais et d'erreur, ou en une demi-seconde prodigieusement inspirée, à la connecter à l'une de ses voisines, que la pièce disparaît, cesse d'exister en tant que pièce : l'intense difficulté qui a précédé ce*

[40] http://www.princeton.edu/%7Edstern/AphidResLinks.htm.
[41] Pérec, G. *La vie mode d'emploi*. Hachette/POL, Paris (1978).

rapprochement, et que le mot puzzle – énigme – désigne si bien en anglais, non seulement n'a plus de raison d'être, mais semble n'en avoir jamais eu, tant elle est devenue évidence : les deux pièces miraculeusement réunies n'en font plus qu'une, à son tour source d'erreur, d'hésitation, de désarroi et d'attente. »

Publications et communications

Publications

REYMOND, N., CHARLES, H., DURET, L., CALEVRO, F., BESLON, G. et FAYARD, J.-M. ROSO: Optimizing oligonucleoitde probes for microarrays. *Bioinformatics*. 2004, vol. 20, n°2, pp. 271-273.

REYMOND, N., CHARLES, H., ROME, S. et MARTY, J. Les données d'expression. In: *Boulicaut J.-F. et Gandrillon, O. Informatique pour l'analyse du transcriptome (traité IC2)*. Lavoisier. Paris: Hermès sciences publication, 2004.

CALEVRO, F., CHARLES, H., **REYMOND, N.,** DUGAS, V., CLOAREC, J.-P., BERNILLON, J., RAHBE, Y., FEBVAY, G. et FAYARD, J.-M. Assessment of 35mer amino-modified oligonucleotide based microarray with bacterial samples. *Journal of Microbiological Methods*. 2004, vol. 57, n°2, pp. 207-218.

REYMOND, N., ALLEN, C., MORIN, N., CALEVRO, C., BERNILLON, J., RAHBÉ, Y., FEBVAY, G., LAUGIER, C., FAYARD, J.-M., DOUGLAS, A. et CHARLES, H. Exploring global gene expression profiles of *Buchnera aphidicola* under essential amino acid rate and saccharose concentration in the diet of its symbiotic partner *Acyrtosiphum pisum*, (en préparation).

CALEVRO, F., **REYMOND, N.,** MORIN, N., BERNILLON, J., RAHBÉ, Y., FEBVAY, G., LAUGIER, C., FAYARD, J.-M. et CHARLES, H. Genome wide transcriptionnal changes associated with nutritional alterations affecting phenylalanine and tyrosine metabolism in *Buchnera aphidicola*, (en préparation).

OCCHIALINI, A., CUNNAC, S., **REYMOND, N.,** GENIN, S. et BOUCHER, C. Genome-wide analysis of gene expression in *Ralstonia solanecearum* reveals that the hrpB gene acts as a regulatory switch controlling multiple virulence pathways, (en préparation).

Communications

REYMOND, N., CHARLES, H., BESLON, G. et FAYARD, J.-M. ROSO: a software to search optimized oligonucleotide probes for microarrays. *ISMB2002: Intelligent Systems for Molecular Biology (Edmonton, Canada)*, 3-7 août 2002.

REYMOND, N., CHARLES, H., BESLON, G. et FAYARD, J.-M. ROSO: a software to search optimized oligonucleotide probes for microarrays. *JOBIM 2002: Journées Ouvertes Biologie Informatiques et Mathématiques (Saint-Malo, France)*, 10-12 juin 2002.

REYMOND, N., CHARLES, H., BESLON, G. et FAYARD, J.-M. Développement d'un logiciel d'optimisation de sondes oligonucléotidiques destinées aux puces à ADN. *JPGD'01 : Journées Post-Génomique de la Doua (Villeurbanne, France)*, 05-06 avril 2001.

CHARLES, H., **REYMOND, N.,** RAHBE, Y., HEDDI, A., FEBVAY, G. et FAYARD, J.-M. Analyse du transcriptome de *Buchnera aphidicola*. *JPGD'01 : Journées Post-Génomique de la Doua (Villeurbanne, France)*, 05-06 avril 2001.

CALEVRO, F., CHARLES, H., **REYMOND, N.,** CLOAREC, J.-P., RAHBÉ, Y., HEDDI, A., FEBVAY, G. et FAYARD, J.-M. Transcriptome analysis in *Buchnera aphidicola*: development of a first DNA chip. *JPGD'02 : Journées Post-Génomique de la Doua (Villeurbanne, France)*, 21-22 mars 2002.

CALEVRO, F., CHARLES, H., **REYMOND, N.,** DUGAS, V., CLOAREC, J.-P., BERNILLON, J., RAHBÉ, Y., FEBVAY, G. et FAYARD, J.-M. A preliminary methodological approach to study the transcriptome of *Buchnera*, the intracellular symbiotic bacteria of aphids. *JPGD'03 : Journées Post-Génomique de la Doua (Villeurbanne, France)*, 14-16 mai 2003.

CALEVRO, F., CHARLES, H., **REYMOND,** N., DUGAS, V., CLOAREC, J.-P., BERNILLON, J., RAHBÉ, Y., FEBVAY, G. et FAYARD, J.-M. Transcriptome analysis of *Buchnera*, the intracellular symbiotic bacteria of aphids: a methodological approach. *CPI 2003 : Quatorzième Congrès de Physiologie de l'Insecte (Amiens, France)*, 14-16 avril 2003.

Partie E
Bibliographie

ABBOT, P. et MORAN, N. A. Extremely low levels of genetic polymorphism in endosymbionts (*Buchnera*) of aphids (*Pemphigus). Molecular Ecology*. 2002, vol. 11, n°12, pp. 2649-2660.

ABISGOLD, J., SIMPSON, S. et DOUGLAS, A. Nutrient regulation in the pea aphid *Acyrthosiphon Pisum* - application of a novel geometric framework to sugar and amino acid consumption. *Physiological Entomology*. 1994, vol. 19, n°2, pp. 95-102.

AGARRABERES, F. et DICE, J. Protein translocation across membranes. *Biochimica Biophysica Acta*. 2001, vol. 1513, n°1, pp. 1-24.

AHRENDT, S., HALACHMI, S., CHOW, J., WU, L., HALACHMI, N., YANG, S., WEHAGE, S., JEN, J. et SIDRANSKY, D. Rapid p53 sequence analysis in primary lung cancer using an oligonucleotide probe array. *Proceedings of the National Academy of Sciences of the USA*. 1999, vol. 96, n°13, pp. 7382-7387.

AKMAN, L. et AKSOY, S. A novel application of gene arrays : *Escherichia coli* array provides insight into the biology of the obligate endosymbiont of tsetse flies. *Proceedings of the National Academy of Sciences of the USA*. 2001, vol. 98, n°13, pp. 7546-7551.

AKUTSU, T., MIYANO, S. et KUHARA, S. Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *Journal of Computional Biology*. 2000, vol. 7, n°3-4, pp. 331-343.

ALLAWI, H. et SANTALUCIA, J. Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry*. 1997, vol. 36, pp. 10581-10594.

ALLAWI, H. et SANTALUCIA, J. Nearest-neighbor thermodynamics of internal A.C mismatches in DNA : sequence dependence and pH effects. *Biochemistry*. 1998a, vol. 37, pp. 9435-9444.

ALLAWI, H. et SANTALUCIA, J. Nearest-neighbor thermodynamics parameters for internal G.A mismatches in DNA. *Biochemistry*. 1998b, vol. 37, pp. 2170-2179.

ALLAWI, H. et SANTALUCIA, J. Thermodynamics of internal C.T mismatches in DNA. *Nucleic Acids Research*. 1998c, vol. 26, n°11, pp. 2694-2701.

ALON, U., BARKAI, N., NOTTERMAN, D., GISH, K., YBARRA, S., MACK, D. et LEVINE, A. Broad pattern of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the USA*. 1999, vol. 96, pp. 6745-6750.

ALTER, O., BROWN, P. et BOTSTEIN, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the USA*. 2000, vol. 97, n°18, pp. 10101-10106.

ALTSCHUL, S., MADDEN, T., SCHÄFFER, A., ZHANG, J., ZHANG, Z., MILLER, W. et LIPMAN, D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. 1997, vol. 25, n°17, pp. 3389-3402.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. et LIPMAN, D. J. Basic local alignment search tool. *Journal of Molecular Biology*. 1990, vol. 215, n°3, pp. 403-410.

ALWINE, J., KEMP, D. et STARK, G. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and

hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the USA*. 1977, vol. 74, n°12, pp. 5350-5354.

ANDERSSON, J. Evolutionary genomics: is *Buchnera* a bacterium or an organelle? *Current Biology*. 2000, vol. 10, n°23, pp. 866-868.

ANDERSSON, J. et ANDERSSON, S. Genome degradation is an ongoing process in *Rickettsia*. *Molecular Biology and Evolution*. 1999, vol. 16, n°9, pp. 1178-1191.

ANDERSSON, S., ALSMARK, C., CANBACK, B., DAVIDS, W., FRANK, C., KARLBERG, O., KLASSON, L., ANTOINE-LEGAULT, B., MIRA, A. et TAMAS, I. Comparative genomics of microbial pathogens and symbionts. *Bioinformatics*. 2002, vol. 18, n°supplement 2, pp. 17.

ANDERSSON, S. et KURLAND, C. Reductive evolution of resident genomes. *Trends in Microbiology*. 1998, vol. 6, pp. 263-268.

ARFIN, S., LONG, A., ITO, E., TOLLERI, L., RIEHLE, M., PAEGLE, E. et HATFIELD, G. Global gene expression profiling in *Escherichia coli* K12 - The effects of integration host factor. *Journal of Biological Chemistry*. 2000, vol. 275, n°38, pp. 29672-29684.

ASHBURNER, M., .BALL, C., BLAKE, J., BOTSTEIN, D., BUTLER, H., CHERRY, J., DAVIS, A., DOLINSKI, K., DWIGHT, S. E., J., HARRIS, M., HILL, D., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J., RICHARDSON, J., RINGWALD, M., RUBIN, G. et SHERLOCK, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*. 2000, vol. 2, n°1, pp. 25-29.

ASHFORD, D., SMITH, W. et DOUGLAS, A. Living on a high sugar diet: the fate of saccharose ingested by a phloem-feeding insect, the pea aphid *Acyrthosiphon pisum*. *Journal of Insect Physiology*. 2000, vol. 46, n°3, pp. 335-341.

ASYALI, M., SHOUKRI, M., DEMIRKAYA, O. et KHABAR, K. Assessment of reliability of microarray data and estimation of signal thresholds using mixture modeling. *Nucleic Acids Research*. 2004, vol. 32, n°8, pp. 2323-2335.

ATLAN, H. La cellule vivante : un paradigme des systèmes naturels complexes. *Médecine/Sciences*. 2002, vol. 18, pp. 764-766.

AWADÉ, A., CLEUZIAT, P., GONZALÈS, T. et ROBERT-BAUDOUY, J. Characterization of the *pcp* gene encoding the pyrrolidone carboxyl peptidase of *Bacillus subtilis*. *FEBS Letters*. 1992, vol. 305, n°1, pp. 67-73.

BAGHDOYAN, S., ROUPIOZ, Y., PITAVAL, A., CASTEL, D., KHOMYAKOVA, E., PAPINE, A., SOUSSALINE, F. et GIDROL, X. Quantitative analysis of highly parallel transfection in cell microarrays. *Nucleic Acids Res*. 2004, vol. 32, n°9, pp. e77.

BAINS, W. Selection of oligonucleotide probes and experimental conditions for multiplex hybridization experiments. *GATA*. 1994, vol. 11, n°3, pp. 49-62.

BALDI, P. et LONG, A. A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*. 2001, vol. 17, n°6, pp. 509-519.

BALDINO, F., CHESSELET, M.-F. et LEWIS, M. High-resolution *in situ* hybridization histochemistry. *Methods in Enzymolology*. 1989, vol. 168, pp. 761-777.

BARCZAK, A., RODRIGUEZ, M., HANSPERS, K., KOTH, L., TAI, Y., BOLSTAD, B., SPEED, T. et ERLE, D. Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Research*. 2003, vol. 13, n°7, pp. 1775-1785.

BAUMANN, L. et BAUMANN, P. Growth kinetics of the endosymbiont Buchnera aphidicola in the aphid Schizaphis graminum. *Applied and Environmental Microbiology*. 1994, vol. 60, n°9, pp. 3440-3443.

BAUMANN, L. et BAUMANN, P. Characterization of ftsZ, the cell division gene of Buchnera aphidicola (endosymbiont of aphids) and detection of the product. *Current Microbiology*. 1998, vol. 36, n°2, pp. 85-89.

BAUMANN, L., BAUMANN, P. et THAO, M. Detection of messenger RNA transcribed from genes encoding enzymes of amino acid biosynthesis in *Buchnera aphidicola* (Endosymbiont of aphids). *Current Microbiology*. 1999, vol. 38, n°2, pp. 135-136.

BAUMANN, P., BAUMANN, L. et CLARK, M. Levels of *Buchnera aphidicola* chaperonin *GroEL* during growth of the aphid *Schizaphis graminum*. *Current Microbiology*. 1996, vol. 32, pp. 279-285.

BAUMANN, P., BAUMANN, L., LAI, C. Y., ROUHBAKHSH, D., MORAN, N. et CLARK, M. Genetics, physiology and evolutionary relationships of the genus *Buchnera*: Intracellular symbionts of aphids. *Annual Review of Microbiology*. 1995, vol. 49, pp. 55-94.

BEATTIE, K., EGGERS, K., SHUMAKER, J., HOGAN, M., VARMA, R., LAMTURE, J., HOLLIS, M., EHRLICH, D. et RATHMAN, D. Genosensor technology. *Clinical Chemistry*. 1993, vol. 39, pp. 719-722.

BELLIS, M. et CASELLAS, P. La puce à ADN : un multi-réacteur de paillasse. *Médecine/Sciences*. 1997, vol. 13, pp. 1317-1324.

BENJAMINI, Y. et HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*. 1995, vol. 85, pp. 289-300.

BERKA, R., CUI, X. et YANOFSKY, C. Genomewide transcriptional changes associated with genetic alterations and nutritional supplementation affecting tryptophan metabolism in *Bacillus subtilis*. *Proceedings of the National Academy of Sciences of the USA*. 2003, vol. 100, n°10, pp. 5682-5687.

BERNAYS, E. et KLEIN, B. Quantifying the symbiont contribution to essential amino acids in aphids: the importance of tryptophan for *Uroleucon ambrosiae*. *Physiological Entomology*. 2002, vol. 27, n°4, pp. 275-284.

BERTUCCI, F., BERNARD, K., LORIOD, B., CHANG, Y., GRANJEAUD, S., BIRNBAUM, D., NGUYEN, C., PECK, K. et JORDAN, B. Sensitivity issues in DNA array-based expression measurements: advantages of nylon membranes. *Human Molecular Genetics*. 1999, vol. 8, n°9, pp. 1715-1722.

BERTUCCI, F., NASSER, V., GRANJEAUD, S., EISINGER, F., ADELAÏDE, J., TAGETT, R., LORIOD, B., GIACONIA, A., BENZIANE, A., DEVILARD, E., JACQUEMIER, J., VIENS, P., NGUYEN, C., BIRNBAUM, D. et HOULGATTE, R. Gene expression profiling of poor-prognosis primary breast cancer correlate with survival. *Human Molecular Genetics*. 2002, vol. 11, n°8, pp. 863-872.

BILBAN, M., BUEHLER, L., HEAD, S., DESOYE, G. et QUARANTA, V. Normalizing DNA microarray data. *Current Issues in Molecular Biology*. 2002, vol. 4, n°2, pp. 57-64.

BIRKLE, L., MINTO, L. et DOUGLAS, A. Relating genotype and phenotype for tryptophan synthesis in an aphid-bacterial symbiosis. *Physiological Entomology*. 2002, vol. 27, pp. 1-17.

BLAKE, W., KAERN, M., CANTOR, C. et COLLINS, J. Noise in eukaryotic gene expression. *Nature*. 2003, vol. 422, n°6932, pp. 633-637.

BLAND, J. et ALTMAN, D. Multiple signifiance tests: the Bonferroni method. *British Medical Journal*. 1995, vol. 310, pp. 170.

BLATTNER, F., PLUNKETT, G., BLOCH, C., PERNA, N., BURLAND, V., RILEY, M., COLLADO-VIDES, J., GLASNER, J., RODE, C., MAYHEW, G., GREGOR, J., DAVIS, N., KIRKPATRICK, H., GOEDEN, M., ROSE, D., MAU, B. et SHAO, Y. The complete genome sequence of *Escherichia coli K-12*. *Science*. 1997, vol. 277, n°5331, pp. 1453-1474.

BO, T., DYSVIK, B. et JONASSEN, I. LSimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Research*. 2004, vol. 32, n°3, pp. e34.

BOMMARITO, S., PEYRET, N. et SANTALUCIA, J. Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic Acids Research*. 2000, vol. 28, n°9, pp. 1929-1934.

BORER, P., DENGLER, B. et TINOCO, I. Stability of ribonucleic acid double-stranded helices. *Journal of Molecular Biology*. 1974, vol. 86, pp. 843-853.

BOULLARD, B. *Guerre et paix dans le règne végétal*. Ellipses. Paris, 1990.

BOZDECH, Z., ZHU, J., JOACHIMIAK, M., COHEN, F., PULLIAM, B. et DERISI, J. Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biology*. 2003, vol. 4, pp. R9.

BRAASCH, D. et COREY, D. Locked nucleic acid (LNA): fine-tuning the recognition of DNA and RNA. *Chemical Biology*. 2001, vol. 8, n°1, pp. 1-7.

BRAENDLE, C., MIURA, T., BICKEL, R., SHINGLETON, A., KAMBHAMPATI, S. et STERN, D. Developmental origin and evolution of bacteriocytes in the Aphid-*Buchnera* symbiosis. *PLOS Biology*. 2003, vol. 1, n°1, pp. 70-76.

BRANDT, O., FELDNER, J., STEPHAN, A., SCHRODER, M., SCHNOLZER, M., ARLINGHAUS, H., HOHEISEL, J. et JACOB, A. PNA microarrays for hybridisation of unlabelled DNA samples. *Nucleic Acids Research*. 2003, vol. 31, n°19, pp. e119.

BRAZMA, A., HINGAMP, P., QUACKENBUSH, J., SHERLOCK, G., SPELLMAN, P., STOECKERT, C., AACH, J., ANSORGE, W., BALL, C., CAUSTON, H., GAASTERLAND, T., GLENISSON, P., HOLSTEGE, F., KIM, I., MARKOWITZ, V., MATESE, J., PARKINSON, H., ROBINSON, A., SARKANS, U., SCHULZE-KREMER, S., STEWART, J., TAYLOR, R., VILO, J. et VINGRON, M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*. 2001, vol. 29, n°4, pp. 365-371.

BRAZMA, A. et VILO, A. Gene expression data analysis. *FEBS Letters*. 2000, vol. 480, pp. 17-24.

BREITKREUTZ, B.-J., STARK, C. et TYERS, M. Osprey: a network visualization system. *Genome biology*. 2003, vol. 4, pp. R22.

BRESLAUER, K., FRANK, R., BLÖCKER, H. et MARKY, L. Predicting DNA duplex stability from the base sequence. *Proceedings of the National Academy of Sciences of the USA*. 1986, vol. 83, pp. 3746-3750.

BROUGH, C. et DIXON, A. Ultrastructural features of egg development in oviparae of the vetch aphid, *Megoura viciae* Buckton. *Tissue Cell*. 1990, vol. 22, n°1, pp. 51-63.

BUCHNER, P. *Aphids*. New York: Interscience, 1965.

BURTON, G., GUAN, Y., NAGARAJAN, R. et MCGEHEE, R. Microarray analysis of gene expression during early adipocyte differentiation. *Gene*. 2002, vol. 293, n°1-2, pp. 21-31.

CALEVRO, F., CHARLES, H., REYMOND, N., DUGAS, V., CLOAREC, J.-P., BERNILLON, J., RAHBE, Y., FEBVAY, G. et FAYARD, J.-M. Assessment of 35mer amino-modified oligonucleotide based microarray with bacterial samples. *Journal of Microbiological Methods*. 2004, vol. 57, n°2, pp. 207-218.

CAMON, E., MAGRANE, M., BARRELL, D., BINNS, D., FLEISCHMANN, W., KERSEY, P., MULDER, N., OINN, T., MASLEN, J., COX, A. et APWEILER, R. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Research*. 2003, vol. 13, n°4, pp. 662-672.

CAO, Z., WU, H., BRUCE, A., WOLLENBERG, K. et PANJWANI, N. Detection of differentially expressed genes in healing mouse corneas, using cDNA microarrays. *Investigative Ophthalmology and Visual Science*. 2002, vol. 43, n°9, pp. 2897-2904.

CASE GREEN, S., MIR, K., PRITCHARD, C. et SOUTHERN, E. Analysing genetic information with DNA arrays. *Current Opinion in Chemical Biology*. 1998, vol. 2, n°3, pp. 404-410.

CHAN, V., GRAVES, D. et MCKENZIE, S. The biophysics of DNA hybridization with immobilized oligonucleotide probes. *Biophysical Journal*. 1995, vol. 69, pp. 2243-2255.

CHANG, P.-C. et PECK, K. Design and assessment of a fast algorithm for identifying specific probes for human and mouse genes. *Bioinformatics*. 2003, vol. 19, n°11, pp. 1311-1317.

CHARLES, H. et ISHIKAWA, H. Physical and genetic map of the genome of *Buchnera*, the primary endosymbiont of the pea aphid *Acyrthosiphon pisum*. *Journal of Molecular Evolution*. 1999, vol. 48, pp. 142-150.

CHARLES, H., MOUCHIROUD, D., LOBRY, J., GONCALVES, I. et RAHBE, Y. Gene size reduction in the bacterial aphid endosymbiont, *Buchnera*. *Molecular Biology and Evolution*. 1999, vol. 16, n°12, pp. 1820-1822.

CHEN, D., MONTLLOR, C. et PURCELL, A. Fitness effects of two facultative endosymbiotic bacteria on the pea aphid, *Acyrthosiphon pisum*, and the blue alfalfa aphid, *A. kondoi*. *Entomologia experimentalis et applicata*. 2000, vol. 95, n°3, pp. 315-323.

CHEN, H. et SHARP, B. Oliz, a suite of Perl scripts that assist in the design of microarrays using 50mer oligonucleotides from 3' untranslated region. *BMC Bioinformatics*. 2002, vol. 3, n°1, pp. 27.

CHEN, H. et ZHU, G. Computer program for calculation the melting temperature of de generate oligonucleotides used in PCR. *Biotechniques*. 1997, vol. 22, pp. 1158-1160.

CHEN, Y., DOUGHERTY, R. et BITTNER, M. Ratio-based decisions and the quantitative analysis of cDNA microarrays images. *Journal of Biomedical Optics*. 1997, vol. 2, pp. 364-374.

CHESTER, N. et MARSHAK, D. Dimethyl sulfoxide-mediated primer Tm reduction: a method for analyzing the role of renaturation temperature in the polymerase chain reaction. *Analytical Biochemistry*. 1993, vol. 209, pp. 284-290.

CHO, J.-C. et TIEDJE, J. Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays. *Applied and Environmental Microbiology*. 2001, vol. 67, n°8, pp. 3677-3682.

CHU, S., DERISI, J., EISEN, M., MULHOLLAND, J., BOTSTEIN, D., BROWN, P. O. et HERSKOWITZ, I. The transcriptional program of sporulation in budding yeast. *Science*. 1998, vol. 282, n°5389, pp. 699-705.

CHUAQUI, R., BONNER, R., BEST, C., GILLESPIE, J., FLAIG, M., HEWITT, S., PHILLIPS, J., KRIZMAN, D., TANGREA, M., AHRAM, M., LINEHAM, W., KNEZEVIC, V. et EMMERT-BUCK, M. Post-analysis follow-up and validation of microarray experiments. *Nature Genetics*. 2002, vol. 32, n°supplement, pp. 509-514.

CHUDIN, E., WALKER, R., KOSAKA, A., WU, S. X., RABERT, D., CHANG, T. et KREDER, D. Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biology*. 2002, vol. 3, n°1, pp. research0005.

CHURCHILL, G. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics*. 2002, vol. 32, n°supplement, pp. 490-495.

CLARK, M., MORAN, N. et BAUMANN, P. Sequence evolution in bacterial endosymbionts having extreme base compositions. *Molecular Biology and Evolution*. 1999, vol. 16, n°11, pp. 1586-1598.

CLARK, M. A., MORAN, N. A., BAUMANN, P. et WERNEGREEN, J. J. Cospeciation between bacterial endosymbionts (*Buchnera*) and a recent radiation of aphids (*Uroleucon*) and pitfalls of testing for phylogenetic congruence. *Evolution*. 2000, vol. 54, n°2, pp. 517-525.

CLEVELAND, W. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*. 1979, vol. 74, n°368, pp. 829-836.

COLANTUONI, C., HENRY, G., ZEGER, S. et PEVSNER, J. Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. *Biotechniques*. 2002, vol. 32, n°6, pp. 1316-1320.

CORBIN, R., PALIY, O., YANG, F., SHABANOWITZ, J., PLATT, M., LYONS, C., ROOT, K., MCAULIFFE, J., JORDAN, M., KUSTU, S., SOUPENE, E. et HUNT, D. Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. *Proceedings of the National Academy of Sciences of the USA*. 2003, vol. 100, n°16, pp. 9232-9237.

COVACCI, A., KENNEDY, G., CORMACK, B., RAPPUOLI, R. et FALKOW, S. From microbial genomics to meta-genomics. *Drug Development Research*. 1997, vol. 41, pp. 180-192.

CROZIER, R. et CROZIER, Y. The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organisation. *Genetics*. 1993, vol. 113, pp. 97-117.

CUI, X. et CHURCHILL, G. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*. 2003, vol. 4, n°4, pp. 210.

CUI, X., HWANG, J., QIU, J., BLADES, N. et CHURCHILL, G. Improved statistical tests for differential gene expression by shrinking variance components [en ligne]. 2004. Disponible sur: http://www.jax.org/staff/churchill/labsite/pubs/index.html.

CUI, X., KERR, K. et CHURCHILL, G. Data transformation for cDNA microarray data [en ligne]. 2002. Disponible sur: http://www.jax.org/staff/churchill/labsite/pubs/index.html.

CULHANE, A., PERRIERE, G., CONSIDINE, E., COTTER, T. et HIGGINS, D. Between-group analysis of microarray data. *Bioinformatics*. 2002, vol. 18, n°12, pp. 1600-1608.

CUMMINGS, C. et RELMAN, D. Using DNA microarrays to study host-microbe interactions. *Genomics*. 2000, vol. 6, n°5, pp. 513-525.

D'HAESELEER, P., LIANG, S. et SOMOGYI, R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*. 2000, vol. 16, n°8, pp. 707-726.

DADD, R. Insect nutrition: current developments and metabolic implications. *Annual Review of Entomology*. 1973, vol. 18, pp. 381-421.

DARBY, A., BIRKLE, L., TURNER, S. et DOUGLAS, A. An aphid-borne bacterium allied to the secondary symbionts of whitefly. *FEMS Microbiology Ecology*. 2001, vol. 36, n°1, pp. 43-50.

DAREL, F. et KEPES, F. *Bioinformatique, génomique et post-génomique*. École Polytechnique, 2002.

DE BARY, H. De la symbiose. *Revue Internationale des Sciences*. 1879, vol. 3, pp. 301-309.

DE JONG, H. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*. 2002, vol. 9, pp. 67-103.

DE SAIZIEU, A., CERTA, U., WARRINGTON, J., GRAY, C., KECK, W. et MOUS, J. Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays. *Nature Biotechnology*. 1998, vol. 16, n°1, pp. 45-48.

DE WILDT, R., MUNDY, C., GORICK, B. et TOMLINSON, I. Antibody arrays for highthroughput screening of antibody-antigen interaction. *Nature Biotechnology*. 2000, vol. 18, pp. 989-984.

DERISI, J., IYER, V. et BROWN, P. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*. 1997, vol. 278, n°5338, pp. 680-686.

DERISI, J., PENLAND, L., BROWN, P., BITTNER, M., MELTZER, P., RAY, M., CHEN, Y., SU, Y. et TRENT, J. Use of cDNA microarrays gene expression patterns in human cancer. *Nature Genetics*. 1996, vol. 14, n°4, pp. 457-460.

DHAR, P., MENG, T., SOMANI, S., YE, L., SAIRAM, A., CHITRE, M., HAO, Z. et SAKHARKAR, K. Cellware-a multi-algorithmic software for computational systems biology. *Bioinformatics*. 2004, vol. 20, n°8, pp. 1319-1321.

DIEHL, F., GRAHLMANN, S., BEIER, M. et HOHEISEL, J. Manufacturing DNA microarrays of high spot homogeneity and reduced background signal. *Nucleic Acids Research*. 2001, vol. 29, n°7, pp. e38.

DIRKS, R., LIN, M., WINFREE, E. et PIERCE, N. Paradigms for computational nucleic acid design. *Nucleic Acids Research*. 2004, vol. 32, n°4, pp. 1392-1403.

DITTMAR, K., MOBLEY, E., RADEK, A. et PAN, T. Exploring the regulation of tRNA distribution on the genomic scale. *Journal of Molecular Biology*. 2004, vol. 337, n°1, pp. 31-47.

DOKTYCZ, M., MORRIS, M., DORMADY, S., BEATTIE, K. et JACOBSON, K. Optical melting of 128 octamer DNA duplexes. *The Journal of Biological Chemistry*. 1995, vol. 270, n°15, pp. 8439-8445.

DONIGER, S., SALOMONIS, N., DAHLQUIST, K., VRANIZAN, K., LAWLOR, S. et CONKLIN, B. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biology*. 2003, vol. 4, n°1, pp. R7.

DOUGLAS, A. Requirement of pea aphids (*Acyrthosiphon pisum*) for their symbiotic bacteria. *Entomologia Experimentalis et Applicata*. 1992, vol. 65, n°2, pp. 195-198.

DOUGLAS, A. Nutritional interactions in insect-microbial symbioses: Aphids and their symbiotic bacteria *Buchnera*. *Annual Review of Entomology*. 1998, vol. 43, pp. 17-37.

DOUGLAS, A. The nutritional physiology of aphids. *Advances in Insect Physiology*. 2003, vol. 31, pp. 73-140.

DOUGLAS, A. et DIXON, A. The mycetocyte symbiosis of aphids: variation with age and morph in virginoparae of *Megoura viciae* and *Acyrthosiphon pisum*. *Journal of Insect Physiology*. 1987, vol. 33, n°2, pp. 109-113.

DOUGLAS, A., MINTO, L. et WILKINSON, T. Quantifying nutrient production by the microbial symbionts in an aphid. *Journal of Experimental Biology*. 2001, vol. 204, n°2, pp. 349-358.

DRUMMOND, M. et STAMPER, J. DNAPROBE, a computer program which generates oligonucleotides probes from protein alignments. *Nucleic Acids Research*. 1999, vol. 27, n°17, pp. 3493.

DUBAQUIE, Y., LOOSER, R., FUNFSCHILLING, U., JENO, P. et ROSPERT, S. Identification of in vivo substrates of the yeast mitochondrial chaperonins reveals overlapping but non-identical requirement for hsp60 and hsp10. *Embo Journal*. 1998, vol. 17, n°20, pp. 5868-5876.

DUDLEY, A., AACH, J., STEFFEN, M. et CHURCH, G. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proceedings of the National Academy of Sciences of the USA*. 2002, vol. 99, n°11, pp. 7554-7559.

DUDOIT, S., YANG, Y., SPEED, T. et CALLOW, M. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*. 2002, vol. 12, n°1, pp. 111-139.

DUGAS, V., CHEVALIER, Y., DEPRET, G., NESME, X. et SOUTEYRAND, E. The immobilisation of DNA strands on silica surface by means of chemical grafting. *Progress in Colloid and Polymer Science*. 2004, vol. 123, pp. 275-279.

DURBIN, B., HARDIN, J., HAWKINS, D. et ROCKE, D. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*. 2002, vol. 18, n°supplement 1, pp. 105-110.

EDWARDS, D. Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics*. 2003, vol. 19, n°7, pp. 825-833.

EFRON, B. et TIBSHIRANI, R. Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiololgy*. 2002, vol. 23, n°1, pp. 70-86.

EICKHOFF, B., KORN, B., SCHICK, M., POUSTKA, A. et VAN DER BOSCH, J. Normalization of array hybridization experiments in differential gene expression analysis. *Nucleic Acids Research*. 1999, vol. 27, n°22, pp. e33.

EISEN, J. et HANAWALT, P. A phylogenomic study of DNA repair genes, proteins, and processes. *Mutation Research*. 1999, vol. 435, n°3, pp. 171-213.

EISEN, M., SPELLMAN, P., BROWN, P. et BOTSTEIN, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the USA*. 1998, vol. 95, pp. 14863-14868.

EMORY, S., BOUVET, P. et BELASCO, J. A 5'-terminal stem-loop structure can stabilize mRNA in *Escherichia coli*. *Genes Development*. 1992, vol. 6, n°1, pp. 135-148.

EMRICH, S., LOWE, M. et DELCHER, A. PROBEmer: a web-based software tool for selecting optimal DNA oligos. *Nucleic Acids Research*. 2003, vol. 31, n°13, pp. 3746-3750.

FARES, M., BARRIO, E., SABATER-MUNOZ, B. et MOYA, A. The evolution of the heat-shock protein *GroEL* from *Buchnera*, the primary endosymbiont of aphids, is governed by positive selection. *Molecular Biology Evolution*. 2002a, vol. 19, n°7, pp. 1162-1170.

FARES, M., RUIZ-GONZALEZ, M., MOYA, A., ELENA, S. et BARRIO, E. Endosymbiotic bacteria: *GroEL* buffers against deleterious mutations. *Nature*. 2002b, vol. 417, n°6887, pp. 398.

FEBVAY, G., DELOBEL, B. et RAHBÉ, Y. Influence of the amino acid balance on the improvement of an artificial diet for a biotype of *Acyrthosiphon pisum* (Homoptera: Aphididae). *Canadian Journal of Zoology*. 1988, vol. 66, pp. 2449-2453.

FEBVAY, G., LIADOUZE, I., GUILLAUD, J. et BONNOT, G. Analysis of energetic aminoacid metabolism in *Acyrthosiphon pisum*: a multidimensional approach to amino acid metabolism in aphids. *Archives of Insect Biochemistry and Physiology*. 1995, vol. 29, pp. 45-69.

FEBVAY, G., RAHBÉ, Y., RYNKIEWICZ, M., GUILLAUD, J. et BONNOT, G. Fate of dietary saccharose and neosynthesis of amino acids in the pea aphid, *Acyrthosiphon pisum*, reared on different diets. *Journal of Experimental Biology*. 1999, vol. 202, n°19, pp. 2639-2652.

FERNANDES, R. et SKIENA, S. Microarray synthesis through multiple-use PCR primer design. *Bioinformatics*. 2002, vol. 18, n°supplement 1, pp. S128-S135.

FEUERSTEIN, B., WILLIAMS, L., BASU, H. et MARTON, L. Implications and concepts of polyamine-nucleic acid interactions. *Journal of Cellular Biochemistry*. 1991, vol. 46, n°1, pp. 37-47.

FIELDEN, M., HALGREN, R., DERE, E. et ZACHAREWSKI, T. GP3: GenePix post-processing program for automated analysis of raw microarray data. *Bioinformatics*. 2002, vol. 18, n°5, pp. 771-773.

FINKELSTEIN, D., GOLLUB, J. et CHERRY, J. Normalization and systematic measurement error in cDNA microarray data [en ligne]. 2002. Disponible sur: http://afgc.stanford.edu/afgc_html/site2Stat.htm.

FISHER, R. *The design of experiments*. 6th edition. London: Oliver and Boyd, 1951.

FLEISCHMANN, R., ADAMS, M., WHITE, O., CLAYTON, R., KIRKNESS, E., KERLAVAGE, A., BULT, C., TOMB, J., DOUGHERTY, B. et MERRICK, J. Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. *Science*. 1995, vol. 269, n°5223, pp. 496-512.

FRASER, C., GOCAYNE, J., WHITE, O., ADAMS, M., CLAYTON, R., FLEISCHMANN, R., BULT, C., KERLAVAGE, A., SUTTON, G., KELLEY, J. et ET AL. The minimal gene complement of *Mycoplasma genitalium*. *Science*. 1995, vol. 270, n°5235, pp. 397-403.

FRANK, A. Über die biologischen Verhaltnisse des Thallus einiger Krusten-Flechten. *Beitrage zur Biologie des Planzen*. 1877, vol. 2, pp. 193.

FREEMAN, W., ROBERTSON, D. et VRANA, K. Fundamentals of DNA hybridization arrays for gene expression analysis. *Biotechniques*. 2000, vol. 29, n°5, pp. 1042-1046.

FREIER, S., KIERZEK, R., JAEGER, J., SUGIMOTO, N., CARUTHERS, M., NEILSON, T. et TURNER, D. Improved free-energy parameters for predictions of RNA duplex stability. *Proceedings of the National Academy of Sciences of the USA*. 1986, vol. 83, pp. 9373-9377.

FRIEDMAN, N., LINIAL, M., NACHMAN, I. et PE'ER, D. Using Bayesian networks to analyze expression data. *Journal of Computional Biology*. 2000, vol. 7, n°3-4, pp. 601-620.

FUKATSU, T. et ISHIKAWA, H. Phylogenetic position of yeast-like symbiont of *Hamiltonaphis styraci* (Homoptera, Aphididae) based on 18S rDNA sequence. *Insect Biochemistry and Molecular Biology*. 1996, vol. 26, n°4, pp. 383-388.

FUKATSU, T., NIKOH, N., KAWAI, R. et KOGA, R. The secondary endosymbiotic bacterium of the pea aphid *Acyrthosiphon pisum* (Insecta, Homoptera). *Applied and environmental Microbiology*. 2000, vol. 66, n°7, pp. 2748-2758.

FUTCHER, B., LATTER, G., MONARDO, P., MCLAUGHLIN, C. et GARRELS, J. A sampling of the yeast proteome. *Molecular and Cellular Biology*. 1999, vol. 19, n°11, pp. 7357-7368.

GELLMAN, S., HAQUE, T. et NEWCOMB, L. New evidence that the hydrophobic effect and dispersion are not major driving forces for nucleotide base stacking. *Biophysical Journal*. 1996, vol. 71, pp. 3523-3526.

GENTLEMAN, R., CAREY, V., BATES, D., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y., GENTRY, J., HORNIK, K., HOTHORN, T., HUBER, W., IACUS, S., IRIZARRY, R., LEISCH, F., LI, C., MAECHLER, M., ROSSINI, A. J., SAWITZKI, G., SMITH, C., SMYTH, G., TIERNEY, L., YANG, J. et ZHANG, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*. 2004, vol. 5, n°10, pp. R80.

GERHOLD, D., RUSHMORE, T. et CASKEY, C. DNA chips: promising toys have become powerful tools. *Trends in Biochemical Sciences*. 1999, vol. 24, pp. 168-173.

GERSON, D. Microarray technology: an array of opportunities. *Nature*. 2002, vol. 416, pp. 885-891.

GETZ, G., LEVINE, E. et DOMANY, E. Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences of the USA*. 2000, vol. 97, n°22, pp. 12079-12084.

GHOSH, D. et CHINNAIYAN, A. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*. 2002, vol. 18, n°2, pp. 275-286.

GIBBS, W. La biologie virtuelle. *Pour la Science*. 2001, vol. 288, pp. 38-44.

GIBSON, G. Microarrays in ecology and evolution: a preview. *Molecular Ecology*. 2002, vol. 11, n°1, pp. 17-24.

GIGLIO-TOS, E. Les phénomènes de la vie. Cagliari, 1903.

GIL, R., SABATER-MUÑOZ, B., LATORRE, A., SILVA, F. et MOYA, A. Extreme genome reduction in *Buchnera spp.*: Toward the minimal genome needed for symbiotic life. *Proceedings of the National Academy of Sciences of the USA*. 2002, vol. 99, n°7, pp. 4454-4458.

GINGERAS, T. et ROSENOW, C. Studying microbial genomes with high-density oligonucleotide arrays. *A.S.M. News*. 2000, vol. 66, n°8, pp. 463-469.

GOEBEL, W. et GROSS, R. Intracellular survival strategies of mutualistic and parasitic prokaryotes. *Trends in Microbiology*. 2001, vol. 9, n°6, pp. 267-273.

GOLFIER, G., DANG, M., DAUPHINOT, L., GRAISON, E., ROSSIER, J. et POTIER, M.-C. VARAN: a web server for variability analysis of DNA microarray experiments. *Bioinformatics*. 2004, vol. 20, n°10, pp. 1641-1643.

GOLUB, T., SLONIM, D., TAMAYO, P., HUARD, C., GASSENBEEK, M., MESIROV, J., COLLER, H., LOH, M., DOWNING, J., CALIGIURI, M., BLOOMFIELD, C. et LANDER, E. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999, vol. 286, pp. 531-537.

GOODSELL, D. Inside a living cell. *Trends in Biochemical Sciences*. 1991, vol. 16, n°6, pp. 203-206.

GORYACHEV, A. et MAC GREGOR, P. Unfolding of microarray data. *Journal of Computational Biology*. 2001, vol. 8, n°4, pp. 443-461.

GOUY, M. et GAUTIER, C. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research*. 1982, vol. 10, n°22, pp. 7055-7074.

GRANJEAUD, S., BERTUCCI, F. et JORDAN, B. Expression profiling: DNA arrays in many guises. *BioEssays*. 1999, vol. 21, n°9, pp. 781-790.

GREENBAUM, D., JANSEN, R. et GERSTEIN, M. Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics*. 2002, vol. 18, n°4, pp. 585-596.

GRIFFIN, T. et SMITH, L. An approach to predicting the stabilities of peptide nucleic acid: DNA duplexes. *Analytical Biochemistry*. 1998, vol. 260, pp. 56-63.

GRIFFITHS, G. et BECK, S. Intracellular symbiotes of the pea aphid, *Acyrthosiphon pisum*. *Journal of Insect Physiology*. 1973, vol. 19, pp. 75-84.

GROEBE, D. et UHLENBECK, O. Characterization of RNA hairpin loop stability. *Nucleic Acids Research*. 1988, vol. 16, n°24, pp. 11725-11735.

GROSU, P., TOWNSEND, J., HARTL, D. L. et CAVALIERI, D. Pathway Processor: a tool for integrating whole-genome expression results into metabolic networks. *Genome Research*. 2002, vol. 12, n°7, pp. 1121-1126.

GUPTA, V., CHERKASSKY, A., CHATIS, P., JOSEPH, R., JOHNSON, A., BROADBENT, J., ERICKSON, T. et DIMEO, J. Directly labeled

mRNA produces highly precise and unbiased differential gene expression data. *Nucleic Acids Research*. 2003, vol. 31, n°4, pp. e13.

GUSCHIN, D., MOBARRY, B., PROUDNIKOV, D., STAHL, D., RITTMAN, B. et MIRZABEKOV, A. Oligonucleotide microchips as genosensors for determinative studies in microbiology. *Applied and Environnemental Microbiology*. 1997, vol. 63, pp. 2397-2402.

GYGI, S., ROCHON, Y., FRANZA, B. et AEBERSOLD, R. Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology*. 1999, vol. 19, n°3, pp. 1720-1730.

HAAS, S., HILD, M., WRIGHT, A., HAIN, T., TALIBI, D. et VINGRON, M. Genome-scale design of PCR primers and long oligomers for DNA microarrays. *Nucleic Acids Research*. 2003, vol. 31, n°19, pp. 5576-5581.

HACIA, J. Resequencing and mutational analysis using oligonucleotide microarrays. *Nature Genetics*. 1999, vol. 21, n°1, pp. 42-47.

HALL, D., ZHU, H., ZHU, X., ROYCE, T., GERSTEIN, M. et SNYDER, M. Regulation of gene expression by a metabolic enzyme. *Science*. 2004, vol. 306, n°5695, pp. 482-484.

HANCOCK, J. et ARMSTRONG, J. SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. *Computer Applications in Biosciences*. 1994, vol. 10, n°1, pp. 67-70.

HARRINGTON, C., ROSENOW, C. et RETIEF, J. Monitoring gene expression using DNA microarrays. *Current Opinion in Microbiology*. 2000, vol. 3, n°3, pp. 285-291.

HAUTEFORT, I. et HINTON, J. Measurement of bacterial gene expression *in vivo*. *Philosophical Transactions of the Royal Society of London Series B Biological Sciences*. 2000, vol. 355, n°1397, pp. 601-611.

HAYNES, S., DARBRY, A., DANIELL, T., WEBSTER, G., VAN VEEN, F., GODFRAY, H., PROSSER, J. et DOUGLAS, A. Diversity of bacteria associated with natural aphid populations. *Applied and Environmental Microbiology*. 2003, vol. 69, n°12, pp. 7216-7223.

HECKER, M. et ENGELMANN, S. Proteomics, DNA arrays and the analysis of still unknown regulons and unknown proteins of *Bacillus subtilis* and pathogenic Gram-positive bacteria. *International Journal of Medical Microbiology*. 2000, vol. 290, n°2, pp. 123-134.

HEDGE, P., QI, R., ABERNATHY, K., GAY, C., DHARAP, S., GASPARD, R., HUGHES, J., SNESRUD, E., LEE, N. et QUACKENBUSH, J. A concise guide to cDNA microarray analysis. *Biotechniques*. 2000, vol. 29, n°3, pp. 548-562.

HEKSTRA, D., TAUSSIG, A., MAGNASCO, M. et NAEF, F. Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Research*. 2003, vol. 31, n°7, pp. 1962-1968.

HELD, G., GRINSTEIN, G. et TU, Y. Modeling of DNA microarray data by using physical properties of hybridization. *Proceedings of the National Academy of Sciences of the USA*. 2000, vol. 100, n°13, pp. 7575-7580.

HELLER, M. DNA microarray technology: devices, systems, and applications. *Annual Review of Biomedical Engineering*. 2002, vol. 4, pp. 129-153.

HELMANN, J., WU, M., KOBEL, P., GAMO, F., WILSON, M., MORSHEDI, M., NAVRE, M. et PADDON, C. Global transcriptional res-

ponse of *Bacillus subtilis* to heat shock. *Journal of Bacteriology*. 2001, vol. 183, n°24, pp. 7318-7328.

HENDRIX, D. et SALVUCCI, M. Polyol metabolism in homopterans at high temperatures: accumulation of mannitol in aphids (Aphididae: Homoptera) and sorbitol in whiteflies (Aleyrodidae: Homoptera). *Comparative Biochemistry and Physiology*. 1998, vol. 120, pp. 487-494.

HESSNER, M., WANG, X., KHAN, S., MEYER, L., SCHLICHT, M., TACKES, J., DATTA, M., JACOB, H. et GHOSH, S. Use of a three-color cDNA microarray platform to measure and control support-bound probe for improved data quality and reproducibility. *Nucleic Acids Research*. 2003, vol. 31, n°11, pp. e60.

HIGGINS, C., DORMAN, C., STIRLING, D., WADDELL, L., BOOTH, I., MAY, G. et BREMER, E. A physiological role for DNA supercoiling in the osmotic regulation of gene expression in *S. typhimurium* and *E. coli*. *Cell*. 1988, vol. 52, n°4, pp. 569-584.

HINDE, R. The control of the mycetome symbiotes of the aphid *Brevicoryne brassicae*, *Myzus persicae* and *Macrosiphum rosae*. *Journal of Insect Physiology*. 1971a, vol. 17, pp. 1791-1800.

HINDE, R. Maintenance of aphid cells and the intracellular symbiotes of aphids *in vitro*. *Journal of Invertebrate Pathology*. 1971b, vol. 17, n°3, pp. 333-338.

HINFRAY, J. Les puces à ADN. *Biofutur*. 1997, vol. 166, pp. 1-15.

HIROTSUNE, S., YOSHIDA, N., CHEN, A., GARRETT, L., SUGIYAMA, F., TAKAHASHI, S., YAGAMI, K., WYNSHAW-BORIS, A. et YOSHIKI, A. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature*. 2003, vol. 423, n°6935, pp. 91-96.

HOHEISEL, J. Oligomer-chip technology. *Trends in Biotechnology*. 1997, vol. 15, pp. 465-469.

HOHEISEL, J. et VINGRON, M. Transcriptional profiling: is it worth the money ? *Research in Microbiology*. 2000, vol. 151, pp. 113-119.

HOLLOWAY, A., VAN LAAR, K., TOTHILL, R. et BOWTELL, D. Options available-from start to finish-for obtaining data from DNA microarrays II. *Nature Genetics*. 2002, vol. 32, n°supplement, pp. 481-489.

HOLTER, N., MITRA, M., MARITAN, A., CIEPLAK, M., BANAVAR, J. et FEDOROFF, N. Fundamental patterns underlying gene expression profiles: simplicity from. *Proceedings of the National Academy of Sciences of the USA*. 2000, vol. 97, n°15, pp. 8409-8414.

HOUK, E. et GRIFFITHS, G. Intracellular symbiotes of the homoptera. *Annual Review of Entomology*. 1980, vol. 25, pp. 161-187.

HUBER, W., VON HEYDEBRECK, A., SULTMANN, H., POUSTKA, A. et VINGRON, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*. 2002, vol. 18, n°supplement 1, pp. 96-104.

HUGHES, D. Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. *Genome Biology*. 2000, vol. 1, n°6, pp. 61-68.

HUGHES, T., MAO, M., JONES, A., BURCHARD, J., MARTON, M., SHANNON, K., LEFKOWITZ, S., ZIMAN, M., SCHELTER, J., MEYER, M., KOBAYASHI, S., DAVIS, C., DAI, H., HE, Y., STEPHANIANTS, S., CAVET, G., WALKER, W., WEST, A., COFFEY, E., SHOEMAKER, D., STOUGHTON, R., BLANCHARD, A., FRIEND, S. et LINSLEY, P. Expression profiling using microarrays

fabricated by an ink-jet oligonucleotide synthetizer. *Nature Biotechnology*. 2001, vol. 19, pp. 342-347.

HUGHES, T., MARTON, M., JONES, A., ROBERTS, C., STOUGHTON, R., ARMOUR, C., BENNETT, H., COFFEY, E., DAI, H., HE, Y., KIDD, M., KING, A., MEYER, M., SLADE, D., LUM, P., STEPANIANTS, S., SHOEMAKER, D., GACHOTTE, D., CHAKRABURTTY, K., SIMON, J., BARD, M. et FRIEND, S. Functional discovery via a compendium of expression profiles. *Cell*. 2000, vol. 102, n°1, pp. 109-126.

HUMPHREYS, N. et DOUGLAS, A. 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. *Applied and Environmental Microbiology*. 1997, vol. 63, n°8, pp. 3294-3296.

HUNG, S., BALDI, P. et HATFIELD, G. Global gene expression profiling in *Escherichia coli* K12. The effects of leucine-responsive regulatory protein. *The Journal of Biological Chemistry*. 2002, vol. 277, n°43, pp. 40309-40323.

HUSMEIER, D. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*. 2003, vol. 19, n°17, pp. 2271-2282.

HYNDMAN, D., COOPER, A., PRUZINSKY, D., COAD, D. et MITSUHASHI, M. Software to determine optimal oligonucleotide sequences based on hybridization simulation data. *Biotechniques*. 1996, vol. 20, pp. 1090-1097.

IDEKER, T., GALITSKI, T. et HOOD, L. A new approach to decoding life: systems biology. *Annual Review of Genomics and Human Genetics*. 2001a, vol. 2, pp. 343-372.

IDEKER, T., THORSSON, V., RANISH, J. A., CHRISTMAS, R., BUHLER, J., ENG, J. K., BUMGARNER, R., GOODLETT, D. R., AEBERSOLD, R. et HOOD, L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*. 2001b, vol. 292, n°5518, pp. 929-934.

ILIOPOULOS, I., TSOKA, S., ANDRADE, M., ENRIGHT, A., CARROLL, M., POULLET, P., PROMPONAS, V., LIAKOPOULOS, T., PALAIOS, G., PASQUIER, C., HAMODRAKAS, S., TAMAMES, J., YAGNIK, A. T., TRAMONTANO, A., DEVOS, D., BLASCHKE, C., VALENCIA, A., BRETT, D., MARTIN, D., LEROY, C., RIGOUTSOS, I., SANDER, C. et OUZOUNIS, C. Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics*. 2003, vol. 19, n°6, pp. 717-726.

ISCOVE, N., BARBARA, M., GU, M., GIBSON, M., MODI, C. et WINEGARDEN, N. Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. *Nature Biotechnology*. 2002, vol. 20, n°9, pp. 940-943.

ISHIKAWA, H. Characterization of the protein species synthesized *in vivo* and *in vitro* by an aphid endosymbiont. *Insect Biochemistry*. 1984, vol. 14, n°4, pp. 417-425.

IYER, V., HORAK, C., SCAFE, C., BOTSTEIN, D., SNYDER, M. et BROWN, P. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*. 2001, vol. 409, pp. 533-538.

JACOB, A., BRANDT, O., STEPHAN, A. et HOHEISEL, J. Peptide nucleic acid microarrays. *Methods in Molecular Biology*. 2004, vol. 283, pp. 283-294.

JACOBS, J. et FODOR, S. Combinatorial chemistry application of light-directed chemical synthesis. *Trends in Biotechnology*. 1994, vol. 12, n°1, pp. 19-26.

JIMENEZ, J., MITCHELL, M. et SGOUROS, J. Microarray analysis of orthologous genes: conservation of the translational machinery across species at the sequence and expression level. *Genome Biology*. 2002, vol. 4, n°1, pp. R4.

JIMENEZ, N., GONZALEZ-CANDELAS, F. et SILVA, F. Prephenate dehydratase from the aphid endosymbiont (*Buchnera*) displays changes in the regulatory domain that suggest its desensitization to inhibition by phenylalanine. *Journal of Bacteriology*. 2000, vol. 182, n°10, pp. 2967-2969.

JIN, W., RILEY, R., WOLFINGER, R., WHITE, K., PASSADOR-GURGEL, G. et GIBSON, G. The contributions of sex, genotype and age to transcritional variance in *Drosophila melanogaster*. *Nature Genetics*. 2001, vol. 29, pp. 389-395.

JOHNSON, S., DOUGLAS, A., WOODWARD, S. et HARYLEY, S. Microbial impacts on plant-herbivore interactions: the indirect effects of a birch pathogen on a birch aphid. *Oecologia*. 2003, vol. 134, pp. 388-396.

KADERALI, L. et SCHLIEP, A. Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics*. 2002, vol. 18, n°10, pp. 1340-1349.

KALLIONIEMI, A., KALLIONIEMI, O., SUDAR, D., RUTOVITZ, D., GRAY, J., WALDMAN, F. et PINKEL, D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*. 1992, vol. 258, n°5083, pp. 818-821.

KÄMPKE, T., KIENINGER, M. et MECKLENBURG, M. Efficient primer design algorithms. *Bioinformatics*. 2001, vol. 17, n°3, pp. 214-225.

KANE, M., JATKOE, T., STUMPF, C., LU, J., THOMAS, J. et MADORE, S. Assessment of the sensitivity and specificity of oligonucleotide (50 mer) microarrays. *Nucleic Acids Research*. 2000, vol. 28, n°22, pp. 4552-4557.

KANEHISA, M., GOTO, S., KAWASHIMA, S. et NAKAYA, A. The KEGG databases at GenomeNet. *Nucleic Acids Research*. 2002, vol. 30, n°1, pp. 42-46.

KARLIN, S. et ALTSCHUL, S. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the USA*. 1990, vol. 87, n°6, pp. 2264-2268.

KARLIN, S. et ALTSCHUL, S. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Sciences of the USA*. 1993, vol. 90, n°12, pp. 5873-5877.

KARP, P., RILEY, M., SAIER, M., PAULSEN, I., PALEY, S. et PELLEGRINI-TOOLE, A. The EcoCyc and MetaCyc databases. *Nucleic Acids Research*. 2000, vol. 28, n°1, pp. 56-59.

KATO-MAEDA, M., GAO, Q. et SMALL, P. Microarray analysis of pathogens and their interaction with hosts. *Cellular Microbiology*. 2001, vol. 3, n°11, pp. 713-719.

KAUFMAN, B., NEWMAN, S., HALLBERG, R., SLAUGHTER, C., PERLMAN, P. et BUTOW, R. *In organello* formaldehyde crosslinking of proteins to mtDNA: identification of bifunctional proteins. *Proceedings of the National Academy of Sciences of the USA*. 2000, vol. 97, n°14, pp. 7772-7777.

KEPES, F. La simulation à l'ère de la génomique. *Biofutur*. 2001, vol. 210, pp. 46-51.

KEPES, F. Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites. *Journal of Molecular Biology*. 2003, vol. 329, n°5, pp. 859-865.

KEPES, F. Periodic transcriptional organization of the *E.coli* genome. *Journal of Molecular Biology*. 2004, vol. 340, n°5, pp. 957-964.

KERR, K., MARTIN, M. et CHURCHILL, G. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*. 2000, vol. 7, n°6, pp. 819-837.

KERR, M., AFSHARI, C., BENNETT, L., BUSHEL, P., MARTINEZ, J., WALKER, N. C. et CHURCHILL, G. Statistical Analysis of a Gene Expression Microarray Experiment with Replication. *Statistica Sinica*. 2002, vol. 12, pp. 203-217.

KERR, M. et CHURCHILL, G. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences of the USA*. 2001a, vol. 98, n°16, pp. 8961-8965.

KERR, M. et CHURCHILL, G. Statistical design and the analysis of gene expression microarray data. *Genetic Research*. 2001b, vol. 77, n°2, pp. 123-128.

KITANO, H. Computational systems biology. *Nature*. 2002a, vol. 420, pp. 206-210.

KITANO, H. Systems biology: a brief overview. *Science*. 2002b, vol. 295, n°5560, pp. 1662-1664.

KOGA, R., TSUCHIDA, T. et FUKATSU, T. Changing partners in an obligate symbiosis: a facultative endosymbiotnt can compensate for loss of the essential endosymbiont *Buchnera* in an aphid. *Proceedings of the Royal Society London B*. 2003, vol. 270, n°1533, pp. 2543-2550.

KOMAKI, K. et ISHIKAWA, H. Intracellular bacterial symbionts of aphids possess many genomic copies per bacterium. *Journal of Molecular Evolution*. 1999, vol. 48, n°6, pp. 717-722.

KOMAKI, K. et ISHIKAWA, H. Genomic copy number of intracellular bacterial symbionts of aphids varies in response to devlopmental stage and morph of their host. *Insect Biochemistry and Molecular Biology*. 2000, vol. 30, pp. 253-258.

KOOPERBERG, C., FAZZIO, T., DELROW, J. et TSUKIYAMA, T. Improved background correction for spotted DNA microarrays. *Journal of Computational Biology*. 2002, vol. 9, n°1, pp. 55-66.

KUBORI, T., MATSUSHIMA, Y., NAKAMURA, D., URALIL, J., LARA-TEJERO, M., SUKHAN, A., GALAN, J. E. et AIZAWA, S. I. Supramolecular structure of the *Salmonella typhimurium* type III protein secretion system. *Science*. 1998, vol. 280, n°5363, pp. 602-605.

KUNITSYN, A., KOCHETKOVA, S., TIMOFEEV, E. et FLORENTIEV, V. Partial thermodynamic parameters for prediction stability and washing behavior of DNA duplexes immobilized on gel matrix. *Journal of Biomolecular Structure and Dynamics*. 1996, vol. 14, n°2, pp. 239-244.

KUPIEC, J.-J. et SONIGO, P. *Ni Dieu ni gène. Pour une autre théorie de l'hérédité*. Seuil: Science ouverte, 2000.

KURLAND, C. et ANDERSSON, S. Origin and evolution of the mitochondrial proteome. *Microbiology and Molecular Biology Reviews*. 2000, vol. 64, n°4, pp. 786-820.

KUSANO, S., DING, Q., FUJITA, N. et ISHIHAMA, A. Promoter selectivity of *Escherichia coli* RNA polymerase E sigma 70 and E sigma 38 holoenzymes. Effect of DNA supercoiling. *Journal of Biological Chemistry*. 1996, vol. 271, n°4, pp. 1998-2004.

LAGODA, P. et REGAD, F. Les puces à ADN: outils pour une nouvelle vision de la diversité et des ressources génétiques. *Cahiers Agriculture*. 2000, vol. 9, pp. 329-340.

LANE, A., MARTIN, S., EBEL, S. et BROWN, T. Solution conformation of deoxynucleotide containing tandem G.A mismatched base pairs and 3'-overhanging ends in $d(GTGAACTT)_2$. *Biochemistry*. 1992, vol. 31, pp. 12087-12095.

LANG, F., PAQUIN, B. et BURGER, G. L'évolution moléculaire et la révolution génomique. *Médecine/Sciences*. 2000, vol. 16, pp. 212-218.

LE NOVERE, N. MELTING, computing the melting temperature of nucleic acid duplex. *Bioinformatics*. 2001, vol. 17, n°12, pp. 1226-1227.

LEE, M., KUO, F. et SKLAR, J. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridation. *Proceedings of the National Academy of Sciences of the USA*. 2000, vol. 97, n°18, pp. 9834-9839.

LEE, M. et WALT, D. A fiber-optic microarray biosensor using aptamers as receptors. *Analytical Biochemistry*. 2000, vol. 282, n°1, pp. 142-146.

LEGAY, J.-M. *L'expérience et le modèle. Un discours sur la méthode*. INRA. Paris, 1996.

LEHNINGER, A. *Biochimie seconde édition : bases moléculaires de la structure et des fonctions cellulaires*. Médecine-Sciences: Flammarion, 1977.

LEMIEUX, B., AHARONI, A. et SCHENA, M. Overview of DNA chip technology. *Molecular Breeding*. 1998, vol. 4, n°4, pp. 277-289.

LENNON, G. et LEHRACH, H. Hybridization analyses of arrayed cDNA libraries. *Trends in Genetics*. 1991, vol. 7, n°10, pp. 314-317.

LEUNG, Y. Unravelling the mystery of microarray data analysis. *Trends in Biotechnology*. 2002, vol. 20, n°9, pp. 366-368.

LEUNG, Y. et CAVALIERI, D. Fundamentals of cDNA microarray data analysis. *Trends in Genetics*. 2003, vol. 19, n°11, pp. 649-659.

LEWIS, S., ASHBURNER, M. et REESE, M. Annotating eukaryote genomes. *Current Opinion in Structural Biology*. 2000, vol. 10, n°3, pp. 349-354.

LEXA, M. et VALLE, G. PRIMEX: rapid identification of oligonucleotide matches in whole genomes. *Bioinformatics*. 2003, vol. 19, n°18, pp. 2486-2488.

LI, F. et STORMO, G. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*. 2001, vol. 17, n°11, pp. 1067-1076.

LIADOUZE, I., FEBVAY, G., GUILLAUD, J. et BONNOT, G. Effect of diet on the free amino acid pools of symbiotic and aposymbiotic pea aphids, *Acyrthosiphon pisum*. *Journal of Insect Physiology*. 1995, vol. 41, n°1, pp. 33-40.

LIADOUZE, I., FEBVAY, G., GUILLAUD, J. et BONNOT, G. Metabolic fate of energetic amino acids in the aposymbiotic pea aphid *Acyrthosi-*

phon pisum (Harris) (Homoptera: Aphididae). *Symbiosis*. 1996, vol. 21, pp. 115-127.

LIANG, P. et PARDEE, A. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*. 1992, vol. 257, n°5072, pp. 967-971.

LINDBLAD-TOH, K., TANENBAUM, D., DALY, M., WINCHESTER, E., LUI, W. O., VILLAPAKKAM, A., STANTON, S., LARSSON, C., HUDSON, T., JOHNSON, B., LANDER, E. et MEYERSON, M. Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nature Biotechnology*. 2000a, vol. 18, n°9, pp. 1001-1005.

LINDBLAD-TOH, K., WINCHESTER, E., DALY, M., WANG, D., HIRSCHHORN, J., LAVIOLETTE, J., ARDLIE, K., REICH, D., ROBINSON, E., SKLAR, P., SHAH, N., THOMAS, D., FAN, J., GINGERAS, T., WARRINGTON, J., PATIL, N., HUDSON, T. et LANDER, E. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genetics*. 2000b, vol. 24, n°4, pp. 381-386.

LIPSHUTZ, R., FODOR, S., GINGERAS, T. et LOCKHART, D. High density synthetic oligonucleotide arrays. *Nature Genetics*. 1999, vol. 21, pp. 20-24.

LIVACHE, T., ROGET, A., DEJEAN, E., BARTHET, C., BIDAN, G. et TEOULE, R. Preparation of a DNA matix via an electrochemically directed copolymerisation of pyrrole and oligonucleotides bearing a pyrrole group. *Nucleic Acids Research*. 1994, vol. 22, pp. 2915-2921.

LO, A., MAGLIANO, D., SIBSON, M., KALITSIS, P., CRAIG, J. et CHOO, K. A novel chromatin immunoprecipitation and array (CIA) analysis identifies a 460-kb CENP-A-binding neocentromere DNA. *Genome Research*. 2001, vol. 11, n°3, pp. 448-457.

LOBRY, J. et GAUTIER, C. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Research*. 1994, vol. 22, n°15, pp. 3174-3180.

LOCKHART, D., DONG, H., BYRNE, M., FOLLETTIE, M., GALLO, M., CHEE, M., MITTMANN, M., WANG, C., KOBAYASHI, M., HORTON, H. et BROWN, E. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*. 1996, vol. 14, n°13, pp. 1675-1680.

LOJKOWSKA, E., MASCLAUX, C., BOCCARA, M., ROBERT-BAUDOUY, J. et HUGOUVIEUX-COTTE-PATTAT, N. Characterisation of the *pelL* gene encoding a novel pectate lyase of *Erwinia chrysantemi* 3937. *Molecular Microbiology*. 1995, vol. 16, n°6, pp. 1183-1195.

LONNSTEDT, I. et SPEED, T. Replicated microarray data. *Statistica Sinica*. 2002, vol. 12, pp. 31-46.

LUDWIG, W., STRUNK, O., WESTRAM, R., RICHTER, L., MEIER, H., YADHUKUMAR, BUCHNER, A., LAI, T., STEPPI, S., JOBB, G., FORSTER, W., BRETTSKE, I., GERBER, S., GINHART, A., GROSS, O., GRUMANN, S., HERMANN, S., JOST, R., KONIG, A., LISS, T., *et al.* ARB: a software environment for sequence data. *Nucleic Acids Research*. 2004, vol. 32, n°4, pp. 1363-1371.

LUEBKE, K., BALOG, R. et GARNER, H. Prioritized selection of oligodeoxyribonucleotide probes for efficient hybridization to RNA transcripts. *Nucleic Acids Research*. 2003, vol. 31, n°2, pp. 750-758.

MACBEATH, G. Protein microarrays and proteomics. *Nature Genetics*. 2002, vol. 32, n°supplement, pp. 526-532.

MAHADEVAPPA, M. et WARRINGTON, J. A high-density probe array sample preparation method using 10- to 100-fold fewer cells. *Nature Biotechnology*. 1999, vol. 17, pp. 1134-1136.

MARAIS, A., BOVE, J. et RENAUDIN, J. Characterization of the *recA* gene regions of *Spiroplasma citri* and *Spiroplasma melliferum*. *Journal of Bacteriology*. 1996, vol. 178, n°23, pp. 7003-7009.

MARGULIS, L. *Symbiosis in cell evolution*. New York: Freeman, W.H., 1993.

MARGULIS, L. et FENSTER, R. *Symbiosis as a source of evolutionnary innovation*. Cambridge: MIT Press, 1991.

MASKOS, U. et SOUTHERN, E. Parallel analysis of oligodeoxyribonucleotide (oligonucleotide) interactions. I. Analysis of factors influencing oligonucleotide duplex formation. *Nucleic Acids Research*. 1992, vol. 20, n°7, pp. 1675-1678.

MASKOS, U. et SOUTHERN, E. A study of oligonucleotide reassociation using large arrays of oligonucleotides synthesized on glass support. *Nucleic Acids Research*. 1993, vol. 21, n°20, pp. 4663-4669.

MATHEWS, D., SABINA, J., ZUCKER, M. et TURNER, D. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*. 1999, vol. 288, pp. 911-940.

MATSON, R., RAMPAL, J. et COASSIN, P. Biopolymer synthesis on polypropylene supports. I. Oligonucleotides. *Analytical Biochemistry*. 1994, vol. 217, n°2, pp. 306-310.

MATTHEWS, H. Polyamines, chromatin structure and transcription. *BioEssays*. 1993, vol. 15, n°8, pp. 561-566.

MATVEEA, O., SHABALINA, S., NEMTSOV, V., TSODIKOV, A., GESTELAND, R. et ATKINS, J. Thermodynamic calculations and statistical correlations for oligo-probes. *Nucleic Acids Research*. 2003, vol. 31, n°14, pp. 4211-4217.

MAUGHAN, N., LEWIS, F. et SMITH, V. An introduction to arrays. *Journal of Pathology*. 2001, vol. 195, pp. 3-6.

MCGALL, G., LABADIE, J., BROCK, P., WALLRAFF, G., NGUYEN, T. et HINSBERG, W. Light-directed synthesis of high-density oligonucleotide arrays using semi-conductor photoresists. *Proceedings of the National Academy of Sciences of the USA*. 1996, vol. 93, pp. 13555-13560.

MCGINNIS, S. et MADDEN, T. BLAST: at the core of a powerful and diverse set of sequence analysis. *Nucleic Acids Research*. 2004, vol. 32, n°*Web Server*, pp. 20-25.

MENDES, P. GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and others systems. *Computer Applications in the Biosciences*. 1993, vol. 9, n°5, pp. 563-571.

MENDES, P. et KELL, D. Non-linear optimization of biochemical pathways : applications to metabolic engineering and parameter estimation. *Bioinformatics*. 1998, vol. 14, n°10, pp. 869-883.

MERCIER, G., BERTHAULT, N., MARY, J., PEYRE, J., ANTONIADIS, A., COMET, J.-P., CORNUEJOLS, A., FROIDEVAUX, C. et DU-

TREIX, M. Biological detection of low radiation doses by combining results of two microarray analysis methods. *Nucleic Acids Research*. 2004, vol. 32, n°1, pp. e12.

MERGNY, J.-L. et LACROIX, L. Des Tms, encore des Tms, toujours des Tms ! *Regard sur la Biochimie*. 2002, vol. 2, pp. 36-52.

MEUNIER-PREST, R., RAVEAU, S., FINOT, E., LEGAY, G., CHERKAOUI-MALKI, M. et LATRUFFE, N. Direct measurement of melting temperature of supported DNA by electrochemical method. *Nucleic Acids Research*. 2003, vol. 31, n°23, pp. e150.

MIR, K. et SOUTHERN, E. Determining the influence of structure on hybridization using oligonucleotide arrays. *Nature Biotechnology*. 1999, vol. 17, n°8, pp. 788-792.

MIRA, A., OCHMAN, H. et MORAN, N. Deletional bias and the evolution of bacterial genomes. *Trends in Genetics*. 2001, vol. 17, n°10, pp. 589-596.

MITSUHASHI, M. Technical report: part. 1. Basic requirements for designing optimal oligonucleotide probe sequences. *Journal of Clinical Laboratory Analysis*. 1996, vol. 10, pp. 277-284.

MITTLER, T. Dietary amino acid requirements of the aphid *Myzus persicae* affected by antibiotic uptake. *Journal of Nutrition*. 1971, vol. 101, pp. 1023-1028.

MONTLLOR, C., MAXMEN, A. et PURCELL, A. Facultative bacterial endosymbionts benefit pea aphids *Acyrthosiphon pisum* under heat stress. *Ecological Entomology*. 2002, vol. 27, n°2, pp. 189-195.

MORAN, N. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences of the USA*. 1996, vol. 93, n°7, pp. 2873-2878.

MORAN, N. Tracing the evolution of gene loss in obligate bacterial symbionts. *Current Opinion in Microbiology*. 2003, vol. 6, n°5, pp. 512-518.

MORAN, N. et BAUMANN, P. Phylogenetics of cytoplasmically inherited microorganisms of arthropods. *Trends in Ecology and Evolution*. 1994, vol. 9, n°1, pp. 15-20.

MORAN, N. et MIRA, A. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biology*. 2001, vol. 2, n°12, pp. 1-12.

MORAN, N., MUNSON, M., BAUMANN, P. et ISHIKAWA, H. A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. *Proceedings of the Royal Society London B*. 1993, vol. 253, pp. 167-171.

MORAN, N. et WERNEGREEN, J. Lifestyle evolution in symbiotic bacteria : insights from genomics. *Tree*. 2000, vol. 15, n°8, pp.

MORIN, E. *La Méthode*. Paris: Seuil, 1977.

MROWKA, R., SCHUCHHARDT, J. et GILLE, C. Oligodb - interactive design of oligo DNA for transcription profiling of human genes. *Bioinformatics*. 2002, vol. 18, n°12, pp. 1686-1687.

MUNSON, M., BAUMANN, L. et BAUMANN, P. *Buchnera aphidicola* (a prokaryotic endosymbiont of aphids) contains a putative 16S rRNA operon unlinked to the 23S rRNA-encoding gene: sequence determination, and promoter and terminator analysis. *Gene*. 1993, vol. 137, n°2, pp. 171-178.

MUNSON, M., BAUMANN, P. et KINSEY, M. *Buchnera* Gen-Nov and *Buchnera-Aphidicola* Sp-Nov, a Taxon Consisting of the Mycetocyte-

Associated, Primary Endosymbionts of Aphids. *International Journal of Systematic Bacteriology*. 1991, vol. 41, n°4, pp. 566-568.

NADON, R. et SHOEMAKER, J. Statistical issues with microarrays: processing and analysis. *Trends in Genetics*. 2002, vol. 18, n°5, pp. 265-271.

NAKABACHI, A. et ISHIKAWA, H. Differential display of mRNAs related to amino acid metabolism in the endosymbiotic system of aphids. *Insect Biochemistry and Molecular Biology*. 1997, vol. 27, n°12, pp. 1057-1062.

NAKABACHI, A. et ISHIKAWA, H. Polyamine composition and expression of genes related to polyamine biosynthesis in an aphid endosymbiont, *Buchnera*. *Applied and Environmental Microbiology*. 2000, n°8, pp. 3305-3309.

NAKABACHI, A. et ISHIKAWA, H. Expression of host S-adenosylmethionine decarboxylase gene and polyamine composition in aphid bacteriocytes. *Insect Biochemistry and Molecular Biology*. 2001, vol. 31, pp. 491-496.

NAKABACHI, A., ISHIKAWA, H. et KUDO, T. Extraordinary proliferation of microorganims in aposymbiotics pea aphids, *Acyrthosiphon pisum*. *Journal of Invertebrate Pathology*. 2003, vol. 82, pp. 152-161.

NAKAHIGASHI, K., YANAGI, H. et YURA, T. Isolation and sequence analysis of *rpoH* genes encoding sigma 32 homologs from gram negative bacteria: conserved mRNA and protein segments for heat shock regulation. *Nucleic Acids Research*. 1995, vol. 23, n°21, pp. 4383-4390.

NARDON, P. Role de la symbiose dans l'adaptation et la speciation. *Bulletin de la Société Zoologique de France*. 1995, vol. 120, n°4, pp. 397-406.

NARDON, P. et GRENIER, A.-M. Symbiose et évolution. *Annales de la Société Entomologique de France*. 1993, vol. 29, n°2, pp. 113-140.

NASH, J. A computer program to calculate and design oligonucleotide primers from amino acid sequences. *Computer Applications in the Biosciences*. 1993, vol. 9, n°4, pp. 469-471.

NASSER, W., AWADE, A., REVERCHON, S. et ROBERT-BAUDOUY, J. Pectate lyase from *Bacillus subtilis*: molecular characterisation of the gene, and properties of the cloned enzyme. *FEBS Letters*. 1993, vol. 335, n°3, pp. 319-326.

NEIDHART, F., CURTISS III, R., INGRAHAM, J., LIN, E., LOW, K., MAGASANIK, W., REZNIKOFF, W., RILEY, M., SCHAECHTER, M. et UMBARGER, H. *Escherichia coli and Salmonella: cellular and molecular biology*. second edition. Washington: ASM Press, 1996.

NEVES, S. et IYENGAR, R. Modeling of signaling networks. *BioEssays*. 2002, vol. 24, n°12, pp. 1110-1117.

NEWTON, M., KENDZIORSKI, C., RICHMOND, F., BLATTNER, F. et TSU, K. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*. 2001, vol. 8, n°1, pp. 37-52.

NIELSEN, H. et KNUDSEN, S. Avoiding cross hybridization by choosing nonredundant targets on cDNAarrays. *Bioinformatics*. 2002, vol. 18, n°2, pp. 321-322.

NIELSEN, H., WERNERSSON, R. et KNUDSEN, S. Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays. *Nucleic Acids Research*. 2003, vol. 31, n°13, pp. 3491-3496.

NING, Z., COX, A. et MULLILKIN, J. SSAHA : afast search method for large DNA databases. *Genome Research*. 2001, vol. 11, pp. 1725-1729.

NOGUEIRA, T. et SPRINGER, M. Post-transcriptional control by global regulators of gene expression in bacteria. *Current Opinion in Microbiology*. 2000, vol. 3, n°2, pp. 154-158.

NUSSINOV, R., SHAPIRO, B., LE, S. et MAIZEL, J. Speeding up the dynamic algorithm for planar RNA folding. *Mathematical Bioscience*. 1990, vol. 100, n°1, pp. 33-47.

OCHMAN, H. et MORAN, N. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science*. 2001, vol. 292, n°5519, pp. 1096-1098.

OGATA, H., GOTO, S., SATO, K., FUJIBUCHI, W., BONO, H. et KANEHISA, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*. 1999, vol. 27, n°1, pp. 29-34.

OHTAKA, C. et ISHIKAWA, H. Effects of heat treatment on the symbiotic system of an aphid mycetocyte. *Symbiosis*. 1991, vol. 11, pp. 19-30.

OLEKSIAK, M., CHURCHILL, G. et CRAWFORD, D. Variation in gene expression within and among natural populations. *Nature Genetics*. 2002, vol. 32, n°2, pp. 261-266.

OUZOUNIS, C. et VALENCIA, A. Early bioinformatics: the birth of a discipline--a personal view. *Bioinformatics*. 2003, vol. 19, n°17, pp. 2176-2190.

PALACIOS, C. et WERNEGREEN, J. A strong effect of AT mutational bias on amino acid usage in *Buchnera* is mitigated at high-expression genes. *Molecular Biology Evolution*. 2002, vol. 19, n°9, pp. 1575-1584.

PAN, W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*. 2002, vol. 18, n°4, pp. 546-554.

PANINA, E., VITRESCHAK, A., MIRONOV, A. et GELFAND, M. Regulation of aromatic amino acid biosynthesis in gamma-proteobacteria. *Journal of Microbiology and Biotechnology*. 2001, vol. 3, n°4, pp. 529-543.

PANINA, E., VITRESCHAK, A., MIRONOV, A. et GELFAND, M. Regulation of biosynthesis and transport of aromatic amino acids in low-GC Gram-positive bacteria. *FEMS Microbiology Letters*. 2003, vol. 222, n°2, pp. 211-220.

PAUSTIAN, M., MAY, B. et KAPUR, V. Transcriptional response of *Pasteurella multocida* to nutrient limitation. *Journal of Bacteriology*. 2002, vol. 184, n°13, pp. 3734-3739.

PEARSON, W. et LIPMAN, D. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the USA*. 1988, vol. 85, n°8, pp. 2444-2448.

PERRU, O. Qu'est-ce que la symbiose ? *Revue des Questions Scientifiques*. 1997, vol. 168, n°2, pp. 113-136.

PERRU, O. *De la société à la symbiose : une histoire des découvertes sur les associations chez les êtres vivants*. Librairie Philosophique J. Vrin. Paris, 2003.

PERVOUCHINE, D., GRABER, J. et KASIF, S. On the normalization of RNA equilibrum free energy to the length of the sequence. *Nucleic Acids Research*. 2003, vol. 31, n°9, pp. e49.

PETALIDIS, L., BHATTACHARYYA, S., MORRIS, G., COLLINS, V., FREEMAN, T. et LYONS, P. Global amplification of mRNA by template-switching PCR: linearity and application to microarray analysis. *Nucleic Acids Research*. 2003, vol. 31, n°22, pp. e42.

PETHO, A., BELTER, J., BOROS, I. et VENETIANER, P. The role of upstream sequences in determining the strength of an rRNA promoter of *E. coli*. *Biochimica et Biophysica Acta*. 1986, vol. 866, n°1, pp. 37-43.

PEYRET, N., SENEVIRATNE, P., ALLAWI, H. et SANTALUCIA, J. Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G and T.T mismatches. *Biochemistry*. 1999, vol. 38, pp. 3468-3477.

PHIMISTER, B. Going global. *Nature Genetics*. 1999, vol. 21, pp. 1.

PILPEL, Y., SUDARSANAM, P. et CHURCH, G. M. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics*. 2001, vol. 29, n°2, pp. 153-159.

PLAGUE, G., DALE, C. et MORAN, N. Low and homogeneous copy number of plasmid-borne symbiont genes affecting host nutrition in *Buchnera aphidicola* of the aphid *Uroleucon ambrosiae*. *Molecular Ecology*. 2003, vol. 12, n°4, pp. 1095-1100.

POLLACK, J. et IYER, V. Characterizing the physical genome. *Nature Genetics*. 2002, vol. 32, pp. 515-521.

POLLACK, J., PEROU, C., ALIZADEH, A., EISEN, M., PERGAMENSCHIKOV, A., WILLIAMS, C., JEFFREY, S., BOTSTEIN, D. et BROWN, P. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*. 1999, vol. 23, n°1, pp. 41-46.

POZHITKOV, A. et TAUTZ, D. An algorithm and program for finding sequence specific oligo-nucleotide probes for species identification. *BMC Bioinformatics*. 2002, vol. 3, n°1, pp. 9.

PROSSER, W. et DOUGLAS, A. A test of the hypotheses that nitrogen is upgraded and recycled in an aphid (*Acyrthosiphon pisum*) symbiosis. *Journal of Insect Physiology*. 1992, vol. 38, n°2, pp. 93-99.

PROSSER, W., SIMPSON, S. et DOUGLAS, A. How an aphid symbiosis responds to variation in dietary nitrogen. *Journal of Insect Physiology*. 1992, vol. 38, n°4, pp. 301-307.

PROUDNIKOV, D., TIMOFEEV, E. et MIRZABEKOV, A. Immobilization of DNA in polyacrylamide gel for the manufacture of DNA and DNA-oligonucleotide microchips. *Analytical Biochemistry*. 1998, vol. 259, n°1, pp. 34-41.

PROUTSKI, V. et HOLMES, E. Primer master: a new program for the design and analysis of PCR primers. *Computer Applications in the Biosciences*. 1996, vol. 12, n°3, pp. 253-255.

RAMDAS, L., COOMBES, K., BAGGERLY, K., ABRUZZO, L., HIGHSMITH, W., KROGMANN, T., HAMILTON, S. et ZHANG, W. Sources of nonlinearity in cDNA microarray expression measurements. *Genome Biology*. 2001, vol. 2, n°11, pp. research0047.

RAMSAY, G. DNA chips: state-of-the-art. *Nature Biotechnology*. 1998, vol. 16, n°1, pp. 40-44.

RECHENMANN, F. From data to knowledge. *Bioinformatics*. 2000, vol. 16, pp. 411.

RECHENMANN, F. et GAUTIER, C. Donner un sens au génome. *La Recherche*. 2000, vol. 332, pp. 38-45.

REED, J., VO, T., SCHILLING, C. et PALSSON, B. An expanded genome-scale model of *Escherichia coli* K-12(iJR904 GSM/GPR). *Genome Biology*. 2003, vol. 4, n°9, pp. R54.

REES, W., YAGER, T., KORTE, J. et VON HIPPEL, P. Betaine can eliminate the base pair composition dependence of DNA melting. *Biochemistry*. 1993, vol. 32, n°1, pp. 137-144.

REINER, A., YEKUTIELI, D. et BENJAMINI, Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*. 2003, vol. 19, n°3, pp. 368-375.

REITH, M. et MUNHOLLAND, J. A High-Resolution Gene Map of the Chloroplast Genome of the Red Alga *Porphyra purpurea*. *Plant Cell*. 1993, vol. 5, n°4, pp. 465-475.

RELOGIO, A., SCHWAGER, C., RICHTER, A., ANSORGE, W. et VALCARCEL, J. Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Research*. 2002, vol. 30, n°11, pp. e51.

RHODES, J., CROGHAN, P. et DIXON, A. Uptake, excretion and respiration of saccharose and amino acids by the pea aphid *Acyrthosiphon pisum*. *Journal of Experimental Biology*. 1996, vol. 199, n°6, pp. 1269-1276.

RHODES, J., CROGHAN, P. et DIXON, A. Dietary saccharose and oligosaccharide synthesis in relation to osmoregulation in the pea aphid, *Acyrthosiphon pisum*. *Physiological Entomology*. 1997, vol. 22, n°4, pp. 373-379.

RICHMOND, C., GLASNER, J., MAU, R., JIN, H. et BLATTNER, F. Genome-wide expression profiling in *Escherichia coli* K-12. *Nucleic Acids Research*. 1999, vol. 27, n°19, pp. 3821-3835.

RILEY, M. Functions of the gene products of *Escherichia coli*. *Microbiological Reviews*. 1993, vol. 57, n°4, pp. 862-952.

RIS, H. et PLAUT, W. Ultrastructure of DNA-containing areas in the chloroplast of Chlamydomonas. *Journal of Cellular Biology*. 1962, vol. 13, pp. 383-391.

RISPE, C., DELMOTTE, F., VAN HAM, R. et MOYA, A. Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. *Genome Research*. 2004, vol. 14, n°1, pp. 44-53.

RISPE, C. et MORAN, N. Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection. *American Naturalist*. 2000, vol. 156, n°4, pp. 425-441.

ROBERTS, C., NELSON, B., MARTON, M., STOUGHTON, R., MEYER, M., BENNETT, H., HE, Y., DAI, H., WALKER, W., HUGHES, T., TYERS, M., BOONE, C. et FRIEND, S. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science*. 2000, vol. 287, n°5454, pp. 873-880.

ROBYR, D., SUKA, Y., XENARIOS, I., KURDISTANI, S., WANG, A., SUKA, N. et GRUNSTEIN, M. Microarray deacetylation maps determine genome-wide functions for yeast. *Cell*. 2002, vol. 109, n°4, pp. 437-446.

ROCHA, E. et DANCHIN, A. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nature Genetics*. 2003a, vol. 34, n°4, pp. 377-378.

ROCHA, E. et DANCHIN, A. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Research*. 2003b, vol. 31, n°22, pp. 6570-6577.

ROCHE, F., HOKAMP, K., ACAB, M., BABIUK, L., HANCOCK, R. et BRINKMAN, F. ProbeLynx: a tool for updating the association of

microarray probes to genes. *Nucleic Acids Research*. 2004, vol. 32, n°Web Server, pp. 471-474.

ROCKE, D. et DURBIN, B. A model for measurement error for gene expression arrays. *Journal of Computional Biology*. 2001, vol. 8, n°6, pp. 557-569.

ROCKETT, J. et DIX, D. DNA arrays: technology, options and toxicological applications. *Xenobiotica*. 2000, vol. 30, n°2, pp. 155-177.

ROGERS, Y., JIANG-BAUCOM, P., HUANG, Z., BOGDANOV, V., ANDERSON, S. et BOYCE-JACINO, M. Immobilisation of oligonucleotides onto a glass support via disulfide bonds: a method for preparation of DNA microarrays. *Analytical Biochemistry*. 1999, vol. 266, pp. 23-30.

ROTH, M., FENG, L., MCCONNELL, K., SCHAFFER, P., GUERRA, C., AFFOURTIT, J., PIPER, K., GUCCIONE, L., HARIHARAN, J., FORD, M., POWELL, S., KRISHNASWAMY, H., LANE, J., INTRIERI, G., MERKEL, J., PERBOST, C., VALERIO, A., ZOLLA, B., GRAHAM, C., HNATH, J., *et al.* Expression profiling using a hexamer-based universal microarray. *Nature Biotechnology*. 2004, vol. 22, n°4, pp. 418-426.

ROUILLARD, J.-M., HERBERT, C. et ZUCKER, M. OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics*. 2002, vol. 18, n°3, pp. 486-487.

ROUILLARD, J.-M., ZUKER, M. et GULARI, E. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Research*. 2003, vol. 31, n°12, pp. 3057-3062.

ROZEN, S. et SKALETSKY, H. Primer3 on the WWW for general users and for biologist programmers. *Methods of Molecular Biology*. 2000, vol. 132, pp. 365-386.

RUTBERG, B. Antitermination of transcription of catabolic operons. *Molecular Microbiolgy*. 1997, vol. 23, n°3, pp. 413-421.

RYCHLIK, W. et RHOADS, R. A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and *in vitro* amplification of DNA. *Nucleic Acids Research*. 1989, vol. 17, n°21, pp. 8543-8551.

SANDSTROM, J. et MORAN, N. Amino acid budgets in three aphid species using the same host plant. *Physiological Entomology*. 2001, vol. 26, n°3, pp. 202-211.

SANDSTROM, J. et PETTERSSON, J. Amino acid composition of phloem sap and the relation to intraspecific variation in pea aphid (*Acyrthosiphon pisum*) performance. *Journal of Insect Physiology*. 1994, vol. 40, n°11, pp. 947-955.

SANDSTROM, J., RUSSELL, J., WHITE, J. et MORAN, N. Independent origins and horizontal transfer of bacterial symbionts of aphids. *Molecular Ecology*. 2001, vol. 10, n°1, pp. 217-228.

SANDSTROM, J., TELANG, A. et MORAN, N. Nutritional enhancement of host plants by aphids - a comparison of three aphid species on grasses. *Journal of Insect Physiology*. 2000, vol. 46, n°1, pp. 33-40.

SANTALUCIA, J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences of the USA*. 1998, vol. 95, pp. 1460-1465.

SANTALUCIA, J., ALLAWI, H. et SENEVIRATNE, P. Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*. 1996, vol. 35, pp. 355-3562.

SANTALUCIA, J. et TURNER, D. Measuring the thermodynamics of RNA secondary structure formation. *Biopolymers*. 1997, vol. 44, pp. 309-319.

SAPIR, M. et CHURCHILL, G. Estimating the posterior probability of differential gene expression from microarray data [en ligne]. 2000. Disponible sur: http://www.jax.org/research/churchill/pubs.

SASAKI, T. et ISHIKAWA, H. Nitrogen recycling in the endosymbiotic system of the pea aphid, *Acyrthosiphon pisum*. *Zoological Science*. 1993, vol. 10, n°5, pp. 779-785.

SASAKI, T. et ISHIKAWA, H. Production of essential amino acids from glutamate by mycetocyte symbionts of the pea aphid, *Acyrthosiphon pisum*. *Journal of Insect Physiology*. 1995, vol. 41, n°1, pp. 41-46.

SATAGOPAN, J. M. et PANAGEAS, K. S. A statistical perspective on gene expression data analysis. *Statistics in Medecine*. 2003, vol. 22, n°3, pp. 481-499.

SATO, S. et ISHIKAWA, H. Expression and control of an operon from an intracellular symbiont which is homologous to the *groE* operon. *Journal of Bacteriology*. 1997, vol. 179, n°9, pp. 2300-2304.

SCHADT, E., LI, C., ELLIS, B. et WONG, W. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*. 2001, vol. 37, pp. 120-125.

SCHAECHTER, M. *Escherichia coli* and *Salmonella* 2000: the view from here. *Microbiology and Molecular Biology Reviews*. 2001, vol. 65, n°1, pp. 119-130.

SCHENA, M. *DNA microarray: a pratical approach*. Paperback: Oxford University Press, 1999.

SCHENA, M., SHALON, D., DAVIS, R. et BROWN, P. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995, vol. 270, pp. 467-470.

SCHENA, M., SHALON, D., HELLER, R., CHAI, A., BROWN, P. et DAVIS, R. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences of the USA*. 1996, vol. 93, pp. 10614-10619.

SCHIMPER, A. Über die Entwickelung der Chlorophyllkörner und Farbkörper. *Botan Z*. 1883, vol. 41, pp. 102-113.

SCHÜTZ, E. et VON AHSEN, N. Spreadsheet software for thermodynamic melting point prediction of oligonucleotide hybridization with and without mismatches. *Biotechniques*. 1999, vol. 27, pp. 1218-1224.

SEGAL, E., SHAPIRA, M., REGEV, A., PE'ER, D., BOTSTEIN, D., KOLLER, D. et FRIEDMAN, N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*. 2003, vol. 34, n°2, pp. 166-176.

SELINGER, D., CHEUNG, K., JOHANSSON, E., RICHMOND, C., BLATTNER, F., LOCKHART, D. et CHURCH, G. RNA expression analysis using a 30 pair resolution *Escherichia coli* genome array. *Nature Biotechnology*. 2000, vol. 18, pp. 1262-1268.

SHANNON, M. et RAO, S. Transcription. Of chips and ChIPs. *Science*. 2002, vol. 296, n°5568, pp. 666-669.

SHAPIRO, L. et LOSICK, R. Dynamic spatial regulation in the bacterial cell. *Cell*. 2000, vol. 100, n°1, pp. 89-98.

SHARP, P. et LI, W. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for "rare" codons. *Nucleic Acids Research*. 1986, vol. 14, n°19, pp. 7737-7749.

SHARP, P. et LI, W. The codon Adaptation Index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*. 1987, vol. 15, n°3, pp. 1281-1295.

SHCHEPINOV, M., CASE-GREEN, S. et SOUTHERN, E. Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays. *Nucleic Acids Research*. 1997, vol. 25, n°6, pp. 1155-1161.

SHCHERBAKOV, D. et GARBER, M. Overlapping genes in bacterial and bacteriophage genomes. *Molecular Biology*. 2000, vol. 34, n°4, pp. 572-583.

SHERIDAN, S., BENHAM, C. et HATFIELD, G. Activation of gene expression by a novel DNA structural transmission mechanism that requires supercoiling-induced DNA duplex destabilization in an upstream activating sequence. *Journal of Biological Chemistry*. 1998, vol. 273, n°33, pp. 21298-21308.

SHIELDS, D. Switches in species-specific codon preferences: the influence of mutation biases. *Journal of Molecular Evolution*. 1990, vol. 31, n°2, pp. 71-80.

SHIGENOBU, S., WATANABE, H., HATTORI, M., SAKAKI, Y. et ISHIKAWA, H. Genome sequence of the endocellular bacterial symbiot of aphids *Buchnera sp. APS*. *Nature*. 2000, vol. 407, pp. 81-86.

SILVA, F., LATORRE, A. et MOYA, A. Genome size reduction through multiple events of gene disintegration in *Buchnera APS*. *Trends in Genetic*. 2001, vol. 17, n°11, pp. 615-618.

SILVA, F., LATORRE, A. et MOYA, A. Why are the genomes of endosymbiotic bacteria so stable ? *Trends in Genetics*. 2003, vol. 19, n°4, pp. 176-180.

SILVA, F., VAN HAM, R., SABATER, B. et LATORRE, A. Structure and evolution of the leucine plasmids carried by the endosymbiont (*Buchnera aphidicola*) from aphids of the family Aphididae. *FEMS Microbiology Letters*. 1998, vol. 168, pp. 43-49.

SMITH, T. et WATERMAN, M. Identification of common molecular subsequences. *Journal of Molecular Biology*. 1981, vol. 147, n°1, pp. 195-197.

SMYTH, G. et SPEED, T. Normalization of cDNA microarray data. *Methods*. 2003, vol. 31, n°4, pp. 265-273.

SMYTH, G., YANG, Y. et SPEED, T. Statistical issues in cDNA microarray data analysis. *Methods in Molecular Biology*. 2003, vol. 224, pp. 111-136.

SOHAIL, M., AKHTAR, S. et SOUTHERN, E. The folding of large RNAs studied by hybridization to arrays of complementary oligonucleotides. *RNA*. 1999, vol. 5, n°5, pp. 646-655.

SOUTHERN, E. An improved method for transferring nucleotides from electrophoresis strips to thin layers of ion-exchange cellulose. *Analytical Biochemistry*. 1974, vol. 62, n°1, pp. 317-318.

SOUTHERN, E., MIR, K. et SHCHEPINOV, M. Molecular interactions on microarrays. *Nature Genetics*. 1999, vol. 21, pp. 5-9.

STECK, T., FRANCO, R., WANG, J. et DRLICA, K. Topoisomerase mutations affect the relative abundance of many *Escherichia coli* proteins. *Molecular Microbiology*. 1993, vol. 10, n°3, pp. 473-481.

STILLMAN, B. et TONKINSON, J. Expression microarray hybridization kinetics depend on length of the immobilized DNA but are independent of immobilization substrate. *Analytical Biochemistry*. 2001, vol. 295, n°2, pp. 149-157.

STOECKERT, C., CAUSTON, H. et BALL, C. Microarray databases: standards and ontologies. *Nature Genetics*. 2002, vol. 32, pp. 469-473.

STOREY, J. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society*. 2004, vol. 66, n°1, pp. 187-205.

SUGIMOTO, N., NAKANO, S.-I., YONEYAMA, M. et HONDA, K.-I. Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Research*. 1996, vol. 24, n°22, pp. 4501-4506.

'T HOEN, P., DE KORT, F., VAN OMMEN, G. et DEN DUNNEN, J. Fluorescent labelling of cRNA for microarray applications. *Nucleic Acids Research*. 2003, vol. 31, n°5, pp. e20.

'T HOEN, P., TURK, R., BOER, J., STERRENBURG, E., DE MENEZES, R., VAN OMMEN, G. et DEN DUNNEN, J. Intensity-based analysis of two-colour microarrays enables efficient and flexible hybridization designs. *Nucleic Acids Research*. 2004, vol. 32, n°4, pp. e41.

TALAAT, A., HOWARD, S., HALE IV, W., LYONS, R., GARNER, H. et JOHNSTON, S. Genomic DNA standards for gene expression profiling in *Mycobacterium tuberculosis*. *Nucleic Acids Research*. 2002, vol. 30, n°20, pp. e104.

TALAAT, A., HUNTER, P. et JOHNSTON, S. Genome-directed primers for selective labeling of bacterial transcripts for DNA microarray analysis. *Nature Biotechnology*. 2000, vol. 18, n°6, pp. 679-682.

TALLA, E., TEKAIA, F., BRINO, L. et DUJON, B. A novel design of whole-genome microarray probes for *Saccharomyces cerevisiae* which minimizes cross-hybridization. *BMC Genomics*. 2003, vol. 4, n°1, pp. 38.

TAMAS, I. Comparative Genomics of endosymbiotic bacteria. *Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology*. 2002, vol. 712, pp. 1-39.

TAMAS, I., KLASSON, L., CANBÄCK, B., NÄSLUND, A., ERIKSSON, A.-S., WERNEGREEN, J., SANDSTRÖM, J., MORAN, N. et ANDERSSON, S. 50 million years of genomic stasis in endosymbiotic bacteria. *Science*. 2002, vol. 296, pp. 2376-2379.

TAMAS, I., KLASSON, L., SANDSTRÖM, J. et ANDERSSON, S. Mutualists and parasites: how to paint yourself into a (metabolic) corner. *FEBS Letters*. 2001, vol. 498, pp. 135-139.

TANAKA, T., JARADAT, S., LIM, M., KARGUL, G., WANG, X., GRAHOVAC, M., PANTANO, S., SANO, Y., PIAO, Y., NAGARAJA, R., DOI, H., WOOD, W., BECKER, K. et KO, M. Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. *Proceedings of the National Academy of Sciences of the USA*. 2000, vol. 97, n°16, pp. 9127-9132.

TANIGUCHI, M., MIURA, K., IWAO, H. et YAMANAKA, S. Quantitative assessment of DNA microarrays - comparison with northern blot analyses. *Genomics*. 2001, vol. 71, n°1, pp. 34-39.

TAO, H., BAUSCH, C., RICHMOND, C., BLATTNER, F. et CONWAY, T. Functional genomics: expression analysis of *Escherichia coli* growing

on minimal and rich media. *Journal of Bacteriology*. 1999, vol. 181, n°20, pp. 6425-6440.

TAVAZOIE, S., HUGHES, J., CAMPBELL, M., CHO, R. et CHURCH, G. Systematic determination of genetic network architecture. *Nature Genetics*. 1999, vol. 22, n°3, pp. 281-285.

TAYLOR, E., COGDELL, D., COOMBES, K., HU, L., RAMDAS, L., TABOR, A., HAMILTON, S. et ZHANG, W. Sequence verification as quality-control step for production of cDNA microarrays. *Biotechniques*. 2001, vol. 31, n°1, pp. 62-65.

TELANG, A., SANDSTROM, J., DYRESON, E. et MORAN, N. Feeding damage by *Diuraphis noxia* results in a nutritionally enhanced phloem diet. *Entomologia experimentalis et applicata*. 1999, vol. 91, n°3, pp. 403-412.

THIEFFRY, D. From global expression data to gene networks. *BioEssays*. 1999, vol. 21, pp. 895-899.

THIEFFRY, D. et DE JONG, H. Modélisation, analyse et simulation de réseaux génétiques. *Médecine/Sciences*. 2002, vol. 18, pp. 492-502.

TOLONEN, A., ALBEANU, D., CORBETT, J., HANDLEY, H., HENSON, C. et MALIK, P. Optimized in situ construction of oligomers on an array surface. *Nucleic Acids Research*. 2002, vol. 30, n°20, pp. e107.

TOLSTRUP, N., NIELSEN, P. S., KOLBERG, J. G., FRANKEL, A. M., VISSING, H. et KAUPPINEN, S. OligoDesign: optimal design of LNA (locked nucleic acid) oligonucleotide capture probes for gene expression profiling. *Nucleic Acids Research*. 2003, vol. 31, n°13, pp. 3758-3762.

TOMII, K. et KANEHISA, M. A comparative analysis of ABC transporters in complete microbial genomes. *Genome Research*. 1998, vol. 8, n°10, pp. 1048-1059.

TOMIUK, S. et HOFMAN, K. Microarray probe strategies. *Briefings in bioinformatics*. 2001, vol. 2, n°4, pp. 329-340.

TROYANSKAYA, O., GARGER, M., BROWN, P., BOTSTEIN, D. et ALTMAN, R. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*. 2002, vol. 18, n°11, pp. 1454-1461.

TSENG, G., OH, M.-K., ROHLIN, L., LIAO, J. et WONG, W. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessement of gene effects. *Nucleic Acids Research*. 2001, vol. 29, n°12, pp. 2549-2557.

TSUCHIDA, T., KOGA, R. et FUKATSU, T. Host plant specialization governed by facultative symbiont. *Science*. 2004, vol. 303, n°5666, pp. 1989.

TSUCHIDA, T., KOGA, R., SHIBAO, H., MATSUMOTO, T. et FUKATSU, T. Diversity and geographic distribution of secondary endosymbiotic bacteria in natural populations of the pea aphid, *Acyrthosiphon pisum*. *Molecular Ecology*. 2002, vol. 11, n°10, pp. 2123-2135.

TUSHER, V., TIBSHIRANI, R. et CHU, G. Signifiance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the USA*. 2001, vol. 98, n°9, pp. 5116-5121.

TYERS, M. et MANN, M. From genomics to proteomic. *Nature*. 2003, vol. 422, pp. 193-197.

UNTERMAN, B., BAUMANN, P. et MCLEAN, D. Pea aphid symbiont relationships established by analysis of 16S rRNAs. *Journal of Bacteriology*. 1989, vol. 171, n°6, pp. 2970-2974.

URBANCZYK-WOCHNIAK, E., LUEDEMANN, A., KOPKA, J., SELBIG, J., ROESSNER-TUNALI, U., WILLMITZER, L. et FERNIE, A. Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Reports*. 2003, vol. 4, n°10, pp. 1-5.

VAN DAM, R. et QUAKE, S. Gene expression analysis with universal n-mer arrays. *Genome Research*. 2002, vol. 12, n°1, pp. 145-152.

VAN DE PEPPEL, J., KEMMEREN, P., VAN BAKEL, H., RADONJIC, M., VAN LEENEN, D. et HOLSTEGE, F. Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Reports*. 2003, vol. 4, n°4, pp. 387-393.

VAN HAL, N., VORST, O., VAN HOUWELINGEN, A., KOK, E., PEIJNENBURG, A., AHARONI, A., VAN TUNEN, A. et KEIJER, J. The application of DNA microarrays in gene expression analysis. *Journal of Biotechnology*. 2000, vol. 78, n°3, pp. 271-280.

VAN HAM, R., KAMERBEEK, J., PALACIOS, C., RAUSELL, C., ABASCAL, F., BASTOLLA, U., FERNANDEZ, J., JIMÉNEZ, L., POSTIGO, M., SILVA, F., TAMAMES, J., VIGUERA, E., LATORRE, A., VALENCIA, A., MORAN, F. et MOYA, A. Reductive genome evolution in *Buchnera aphidicola*. *Proceedings of the National Academy of Sciences of the USA*. 2003, vol. 100, n°2, pp. 581-586.

VAN HELDEN, J., ANDRE, B. et COLLADO-VIDES, J. A web site for the computational analysis of yeast regulatory sequences. *Yeast*. 2000a, vol. 16, n°2, pp. 177-187.

VAN HELDEN, J., NAIM, A., MANCUSO, R., ELDRIDGE, M., WERNISCH, L., GILBERT, D. et WODAK, S. J. Representing and analysing molecular and cellular function using the computer. *Biological Chemistry*. 2000b, vol. 381, n°9-10, pp. 921-935.

VENTER, J. C., ADAMS, M. D., MYERS, E. W., LI, P. W., MURAL, R. J., SUTTON, G. G., SMITH, H. O., YANDELL, M., EVANS, C. A., HOLT, R. A., GOCAYNE, J. D., AMANATIDES, P., BALLEW, R. M., HUSON, D. H., WORTMAN, J. R., ZHANG, Q., KODIRA, C. D., ZHENG, X. H., CHEN, L., SKUPSKI, M., *et al.* The sequence of the human genome. *Science*. 2001, vol. 291, n°5507, pp. 1304-1351.

VOIT, E. et RILEY, M. Extending knowledge of *Escherichia coli* metabolism by modeling and experiment. *Genome Biology*. 2003, vol. 4, n°11, pp. 235.

VON AHSEN, N., OELLERICH, M., ARMSTRONG, W. et SCHÜTZ, E. Application of a thermodynamic nearest-neighbor model to estimate nucleic acid stability and optimize probe design: prediction of melting points of multiple mutations of apolipoprotein B-3500 and factor V with a hybridization probe genotyping assay on the LightCycler. *Clinical Chemistry*. 1999, vol. 45, n°12, pp. 2094-2101.

VON DOHLEN, C., KOHLER, S., ALSOP, S. et MCMANUS, W. Mealybug beta-proteobacterial endosymbionts contain gamma-proteobacterial symbionts. *Nature*. 2001, vol. 412, n°6845, pp. 433-436.

WADDINGTON, G. *Towards a theoretical biology*. Chicago: Aldine, 1967.

WALLACE, T., SHAFFER, J., MURPHY, R., BONNER, J., HIROSE, T. et ITAKURA, K. Hybridization of synthetic oligodeoxyribonucleotides to

phi chi 174 DNA : the effect of single base pair mismatch. *Nucleic Acids Research*. 1979, vol. 6, pp. 3543-3557.

WALTER, A., TURNER, D., KIM, J., LYTTLE, M., MÜLLER, P., MATHEWS, D. et ZUCKER, M. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proceedings of the National Academy of Sciences of the USA*. 1994, vol. 91, pp. 9218-9222.

WANG, E., MILLER, L., OHNMACHT, G., LIU, E. et MARINCOLA, F. High-fidelity mRNA amplification for gene profiling. *Nature Biotechnology*. 2000, vol. 18, n°4, pp. 457-459.

WANG, H.-Y., MALEK, R., KWITEK, A., GREENE, A., LUU, T., BEHBAHANI, B., FRANK, B., QUACKENBUSH, J. et LEE, N. Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays. *Genome Biology*. 2003, vol. 4, pp. R5.

WANG, P., DING, F., CHIANG, H., THOMPSON, R., WATSON, S. et MENG, F. ProbeMatchDB - a web database for finding equivalent probes across microarray platforms and species. *Bioinformatics*. 2002, vol. 18, n°3, pp. 488-489.

WANG, X. et SEED, B. Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics*. 2003, vol. 19, n°7, pp. 796-802.

WATSON, A., MAZUMDER, A., STEWART, M. et BALASUBRAMANIAN, S. Technology for microarray analysis of gene expression. *Current Opinion in Biotechnology*. 1998, vol. 9, pp. 609-614.

WEIL, M., MACATEE, T. et GARNER, H. Toward a universal standard: comparing two methods for standardizing spotted microarray data. *Biotechniques*. 2002, vol. 32, n°6, pp. 1310-1314.

WERNEGREEN, J. Genome evolution in bacterial endosymbionts of insects. *Nature Review Genetic*. 2002, vol. 3, n°11, pp. 850-861.

WERNEGREEN, J. et MORAN, N. Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes. *Molecular Biology Evolution*. 1999, vol. 16, n°1, pp. 83-97.

WERNEGREEN, J., OCHMAN, H., JONES, I. et MORAN, N. Decoupling of genome size and sequence divergence in a symbiotic bacterium. *Journal of Bacteriology*. 2000, vol. 182, n°13, pp. 3867-3869.

WESTERHOFF, H., O'DEA, M. S., MAXWELL, A. et GELLERT, M. DNA supercoiling by DNA gyrase. A static head analysis. *Cell Biophysics*. 1988, vol. 12, pp. 157-181.

WESTFALL, P. et YOUNG, S. *Resampling-based multiple testing: examples and methods for p-value adjustment*. Wiley: John Wiley & Sons Inc, 1993.

WHITEHEAD, L. et DOUGLAS, A. Populations of symbiotic bacteria in the parthenogenetic pea aphid (*Acyrthosiphon pisum*) symbiosis. *Proceedings of the Royal Society London B*. 1993a, vol. 254, n°1339, pp. 29-32.

WHITEHEAD, L. F. et DOUGLAS, A. E. A metabolic study of *Buchnera*, the intracellular bacterial symbionts of the pea aphid *Acyrthosiphon pisum*. *Journal of general Microbiology*. 1993b, vol. 139, pp. 821-826.

WHITEHEAD, L. F., WILKINSON, T. L. et DOUGLAS, A. E. Nitrogen recycling in the pea aphid (Acyrthosiphon pisum) symbiosis. *Proceedings of the Royal Society London B*. 1992, vol. 250, n°1328, pp. 115-117.

WILCOX, J., DUNBAR, H., WOLFINGER, R. et MORAN, N. Consequences of reductive evolution for gene expression in an obligate endosymbiont. *Molecular Microbiology*. 2003, vol. 48, n°6, pp. 1491-1500.

WILKINSON, T., ADAMS, D., MINTO, L. et DOUGLAS, A. The impact of host plant on the abundance and function of symbiotic bacteria in an aphid. *Journal of Experimental Biology*. 2001, vol. 204, n°17, pp. 3027-3038.

WILKINSON, T., ASHFORD, D., PRITCHARD, J. et DOUGLAS, A. Honeydew sugars and osmoregulation in the pea aphid *Acyrthosiphon pisum*. *Journal of Experimental Biology*. 1997, vol. 200, n°15, pp. 2137-2143.

WILKINSON, T. et ISHIKAWA, H. The assimilation and allocation of nutrients by symbiotic and aposymbiotic pea aphids, *Acyrthosiphon pisum*. *Entomologia Experimentalis et Applicata*. 1999, vol. 91, n°1, pp. 195-201.

WILLIAMS, B., GWIRTZ, R. et WOLD, B. Genomic DNA as a cohybridization standard for a mammalian microarray measurements. *Nucleic Acids Research*. 2004, vol. 32, n°10, pp. e81.

WILLIAMS, J., CASE-GREEN, S., MIR, K. et SOUTHERN, E. Studies of oligonucleotide interactions by hybridization to arrays: the influence of dangling ends on duplex yield. *Nucleic Acids Research*. 1994, vol. 22, n°8, pp. 1365-1367.

WILSON, D., BUCKLEY, M., HELLIWELL, C. et WILSON, I. W. New normalization methods for cDNA microarray data. *Bioinformatics*. 2003, vol. 19, n°11, pp. 1325-1332.

WODICKA, L., DONG, H., MITTMANN, M., HO, M. H. et LOCKHART, D. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnology*. 1997, vol. 15, n°13, pp. 1359-1367.

WOLFINGER, R., GIBSON, G., WOLFINGER, E., BENNETT, L., HAMADEH, H., BUSHEL, P., AFSHARI, C. et PAULES, R. Assessing gene signifiance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*. 2001, vol. 8, pp. 625-637.

WOLKENHAUER, O. Mathematical modelling in the post-genome area: understanding genome expression and regulation-a sytem theoretic approach. *Biosystems*. 2002, vol. 65, pp. 1-18.

WONG, H. et CHANG, S. Identification of a positive retroregulator that stabilizes mRNAs in bacteria. *Proceedings of the National Academy of Sciences of the USA*. 1986, vol. 83, n°10, pp. 3233-3237.

WOOLFIT, M. et BROMHAM, L. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Molecular Biology and Evolution*. 2003, vol. 20, n°9, pp. 1545-1555.

WOOTTON, J. et FEDERHEN, S. Analysis of compositionally biased regions in sequence databases. *Methods in Enzymology*. 1996, vol. 266, pp. 554-571.

WORKMAN, C., JENSEN, L., JARMER, H., BERKA, R., GAUTIER, L., NIELSER, H., SAXILD, H., NIELSEN, C., BRUNAK, S. et KNUDSEN, S. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology*. 2002, vol. 3, n°9, pp. research0048.

WU, H., KERR, K. et CUI, X. MAANOVA: a software package for the analysis of spotted cDNA microarray experiments. In: *The analysis of gene expression data: methods and software*: Springer, 2002.

XIA, T., SANTALUCIA, J. J., BURKARD, M., KIERZEK, R., SCHROEDER, S., JIAO, X., COX, C. et TURNER, D. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*. 1998, vol. 37, pp. 14719-14735.

XU, D., LI, G., WU, L., ZHOU, J. et XU, Y. PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. *Bioinformatics*. 2002, vol. 18, n°11, pp. 1432-1437.

YANG, M., RUAN, Q., YANG, J., ECKENRODE, S., WU, S., MCINDOE, R. et SHE, J. A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. *Physiological Genomics*. 2001, vol. 7, n°1, pp. 45-53.

YANG, Y., BUCKLEY, M., DUDOIT, S. et SPEED, T. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*. 2002a, vol. 11, n°1, pp. 1-40.

YANG, Y., DUDOIT, S., LUU, P., LIN, D., PENG, V., NGAI, J. et SPEED, T. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*. 2002b, vol. 30, n°4, pp. e15.

YANG, Y. et SPEED, T. Design issues for cDNA microarray experiments. *Nature Reviews Genetics*. 2002, vol. 3, n°8, pp. 579-588.

YANG, Y. et SPEED, T. Design and analysis of comparative microarray experiments. In: *Statistical analysis of gene expression microarray data*. Hardcover: Chapman and Hall, 2003.

YE, R., WANG, T., BEDZYK, L. et CROKER, K. Applications of DNA microarrays in microbial systems. *Journal of Microbiological Methods*. 2001, vol. 47, pp. 257-272.

YERSHOV, G., BARSKY, V., BELGOVSKIY, A., KIRILLOV, E., KREINDLIN, E., IVANOV, I., PARINOV, S., GUSCHIN, D., DROBISHEY, A., DUBILEY, S. et MIRZABEKOV, A. DNA analysis and diagnostics on oligonucleotide microchips. *Proceedings of the National Academy of Sciences of the USA*. 1996, vol. 93, pp. 4913-4918.

YOUNG, G., SCHMIEL, D. et MILLER, V. A new pathway for the secretion of virulence factors by bacteria: the flagellar export apparatus functions as a protein-secretion system. *Proceedings of the National Academy of Sciences of the USA*. 1999, vol. 96, n°11, pp. 6456-6461.

YU, J., OTHMAN, M., FARJO, R., ZAREPARSI, S., MACNEE, S., YOSHIDA, S. et SWAROOP, A. Evaluation and optimization of procedures for target labeling and hybridization of cDNA microarrays. *Molecular Vision*. 2002, vol. 8, pp. 130-137.

YUZAWA, H., NAGAI, H., MORI, H. et YURA, T. Heat induction of sigma 32 synthesis mediated by mRNA secondary structure: a primary step of the heat shock response in *Escherichia coli*. *Nucleic Acids Research*. 1993, vol. 21, n°23, pp. 5449-5455.

ZHANG, Z., WILLSON, R. et FOX, G. Identification of characterisitc oligonucleotides in the bacterial 16S ribosomal RNA sequence dataset. *Bioinformatics*. 2002, vol. 18, n°2, pp. 244-250.

ZHU, H., KLEMIC, J.-F., CHANG, S., BERTONE, P., CASAMAYOR, A., KLEMIC, K., SMITH, D., GERSTEIN, M., REED, M. et SNYDER, M. Analysis of yeast protein kinases using protein chips. *Nature Genetics*. 2000, vol. 26, pp. 283-289.

ZIAUDDIN, J. et SABATINI, D. Microarrays of cells expressing defined cDNAs. *Nature*. 2001, vol. 411, pp. 107-110.

ZIEN, A., AIGNER, T., ZIMMER, R. et LENGAUER, T. Centralization: a new method for the normalization of gene expression data. *Bioinformatics*. 2001, vol. 17, n°supplement 1, pp. 323-331.

ZIENTZ, E., SILVA, F. et GROSS, R. Genome interdependence in insect-bacterium symbioses. *Genome Biology*. 2001, vol. 2, n°12, pp. 321-326.

ZUKER, M. Mfold web server for nucleic acid folding hybridization prediction. *Nucleic Acids Research*. 2003, vol. 31, n°13, pp. 3406-3415.

ZUKER, M., MATHEWS, D. et TURNER, D. *Algorithms and thermodynamics for RNA secondary structure prediction: a pratical guide*. Dordrecht: Kluwer, 1999.

Partie F
Annexes

1

1 Paramètres thermodynamiques du modèle du plus proche voisin

1.1 Termes d'initiation et de symétrie

Le calcul des enthalpies, des entropies et des énergies libres d'hybridation entre deux brins d'ADN intègre un paramètre d'initiation et un paramètre de symétrie qui sont rassemblés dans le **Tableau F.1.1**.

Tableau F.1.1 Table des termes de symétrie et d'initiation pour le calcul des énergies d'hybridation d'une solution à pH 7 contenant 1 M de NaCl (d'après Santalucia, 1998).

	ΔH en kcal.mol^{-1}	ΔS en cal.mol^{-1}	ΔG en kcal.mol^{-1}
Initiation	0,1	-2,8	0,98
Initiation (séquences ne contenant que des A et des T)	2,3	4,1	1,03
Symétrie	0	-1,4	-0,4

1.2 Cas des hybrides homologues

Les paramètres d'hybridation entre deux brins parfaitement homologues, utilisés pour le calcul des Tm, sont intégrés dans le **Tableau F.1.2**. L'explication de la lecture de ce tableau est présentée sur la **Figure F.1.1**.

Tableau F.1.2 Table des énergies d'interaction entre bases homologues pour une solution à pH 7 contenant 1 M de NaCl. L'enthalpie (ΔH) est exprimée en kcal.mol^{-1} et l'entropie (ΔS) en cal.mol^{-1} (d'après Santalucia, 1998). La première base d'un couple de nucléotides au sein d'une séquence est lue en ligne et la seconde en colonne.

	A		T		C		G	
	ΔH	ΔS	ΔH	ΔS	ΔH	ΔS	ΔH	ΔS
A	-7,9	-22,2	-7,2	-20,4	-8,4	-22,4	-7,8	-21,0
T	-7,2	-21,3	-7,9	-22,2	-8,2	-22,2	-8,5	-22,7
C	-8,5	-22,7	-7,8	-21,0	-8,0	-19,9	-10,6	-27,2
G	-8,2	-22,2	-8,4	-22,4	-9,8	-24,4	-8,0	-19,9

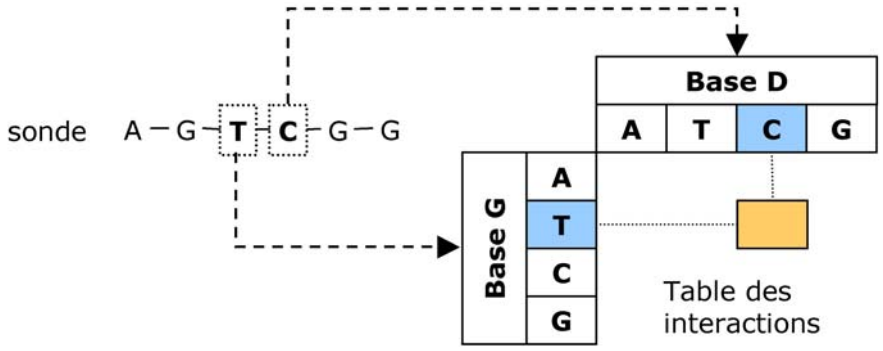


Figure F.1.1 Exemple de lecture d'un couple de nucléotides dans la table contenant les données thermodynamiques d'énergie d'interaction entre bases parfaitement homologues. Sur la séquence, pour chaque couple de bases contiguës, la base située à gauche (base G) est lue en ligne et la base située à droite (base D) en colonne.

1.3 Cas des hybrides avec mésappariements

Les paramètres d'hybridation entre deux brins présentant des mésappariements ont été associés à ceux des hybridations entre brins homologues dans le **Tableau F.1.3**. Ce tableau est utilisé pour le calcul des énergies libres de formation des tiges des épingles à cheveux et des homodimères. L'explication de sa lecture est présentée sur la **Figure F.1.2**. Un tableau supplémentaire (**Tableau F.1.4**) contient les paramètres associés à la formation des boucles dans les épingles à cheveux.

Tableau F.1.3 Table des enthalpies (ΔH en kcal.mol^{-1}) et des entropies (ΔS en cal.mol^{-1}) d'hybridation pour une solution à pH 7 contenant 1 M de NaCl. Les deux dimensions de la table utilisée dans ROSO correspondent aux deux lignes de chaque case de la table. La première ligne contient les valeurs des enthalpies et la seconde les valeurs des entropies. La valeur 0 a été utilisée pour l'enthalpie comme pour l'entropie lorsqu'il n'existe pas de données thermodynamiques disponibles. Le couple de nucléotides du brin 3' est lu en ligne et le couple du brin 5' en colonne. (d'après Allawi et Santalucia, 1997 ; Allawi et Santalucia, 1998a ; Allawi et Santalucia, 1998b ; Allawi et Santalucia, 1998c ; Peyret *et al.*, 1999 ; Santalucia, 1998)

	AA	AT	AC	AG	TA	TT	TC	TG	CA	CT	CC	CG	GA	GT	GC	GG
AA	0	4,7 12,9	0	0	1,2 1,7	0	0	0	0	0	0	-2,9 -9,8	0	0	-0,9 -4,2	0
AT	1,2 1,7	-7,9 -22,2	2,3 4,6	-0,6 -2,3	-7,2 -20,4	-2,7 -10,8	-1,2 -6,2	-2,5 -8,3	5,3 14,6	0,7 0,2	0,0 -4,4	-8,4 -22,4	-0,7 -2,3	1,0 0,9	-7,8 -21,0	-3,1 -9,5
AC	0	7,6 20,2	0	0	5,3 14,6	0	0	0	0	0	0	-0,7 -3,8	0	0	0,6 -0,6	0
AG	0	3,0 7,4	0	0	-0,7 -2,3	0	0	0	0	0	0	0,5 3,2	0	0	-4,1 -13,2	0
TA	4,7 12,9	-7,2 -21,3	3,4 8,0	0,7 0,7	-7,9 -22,2	0,2 -1,5	1,0 0,7	1,3 -5,3	7,6 20,2	1,2 0,7	6,1 16,4	-8,2 -22,2	3,0 7,4	-0,1 -1,7	-8,5 -22,7	1,6 3,6
TT	0	0,2 -1,5	0	0	-2,7 -10,8	0	0	0	0	0	0	-2,2 -8,4	0	0	-5,0 -15,8	5,8 16,3
TC	0	1,2 0,7	0	0	0,7 0,2	0	0	0	0	0	0	2,3 5,4	0	0	-0,8 -',5	0
TG	0	-0,1 -1,7	0	0	1,0 0,9	0	0	0	0	0	0	3,3 10,4	0	-1,4	-4,1 -11,7	0
CA	0	3,4 8,0	0	0	2,3 4,6	0	0	0	0	0	0	5,2 14,2	0	0	1,9 3,7	0
CT	0	1,0 0,7	0	0	-1,2 -6,2	0	0	0	0	0	0	5,2 13,5	0	0	-1,5 -6,1	0
CC	0	6,1 16,4	0	0	0,0 -4,4	0	0	0	0	0	0	3,6 8,9	0	0	-1,5 -7,2	0
CG	-0,9 -4,2	-8,5 -22,7	1,9 3,7	-0,7 -2,3	-7,8 -21,0	-5,0 -15,8	-1,5 -6,1	-2,8 -8,0	0,6 -0,6	-0,8 -4,5	-1,5 -7,2	-8,0 -19,9	-4,0 -13,2	-4,1 -11,7	-10,8 -27,2	-4,9
GA	0	0,7 0,7	0	0	-0,6 -2,3	0	0	0	0	0	0	-0,6 -1,0	0	0	-0,7 -2,3	0
GT	0	-1,3 -5,3	0	0	-2,5 -8,3	0	0	4,1 9,5	0	0	0	-4,4 -12,3	0	5,8 16,3	-2,8 -8,0	0
GC	-2,9 -9,8	-8,2 -22,2	5,2 -22,2	-0,6 -14,2	-8,4 -22,4	-2,2 -8,4	5,2 13,5	-4,4 -12,3	-0,7 -3,8	2,3 5,4	3,6 8,9	-9,8 -24,4	0,5 3,2	3,3 10,4	-8,0 -19,9	-6,0 -15,8
GG	0	1,6 3,6	0	0	-3,1 -9,5	0	0	0	0	0	0	-6,0 -15,8	0	0	-4,9 -15,3	0

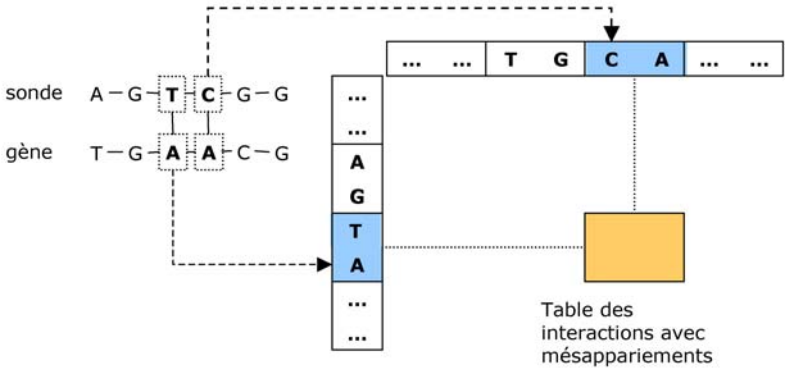


Figure F.1.2 Exemple de lecture d'un couple de nucléotides dans la table contenant les données thermodynamiques d'énergie d'interaction entre bases pouvant présenter des mésappariements.

1.4 Paramètres d'énergie libre de formation des boucles

Les paramètres de formation des boucles dans les épingles à cheveux ont été rassemblés dans le **Tableau F.1.4**.

Tableau F.1.4 Énergies libres de formation des boucles des épingles à cheveux (en kcal.mol^{-1}). L'énergie libre de formation d'une boucle dépend uniquement de sa taille (d'après Freier *et al.*, 1986 ; Groebe et Uhlenbeck, 1988).

Taille de la boucle (en bases)	ΔG (en kcal.mol^{-1})	Taille de la boucle (en bases)	ΔG (en kcal/mol)
3	5,2	23	7,7
4	4,5	24	7,9
5	4,4	25	8,1
6	4,3	26	8,3
7	4,1	27	8,4
8	4,1	28	8,6
9	4,2	29	8,8
10	4,3	30	8,9
11	4,5	31	9,1
12	4,9	>32	9,7
13	5,2	>40	10,2
14	5,6	>45	10,6
15	5,8	>50	11,0
16	6,1	>55	11,3
17	6,4	>60	11,5
18	6,7	>65	11,7
19	6,9	>70	11,9
20	7,1	>80	12,1
21	7,3	>90	12,2
22	7,5	>100	12,3

2

2 Protocoles d'utilisation des lames QUANTIFOIL® et ROSA® en hybridation manuelle

2.1 Traitement des lames

Le protocole de traitement des lames QUANTIFOIL® est donné dans la partie « matériel et méthodes » du chapitre 3.

En ce qui concerne les lames ROSA®, une première étape de traitement permet la fixation des sondes sur les lames immédiatement après leur dépôt. Pour cela, les lames sont séchées et incubées une nuit à température ambiante dans le tampon *RosaBlock1*^{TM} (*Rosa Tech*). Les lames sont ensuite lavées à température ambiante, 2 fois 2 minutes dans de l'éthanol absolu, et 1 fois 2 minutes dans de l'eau ultra-pure. Elles sont ensuite rincées, 3 fois 1 minute dans du SDS 0,1 %, (*Flüka,* Buchs, Suisse), et 1 fois 2 minutes dans de l'eau ultra-pure, de façon à éliminer les sondes qui ne sont pas liées à la lame.

Une seconde étape permet de bloquer tous les sites d'hybridation aspécifique sur les lames. Les lames sont incubées 20 minutes à température ambiante et en agitation, dans une solution [anhydride succinique (*Flüka*), 6 g ; 1-methyl-2-pyrrolidone (*Flüka*), 325 ml ; borate de sodium 1 M, pH 8, 25 ml]. Les lames sont ensuite rincées à température ambiante dans de l'eau distillée portée à 95 °C, puis 1 fois 1 minute dans de l'éthanol 95 % et finalement séchées par centrifugation 1 minute à 2000 g. Une fois ce traitement réalisé, les lames sont conservées à 4 °C et sous dessiccateur.

2.2 Pré-hybridation

L'étape de pré-hybridation permet de réduire l'adsorption non spécifique des cibles fluorescentes sur les lames. Elle est réalisée juste avant l'hybridation. Le protocole diffère légèrement en fonction du type de lame utilisée.

Les lames QUANTIFOIL® sont incubées 60 minutes, à température ambiante et en agitation, dans un tampon de pré-hybridation [SSC 3X ; SDS

0,2 % ; BSA 1 %], puis rincées 3 fois 2 minutes dans l'eau ultra-pure et finalement séchées par centrifugation 1 minute à 2000 g.

Les lames ROSA® sont incubées 60 minutes à 42 °C et en agitation, dans une solution [SSC 5X ; Albumine de Sérum Bovin 1 % (BSA, fraction V) (*Miles Inc.*, Kankakee, IL, USA) ; SDS 0.1 %]. Elles sont ensuite rincées 2 minutes dans l'eau ultra-pure et 30 secondes dans l'isopropanol, puis finalement séchées par centrifugation 1 minutes à 2000 g.

2.3 Hybridation et lavages

Les cibles marquées sont re-suspendues dans 50 µl de tampon d'hybridation *Quantifoilb*^{M} (*Quantifoil*) pour les lames QUANTIFOIL® ou dans 15 µl de tampon d'hybridation *Arrayhyb*^{M} ^{T}*low temperature* (*Sigma-Aldrich*) pour les lames ROSA®. Elles sont ensuite dénaturées par incubation 5 minutes à 95 °C et centrifugées 5 minutes à 12000 g. Elles sont finalement refroidies 10 minutes à température ambiante, puis déposées sur les lames.

Les lames sont ensuite recouvertes d'une lamelle (*Sigma-Aldrich*), placées dans une chambre à hybridation (*Proteigene*, Saint Marcel, France) et incubées 16 heures à 42 °C. Une fois l'hybridation terminée, les lames subissent un cycle de lavages comprenant plusieurs étapes. La lamelle est décollée par agitation dans un bain [SSC 2X, SDS 0,1 %]. Les lames sont lavées successivement dans le bain n°1 [SSC 2 X, SDS 0,1 %], 5 minutes à 42 °C, le bain n°2 [SSC 0,1 X, SDS 0,1 %], 2 fois 10 minutes à température ambiante et le bain n°3 [SSC 0,1 X], 2 fois 2 minutes à température ambiante. Elles sont finalement rincées rapidement dans de l'eau ultra-pure et séchées par centrifugation 1 minute à 2000 g.

3

3 Préparation des milieux artificiels pour l'élevage des pucerons

3.1 Milieu AP3

Le milieu utilisé est le milieu complet AP_3 dont la composition a été définie à partir de l'analyse biochimique de carcasses de pucerons (Febvay *et al.*, 1988). Il permet un développement correct des pucerons sur une génération entière bien que le poids des adultes obtenus soit légèrement inférieur à celui des pucerons élevés sur plante hôte. Le rapport saccharose/acides aminés de ce milieu est de 2,25.

La composition complète du milieu est présentée dans le **Tableau F.3.1**. La solution de micro-composants est préparée avec les vitamines (4), les métaux (5) et autres composés (6) qui sont pesés sur une microbalance et dissous dans 200 ml d'eau ultra-pure. Les acides aminés et le saccharose sont pesés individuellement sur une microbalance puis dissous dans 40 ml d'eau ultra-pure et 20 ml de la solution de micro-composants (diluée 5 fois). Le KH_2PO_4 est ensuite ajouté et dissous par agitation. Une solution de KOH (5 M) est utilisée pour obtenir un pH de 7,5 et le volume final est ajusté avec de l'eau distillée à 100 ml. Le milieu est finalement filtré à travers une membrane de pores de 0,45 µm de diamètre (*Millipore*) et stocké à –20°C.

Tableau F.3.1 Composition du milieu AP3 (pour 100 ml de milieu).

N°	Produit	PM	mM	masse (en mg)
1	Saccharose	342,3	584,28	20000,00
2	Acides aminés L			
	Alanine	89,09	20,06	178,71
	β-alanine	89,10	0,70	6,22
	Arginine	174,20	14,06	244,90
	Asparagine H_20	150,14	19,88	298,55
	Acide aspartique	133,11	6,63	88,25
	Cystéine	121,16	2,44	29,59
	Acide glutamique	147,13	10,15	149,36
	Glutamine	146,15	30,49	445,61
	Glycine	75,07	22,19	166,56
	Histidine HCl H_2O	209,63	6,49	136,02
	Isoleucine	131,18	12,56	164,75
	Leucine	131,18	17,65	231,56
	Lysine Hcl	182,65	19,22	351,09
	Méthionine	149,21	4,85	72,35
	Ornithine HCl	168,62	0,56	9,41
	Phénylalanine	165,19	17,83	294,53
	Proline	115,13	11,23	129,33
	Sérine	105,09	11,83	124,28
	Thréonine	119,12	10,67	127,16
	Tryptophane	204,23	2,09	42,75
	Tyrosine	181,19	2,13	38,63
	Valine	117,15	16,29	190,85
	Total		260,01	3520,45
3	Divers			
	Citrate de calcium			10,00
	Benzoate de cholestérol			2,50
	$MgSO_4$ $7H_20$			242,00
4	Vitamines			
	Acide p-aminobenzoique			10,00
	Acide L-ascorbique			100,00
	Biotine			0,10
	D-panthothénate de calcium			5,00
	Chlorure de choline			50,00
	Acide folique			1,00
	Inositol anhydre			42,00
	Amide nicotinique			10,00
	Pyridoxine HCl			2,50
	Riboflavine			0,50
	Thiamine HCl			2,50
5	Métaux traces			
	$CuSO_4$ $5H_20$			0,47
	$FeCl_3$ $6H_2O$			4,45
	$MnCl_2$ $4H_2O$			0,65
	NaCl			2,54
	$ZnCl_2$			0,83
6	KH_2PO_4			250,00

3.2 Milieux utilisés à l'Université d'York

Les solutions d'acides aminés, de minéraux et de vitamines sont préparées selon la composition donnée dans le **Tableau F.3.2** et conservées à –20 °C. La solution de minéraux est préparée par dilution des composés dans 10 ml d'eau distillée, puis divisée en aliquotes de 0,1 ml, et celle de vitamines par dilution dans 5 ml d'eau distillée et formation d'aliquotes de 0,5 ml. La solution d'acides aminés contenant 50 % d'acides aminés essentiels est préparée en diluant l'ensemble des acides aminés dans 50 ml d'eau distillée. La solution à 25 % d'acides aminés essentiels est préparée en diluant les acides aminés essentiels et non essentiels, respectivement, dans 12,5 ml et 37,5 ml d'eau distillée. Les deux solutions d'acides aminés sont stockées en aliquotes de 5 ml.

Les milieux sont préparés en mélangeant 5 ml de solution d'acides aminés, 0,1 ml de solution de minéraux et 0,1 ml de solution de vitamines. La solution de saccharose est préparée dans 3 ml d'eau distillée et ajoutée au mélange précédent. La solution de phosphate est préparée par dissolution du K_2HPO_4 dans 1 ml d'eau distillée. Son pH est contrôlé entre 7,0 et 7,5 puis la solution est ajoutée au mélange précédent et le volume total est ajusté à 10 ml avec de l'eau distillée. Le milieu obtenu est finalement filtré à travers une membrane de pores de 0,45 µm de diamètre (*Millipore*).

Tableau F.3.2 Composition des milieux.

Produits (*Sigma-Aldrich*)	Masse (en mg)
Acides aminés L non essentiels	25 ml (50 %) ou 37,5 ml (25 %)
Alanine	52
Asparagine	190
Acide aspartique	191
Cystéine	35
Acide glutamique	126
Glutamine	253
Glycine	12
Proline	65
Sérine	61
Tyrosine	12
Arginine	252
Acides aminés L essentiels	25 ml (50 %) ou 12,5 ml (25 %)
Histidine	135
Isoleucine	114
Leucine	114
Lysine	127
Méthionine	42
Phénylalanine	48
Thréonine	103
Tryptophane	59
Valine	103
Minéraux	10 ml
$FeCl_3$ $6H_2O$	11
$CuCl_2$ $4H_2O$	2
$MnCl_2$ $6H_20$	4
$ZnSO_4$	17
Vitamines	5 ml
Biotine	0,1
Pantothénate	5
Acide folique	2
Acide nicotinique	10
Pyridoxine	2,5
Thiamine	2,5
Choline	50
Myo-inositol	50
Saccharose	3 ml
Acide ascorbique	10
Acide citrique	1
$MgSO_4$ $7H_2O$	20
Saccharose	1,7 (0,5 M) ou 3.4 (1 M)
Phosphate	1 ml
K_2HPO_4	115

4

4 Listes des gènes exprimés de façon différentielle

Les liste de gènes obtenus à partir des tests *Fs* et F_3, réalisés sur le terme d'interaction *aa : saccharose* et sur les facteurs *aa* et *saccharose* sont présentées respectivement dans les **Tableau F.4.1**, **Tableau F.4.2** et **Tableau F.4.3**

Tableau F.4.1 Liste des gènes exprimés de façon différentielle pour le terme d'interaction *aa : saccharose* (obtenue avec les tests F_3 et *Fs,* seuil de significativité de la probabilité associée à chaque gène : 0,05). M25 et M50 représentent les rapports M estimés pour les concentrations en saccharose (0,5 M : 1 M), respectivement pour des taux d'acides aminés essentiels de 25 et 50 %. M1 et M0,5 représentent les rapports M estimés pour les taux d'acides aminés essentiels (50 % : 25 %), respectivement pour des concentrations en saccharose de 1 et 0,5 M.

BU	Nom	M25	M50	M1	M0,5
BU579	HslU	0,09705212	-0,1626623	0,12709398	-0,1326204
BU279	trpC	-0,1099045	0,15906373	-0,1383596	0,13060865
BU311	aroA	-0,08471	0,1770068	-0,123181	0,13853585
BU534	argD	-0,0907828	0,14758327	-0,1492845	0,08908154
BU044	tRNA-Thr	0,32547765	-0,8554664	0,51860559	-0,6623385
BU244	tRNA-Ile	-1,0745349	0,86281359	-0,9023197	1,03502873
BU379	tRNA-Leu	0,50601243	-0,8112807	0,62872002	-0,6885731
BU540	tRNA-Ser	0,28565275	-0,4835344	0,14902686	-0,6201603
BU557	tRNA-Asn	-0,2559035	0,47608924	-0,5487138	0,18327894
BU558	tRNA-Met	0,66586992	-1,3323556	0,80138326	-1,1968422
BU593	tRNA-Pro	0,24431826	-0,6616692	0,38611813	-0,5198694
BU601	tRNA-Trp	0,28278539	-0,5943955	0,27954108	-0,5976398
BU059	ribB	0,13845907	-0,1765103	0,13856096	-0,1764084
BU591	hemC	0,7472029	-0,9579783	0,86243616	-0,842745
BU592	hemD	0,41663982	-0,7642888	0,4097793	-0,7711493
BU024	ftsY	-0,112162	0,221432	-0,2371273	0,09646667
BU110	mesJ	0,12708539	-0,1774909	0,22196489	-0,0826114
BU080	fliM	-0,0678536	0,14813739	-0,1157628	0,10022819
BU264	mltE	-0,2316862	0,19041902	-0,2015599	0,22054527
BU337	flgB	-0,206313	0,43843071	-0,3747814	0,26996234
BU339	flgD	-0,1184826	0,20882744	-0,1945512	0,13275881
BU341	flgF	-0,1499883	0,21938706	-0,1718643	0,19751102
BU359	ompF	0,04972747	-0,1465001	0,10109459	-0,095133
BU252	grpE1	0,30384588	-0,6639849	0,49133784	-0,4764929
BU605	hscA	0,31523145	-0,3474767	0,38669774	-0,2760104
BU168	cvpA	0,01578941	-0,3057017	0,15808862	-0,1634025
BU188	rnt	-0,2299516	0,22783618	-0,2197471	0,23804061
BU283	sohB	-0,0681495	0,11788727	-0,113994	0,07204275
BU347	rne	0,46812447	-0,5002305	0,46551468	-0,5028403

BU	Nom	M25	M50	M1	M0,5
BU016	thdF	-0,0643575	0,13088715	-0,0952863	0,09995828
BU007	atpG	0,3380585	-0,6200274	0,51438014	-0,4437058
BU094	tktB	0,17602393	-0,1535077	0,15452083	-0,1750108
BU299	fldA	0,40094243	-0,4873701	0,47931869	-0,4089939
BU417	eno	-0,099867	0,19146809	-0,1496965	0,14163866
BU450	pgk	-0,0806586	0,15206132	-0,1323069	0,1004131
BU469	cyoD	0,26268878	-0,4787518	0,33844811	-0,4029925
BU573	pgi	0,1063305	-0,1606651	0,1606407	-0,1063549
BU092	fabB	0,22724639	-0,2462977	0,24002035	-0,2335237
BU350	fabD	0,50601297	-0,2262689	0,34591579	-0,386366
BU351	fabG	-0,0725135	0,28643103	-0,1725229	0,18642165
BU052	yibN	0,25583901	-0,453038	0,3250365	-0,3838405
BU078	yba1	-0,3971264	0,36912375	-0,296276	0,46997415
BU087	ytfN	-0,1704642	0,25223838	-0,306508	0,11619464
BU113	rnfA	0,06785624	-0,1703367	0,12930827	-0,1088846
BU140	surA	-0,0039792	0,7268086	-0,0971252	0,63366262
BU172	hemK	-0,1403155	0,12981838	-0,1747958	0,0953381
BU181	yba2	-0,1238308	0,24712695	-0,2182038	0,15275397
BU191	ychF	-0,1572968	0,14499141	-0,173918	0,12837021
BU235	dxr	-0,1179213	0,13367813	-0,1306556	0,12094374
BU237	yaeT	-0,0989243	0,16738587	-0,1325884	0,1337218
BU363	ycbY	-0,0412308	0,17995256	-0,1717166	0,0494667
BU385	yrbA	-0,0784864	0,20252461	-0,1792734	0,10173763
BU395	rimM	0,58943703	-0,558532	0,54435743	-0,6036116
BU401	rluD	0,32290847	-0,5259756	0,44656906	-0,402315
BU441	yleA	-0,1515055	0,25970737	-0,1708604	0,24035244
BU442	ybeY	-0,1318103	0,24209496	-0,1643929	0,20951228
BU446	ybeN	-0,3177308	0,50826887	-0,429985	0,39601467
BU452	yggB	0,09582978	-0,1604586	0,1579466	-0,0983418
BU467	yccK	-0,0697851	0,16805296	-0,1226217	0,11521635
BU209	speE	0,1372251	-0,1027009	0,13050217	-0,1094239
BU393	ffh	-0,1487827	0,20387031	-0,1836014	0,16905161
BU108	dcd	0,08599176	-0,1846689	0,13279706	-0,1378636
BU144	carB	-0,1865822	0,29155967	-0,1866251	0,29151675
BU314	trxB	-0,084706	0,2185284	-0,1547937	0,14844067
BU434	gmk	-0,1049764	0,22240294	-0,2135506	0,11382872
BU039	nusG	0,18688908	-0,1466885	0,24003034	-0,0935473
BU085	rpmG	-0,1474056	0,32958884	-0,2440729	0,23292149
BU086	rpmB	-0,1860936	0,34584919	-0,2611079	0,27083487
BU128	rplT	0,11823113	-0,0737047	0,0519067	-0,1400292
BU130	pheT	-0,378088	0,72114727	-0,5358828	0,56335248
BU171	prfA	-0,0945674	0,10904048	-0,1082292	0,09537863
BU230	map	-0,0997816	0,13464748	-0,1048935	0,12953553
BU232	tsf	-0,1425581	0,1615182	-0,1287497	0,17532655
BU315	infA	-0,1608669	0,17939899	-0,1574249	0,182841
BU349	rpmF	-0,1187746	0,3013234	-0,211532	0,20856597
BU366	valS	1,13416695	-0,4224792	0,93506155	-0,6215846
BU387	rplU	-0,1580121	0,25545074	-0,1745158	0,23894704
BU397	rplS	-0,4939597	0,42667329	-0,4711862	0,4494468
BU403	alaS	-0,1082988	0,28459463	-0,245302	0,14759137
BU439	lgt	-0,1130908	0,20318738	-0,1629268	0,15335142
BU445	holA	-0,392443	0,50643764	-0,0385954	0,86028523
BU463	nusB	-0,4279242	0,35433863	-0,2307966	0,55146622
BU499	rpoA	0,13718551	-0,1243114	0,15329808	-0,1081989
BU546	dnaB	0,29539492	-0,5856258	0,64358378	-0,2374369
BU552	mutY	0,26556	-0,4689229	0,41927055	-0,3152124
BU564	rpsF	0,29661975	-0,4864841	0,43352491	-0,3495789
BU569	miaA	-0,5637009	0,69379065	-0,4248838	0,8326077
BU296	ycfV	-0,4335832	0,77409851	-0,5782855	0,62939613

BU	Nom	M25	M50	M1	M0,5
BU318	znuC	-0,1582512	0,13411421	-0,1233998	0,16896561
BU480	mdlB	0,04174204	-0,1285777	0,0882827	-0,0820371
BU587	pitA	-0,2262842	0,17439397	-0,2066239	0,1940543
BU588	ynfM	-0,1633473	0,12909408	-0,1566688	0,1357726

Tableau F.4.2 Liste des gènes exprimés de façon différentielle pour le facteur *aa* (obtenue avec les tests F_3 et *Fs*, seuil de significativité de la probabilité associée à chaque gène : 0,05).

BU	Nom	M*aa* (50 % : 25 %)
BU556	yeeX	-0,8688607
BU554	murI	-0,7312525
BU336	flgA	-0,6624382
BU582	yjeA	-0,5944439
BU028	yigL	-0,5682298
BU431	polA	-0,5295391
BU331	tRNA-Ser	-0,5122454
BU389	yhbZ	-0,4684599
BU041	tRNA-Thr	-0,4496004
BU413	tRNA-Leu	-0,4389152
BU043	tRNA-Tyr	-0,3494168
BU014	rnpA	-0,288758
BU249	tRNA-Asp	-0,2840372
BU067	lig	-0,16239
BU549	yggS	0,10361251
BU250	lpcA	0,15008821
BUpL07	leuD	0,26551337
BU473	bolA	0,52309674
BU348	rluC	0,54424569
BU562	rplI	0,54702012
BU491	rrl	0,55674428
BU551	yggH	0,72992775
BU020	efp	0,9283778
BU418	nlpD	0,97183708
BU454	recB	1,23067351

Tableau F.4.3 Liste des gènes exprimés de façon différentielle pour le facteur *saccharose* (obtenue avec les tests F_3 et *Fs*, seuil de significativité de la probabilité associée : 0,05).

BU	Nom	M*saccharose* (0,5 : 1)
BU066	cysK	0,19385416
BU096	dapA	0,09412876
BU194	thrA	0,16826666
BU278	trpB	0,15321112
BU312	serC	-0,1240491
BU438	lysA	0,17248952
BU493	aroE	-0,1348901
BU538	aroB	-0,14329
BU042	tRNA-Gly	-0,4272141
BU068	tRNA-Lys	-0,4303599
BU069	tRNA-Val	-0,4472085
BU071	tRNA-Ala	-0,4451801
BU249	tRNA-Asp	-0,4491716
BU406	tRNA-Arg	-0,3869276
BU414	tRNA-Met	-0,3872472
BU457	tRNA-Met	-0,3811693
BU492	tRNA-Glu	-0,616071
BU043	tRNA-Tyr	-0,224632
BU111	tRNA-Val	-0,3325759
BU245	tRNA-Ala	-0,2963255
BU575	tRNA-Gly	-0,2766839
BU594	tRNA-His	-0,3252207
BU595	tRNA-Arg	-0,3190797
BU268	lipB	0,28939078
BU361	pncB	-0,1536413
BU407	gshA	0,18017362
BU425	cysG	0,15992762
BU460	thiL	0,2351064
BU222	ftsI	0,24452958
BU223	ftsL	0,33561413
BU326	minD	-0,141365
BU216	murG	-0,3801495
BU218	murD	-0,4201848
BU332	ompA	0,41005959
BU418	nlpD	1,09554663
BU045	murB	0,30821689
BU075	fliH	-0,1533327
BU077	fliJ	-0,3349085
BU186	smpA	-0,1247948
BU215	murC	0,18120857
BU250	lpcA	-0,1375309
BU338	flgC	0,22452568
BU576	amiB	0,269791
BU422	cysC	0,19704173
BU427	cysI	-0,1288922
BU478	ppiD	0,50445839
BU019	mopA	-0,135113
BU152	dnaJ	0,24123201
BU455	recD	0,78668665
BU119	nth	0,29028526
BU453	recC	-0,1666726
BU182	ahpC	-0,1514795
BU002	atpB	0,12939092
BU107	gnd	0,16182587
BU158	nuoF	0,18279503

BU	Nom	M*saccharose* (0,5 : 1)
BU305	pfkA	0,12985156
BU320	zwf	0,11830488
BU470	cyoC	-0,3109655
BU118	ydgQ	-0,6567922
BU254	smpB	-0,4522794
BU323	yoaE	-0,4911209
BU324	yeaZ	0,65186747
BU358	ycfM	-0,4231385
BU389	yhbZ	0,49037693
BU419	ygbB	-0,5322775
BU115	rnfC	0,10741192
BU123	yb1688	-0,2283002
BU187	ydhD	0,20826006
BU274	yciA	0,09159302
BU275	yciB	0,15907489
BU282	yciL	0,26224008
BU286	yfgB	-0,1538664
BU328	yjjT	-0,1342333
BU355	ycfH	0,20945587
BU532	yheN	0,21233387
BU549	yggS	-0,0940282
BU585	yba4	0,18088964
BUpL02	yqhA	-0,2716902
BU259	lepB	0,15614437
BU031	purH	0,15643654
BU204	guaC	0,33601772
BU362	pyrD	0,34892287
BU416	pyrG	0,13551846
BU541	deoD	-0,176355
BU012	dnaA	0,63097794
BU020	efp	0,95456944
BU120	priA	-0,3716237
BU431	polA	-0,7780166
BU521	rplB	-0,579513
BU562	rplI	0,73723368
BU035	rplL	0,14021943
BU056	dnaG	0,30894526
BU121	tyrS	0,23626291
BU131	himA	0,10979557
BU300	phrB	-0,2289112
BU375	truB	-0,1506802
BU377	infB	-0,219146
BU390	rpsI	-0,089172
BU400	fis	0,21543908
BU444	leuS	0,19039806
BU497	fmt	0,28494746
BU517	rplP	-0,1698744
BU065	ptsH	0,19502815
BU572	mtlA	0,262341

FOLIO ADMINISTRATIF

THESE SOUTENUE DEVANT L'INSTITUT NATIONAL DES SCIENCES APPLIQUEES DE LYON

NOM : REYMOND
(avec précision du nom de jeune fille, le cas échéant)

Prénoms : Nancie

DATE de SOUTENANCE : 16 décembre 2004

TITRE : Bioinformatique des puces à ADN et application à l'analyse du transcriptome de *Buchnera aphidicola*

NATURE : Doctorat

Numéro d'ordre : 04 ISAL 0100

Ecole doctorale : Évolution, Écosystèmes, Microbiologie, Modélisation (E2M2)

Spécialité : Analyse et Modélisation des Systèmes Biologiques
Bioinformatique

Cote B.I.U. - Lyon : T 50/210/19 / et bis

CLASSE :

RESUME :

L'objectif de cette thèse est l'étude par la technologie des puces à ADN, du transcriptome de la bactérie *Buchnera aphidicola,* qui vit en symbiose avec le puceron du pois *Acyrthosiphon pisum*. Le génome extrêmement réduit de *Buchnera* a perdu l'essentiel de ses gènes de régulation, mais conserve les voies de biosynthèse permettant la production des acides aminés essentiels pour son hôte. La première partie de cette thèse concerne le développement du logiciel ROSO, qui permet de déterminer les sondes oligonucléotidiques destinées aux puces. Une interface a été conçue pour son utilisation en ligne (http://pbil.univ-lyon1.fr/roso). Dans une seconde partie, la conception et l'utilisation d'une puce dédiée à *Buchnera* ont permis d'étudier le transcriptome de la bactérie, lorsque son hôte subit un stress nutritionnel et osmotique. Cette analyse montre que *Buchnera* est capable de réguler son expression génique et de réorienter son métabolisme, pour répondre à la demande changeante de son hôte.

MOTS-CLES :

bioinformatique – puce à ADN – transcriptome – symbiose – insecte – puceron – *Acyrthosiphum pisum* – bactérie – *Buchnera* – sonde – oligonucléotide.

Laboratoire (s) de recherches : UMR INRA/INSA de Lyon, Biologie Fonctionnelle Insectes et Interactions (BF2I)

Directeurs de thèse :

J.-M. Fayard
H. Charles

Président de jury : C. Gautier

Composition du jury :
H. Charles
J.-M. Fayard
G. Febvay
C. Gautier
M.-C. Potier
D. Tagu
A. Trubuil

Bioinformatique des puces à ADN et application à l'analyse du transcriptome de *Buchnera aphidicola*

Résumé

L'objectif de cette thèse est l'étude par la technologie des puces à ADN, du transcriptome de la bactérie *Buchnera aphidicola,* qui vit en symbiose avec le puceron du pois *Acyrthosiphon pisum*. Le génome extrêmement réduit de *Buchnera* a perdu l'essentiel de ses gènes de régulation, mais conserve les voies de biosynthèse permettant la production des acides aminés essentiels pour son hôte. La première partie de cette thèse concerne le développement du logiciel ROSO, qui permet de déterminer les sondes oligonucléotidiques destinées aux puces. Une interface a été conçue pour son utilisation en ligne (http://pbil.univ-lyon1.fr/roso). Dans une seconde partie, la conception et l'utilisation d'une puce dédiée à *Buchnera* ont permis d'étudier le transcriptome de la bactérie, lorsque son hôte subit un stress nutritionnel et osmotique. Cette analyse montre que *Buchnera* est capable de réguler son expression génique et de réorienter son métabolisme, pour répondre à la demande changeante de son hôte.

Mots-clés : bioinformatique – puce à ADN – transcriptome – symbiose – insecte – puceron – *Acyrthosiphum pisum* – bactérie – *Buchnera* – sonde – oligonucléotide.

Bioinformatics of microarrays and application to the transcriptome analysis of *Buchnera aphidicola*

Abstract

The aim of this thesis is the study of the transcriptome of the bacterium *Buchnera aphidicola*, the primary endosymbiont of the pea aphid, *Acyrthosiphon pisum,* by microarray technology. The high interdependance between *Buchnera* and the aphid caused modifications of the bacterial genome, including loss of most regulatory genes. However, *Buchnera* retained the biosynthesis pathways, for the production of essential amino acids to its host. The first part of this thesis relates to the development of ROSO, a software to design optimized oligonucleotide probes for microarrays. An interface was conceived to allow its use on line (http://pbil.univ-lyon1.fr/roso). In the second part, the design and the use of a chip dedicated to *Buchnera* allowed to study the bacterial transcriptome, under nutritionnal and osmotic stress in the diet of the aphid. This analysis shows that *Buchnera* is able to modify its gene expression and to adapt its metabolism to answer to changing demand of its host.

Keywords : bioinformatic – microarray – transcriptome – symbiosis – insect – aphid – *Acyrthosiphum pisum* – bacteria – *Buchnera* – probe – oligonucleotide.