Exploration et inférence du réseau de régulation de la transcription de la bactérie symbiotique intracellulaire à génome réduit Buchnera aphidicola

Lilia Brinza

N° d'ordre 2010-ISAL-0102
Année 2010

Thèse

Exploration et inférence du réseau de régulation de la transcription de la bactérie symbiotique intracellulaire à génome réduit, *Buchnera aphidicola*

présentée devant
L'Institut National des Sciences Appliquées de Lyon

pour obtenir
le grade de Docteur

Ecole doctorale : Evolution, Ecosystèmes, Microbiologie, Modélisation
Spécialité : Méthodes en bioinformatique moléculaire

Par

Lilia Brinza

Soutenue le 08/12/2010 devant la Commission d'examen

Jury

J. Geiselmann	Professeur	Rapporteur
G. Fichant	Professeur	Rapporteur
F. Calevro	Maître de Conférences	Examinateur
S. Reverchon-Pescheux	Maître de Conférences	Examinateur
H. Vidal	Directeur de recherche	Examinateur
H. Charles	Maître de Conférences	Directeur
C. Gautier	Professeur	Co-directeur

Laboratoire de recherche : Biologie Fonctionnelle, Insectes et Interaction

INSA Direction de la Recherche - Ecoles Doctorales – Quadriennal 2007-2010

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **http://sakura.cpe.fr/ED206** M. Jean Marc LANCELIN Insa : R. GOURDON	M. Jean Marc LANCELIN Université Claude Bernard Lyon 1 Bât CPE 43 bd du 11 novembre 1918 69622 VILLEURBANNE Cedex Tél : 04.72.43 13 95 Fax : lancelin@hikari.cpe.fr
E.E.A.	**ELECTRONIQUE, ELECTROTECHNIQUE, AUTOMATIQUE** **http://www.insa-lyon.fr/eea** M. Alain NICOLAS Insa : C. PLOSSU ede2a@insa-lyon.fr Secrétariat : M. LABOUNE AM. 64.43 – Fax : 64.54	M. Alain NICOLAS Ecole Centrale de Lyon Bâtiment H9 36 avenue Guy de Collongue 69134 ECULLY Tél : 04.72.18 60 97 Fax : 04 78 43 37 17 eea@ec-lyon.fr Secrétariat : M.C. HAVGOUDOUKIAN
E2M2	**EVOLUTION, ECOSYSTEME, MICROBIOLOGIE, MODELISATION** **http://biomserv.univ-lyon1.fr/E2M2** M. Jean-Pierre FLANDROIS Insa : H. CHARLES	M. Jean-Pierre FLANDROIS CNRS UMR 5558 Université Claude Bernard Lyon 1 Bât G. Mendel 43 bd du 11 novembre 1918 69622 VILLEURBANNE Cédex Tél : 04.26 23 59 50 Fax 04 26 23 59 49 06 07 53 89 13 e2m2@biomserv.univ-lyon1.fr
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTE** Sec : Safia Boudjema M. Didier REVEL Insa : M. LAGARDE	M. Didier REVEL Hôpital Cardiologique de Lyon Bâtiment Central 28 Avenue Doyen Lépine 69500 BRON Tél : 04.72.68 49 09 Fax :04 72 35 49 16 Didier.revel@creatis.uni-lyon1.fr
INFOMATHS	**INFORMATIQUE ET MATHEMATIQUES** http://infomaths.univ-lyon1.fr M. Alain MILLE	M. Alain MILLE Université Claude Bernard Lyon 1 LIRIS - INFOMATHS Bâtiment Nautibus 43 bd du 11 novembre 1918 69622 VILLEURBANNE Cedex Tél : 04.72. 44 82 94 Fax 04 72 43 13 10 infomaths@bat710.univ-lyon1.fr - alain.mille@liris.cnrs.fr
Matériaux	**MATERIAUX DE LYON** M. Jean Marc PELLETIER Secrétariat : C. BERNAVON 83.85	M. Jean Marc PELLETIER INSA de Lyon MATEIS Bâtiment Blaise Pascal 7 avenue Jean Capelle 69621 VILLEURBANNE Cédex Tél : 04.72.43 83 18 Fax 04 72 43 85 28 Jean-marc.Pelletier@insa-lyon.fr
MEGA	**MECANIQUE, ENERGETIQUE, GENIE CIVIL, ACOUSTIQUE** M. Jean Louis GUYADER Secrétariat : M. LABOUNE PM : 71.70 –Fax : 87.12	M. Jean Louis GUYADER INSA de Lyon Laboratoire de Vibrations et Acoustique Bâtiment Antoine de Saint Exupéry 25 bis avenue Jean Capelle 69621 VILLEURBANNE Cedex Tél :04.72.18.71.70 Fax : 04 72 43 72 37 mega@lva.insa-lyon.fr
ScSo	**ScSo*** M. OBADIA Lionel Insa : J.Y. TOUSSAINT	M. OBADIA Lionel Université Lyon 2 86 rue Pasteur 69365 LYON Cedex 07 Tél : 04.78.77.23.88 Fax : 04.37.28.04.48 Lionel.Obadia@univ-lyon2.fr

*ScSo : Histoire, Geographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

A mes parents, Maria et Petru
A ma sœur, Aliona

Ce travail n'aurait pu voir le jour ni prendre cette forme sans la présence et le concours des aides précieuses. Je souhaite ici exprimer ma gratitude à tous ceux qui ont contribué d'une façon ou d'une autre à ce que je puisse mener à bien ce projet.

Je remercie

Hubert Charles et Christian Gautier pour leur confiance, leur disponibilité, la grande liberté qu'ils m'ont accordée et l'intranquillité de la pensée sur laquelle ils ne sauraient céder, pour leur approche du monde de la recherche que je garderai en modèle,

Gwennaelle Fichant, Hans Gheiselmann, Sylvie Reverchon et Hubert Vidal pour avoir accepté de lire ce travail, et lui donner le temps de la réponse,

Vincent Navratil, Hidde de Jong, Stéphane Robin, Gérard Febvay et Stéphane Genieys, pour leurs conseils et patience lors des comités de pilotage et en dehors,

Féderica Calevro, Jean-Michel Fayard et Yvan Rahbé pour avoir soutenu, chacun à leur manière, mes activités d'enseignant et de chercheur,

Karen Gaget, Pedro da Silva et Fréderic Gressent pour leurs amitié et pour les mémorables moments que nous avons passés dans la fameux bureau « Brisefer »,

Panagiotis Sapountzis pour son amitié, pour ses encouragement et sa merveilleuse cuisine grecque,

Tous les membres du laboratoire pour leur accueil, leur aide, leurs conseils et pour les excellents moments sans ombre durant ces quatre années,

Jacqueline Marvel pour son accueil et soutien, bien sûr, et la stimulation intellectuelle qu'elle suscite, continûment,

Mireille Radzyner et Michel Labrosse, pour leur patience et soutien discret et élégant durant ces années,

La famille Ogier : Maelle, Ghislaine, Michel, Christophe, Claire Dandurand et Gregory Thomas, pour leur accueil et leur générosité sans fond,

Elena et Robert Rizea, Adrian et Daiana Basarab, Alina Firicel, Alexandru Sirbu, Ioan Burciu, Amira ben Hamida, Iulius Dragonu, Nicoleta Baxan, Marius Stan, Costin et Alexandru Caciu, Celine Brunet Dunand,

Adrien Richard, pour d'innombrables raisons dont leur présence durable et leur énergie rayonnante,

Onu Nechita, la liste serait trop longue et encore incomplète, je le remercierai ici surtout pour avoir été toujours à mes côtés et pour incarner l'ami idéal,

Lynda Maurice, pour son amitié spontanée et inespérée, pour son soutien inconditionnel et pour les bonnes idées qu'elle sait m'inspirer,

Lucas Boucinha, pour sa présence et son aide surtout dans les moments plus difficils, en fin de thèse,

Mes parents et ma sœur pour avoir immanquablement soutenu, depuis ses prémices, l'engagement dans un tel chemin, pour avoir déterminé l'orientation de ce chemin et m'avoir transmis la persévérance pour l'endurer.

Table des matières

Liste des Figures

Liste des Tableaux

Avant-propos

Le puceron du pois, *Acyrthosiphon pisum* est étudié au sein de l'UMR INRA-INSA de Lyon « Biologie fonctionnelle, Insectes et Interactions » (BF2I) depuis 1985, Yvan Rahbé et Gérard Febvay étant les pionniers de ces études. Vers les années 90, les études métaboliques qu'ils faisaient sur le puceron ont conduit les deux chercheurs à s'intéresser à *Buchnera aphidicola*, une Entérobactériacée proche d'*Escherichia coli*, vivant en symbiose intracellulaire obligatoire avec le puceron. Cette bactérie permet au puceron de se développer et de se multiplier très efficacement sur de nombreuses plantes cultivées, notamment des légumineuses d'intérêt agronomique. Elle est donc en partie responsable des pertes économiques considérables liées au puceron dans les régions tempérées. Ainsi, des analyses physiologiques (Febvay *et al.,* 1995; Febvay *et al.,* 1999) ont mis en évidence la forte interaction métabolique dans le couple symbiotique, *Buchnera* synthétisant et fournissant à son hôte les acides aminés essentiels que lui-même ne peut pas synthétiser, ni trouver dans son alimentation naturelle très déséquilibrée, la sève phloémienne. Le séquençage du premier génome de *Buchnera* par l'équipe japonaise de Shigenobu et al. (2000), suivi ensuite par le séquençage des génomes des *Buchnera* de 3 autres espèces de puceron (*Schizaphis graminum*, *Baizongia pistaciae* et *Cinara cedri*) ont mis en évidence des propriétés de ce modèle bactérien liées très fortement au mode de vie intracellulaire, comme par exemple un génome fortement réduit caractérisé par un fort biais compositionnel en bases A et T, ou encore, la perte d'une partie importante des protéines intervenant dans la réparation et la recombinaison de l'ADN. Des analyses génomiques accompagnées par des analyses d'expression des gènes ont enfin révélé une autre particularité de cette bactérie : la perte de la plupart des régulateurs de la transcription et une faible réponse transcriptionnelle aux conditions de stress (Moran *et al.,* 2003; Reymond *et al.,* 2006). Ainsi l'étude de la régulation de la transcription des gènes chez *Buchnera* en réponse aux besoins du puceron constitue un sujet à double intérêt : appliqué d'une part, car une meilleure compréhension des mécanismes moléculaires de l'association du couple symbiotique pourrait permettre de développer de nouveaux moyens de lutte contre les pucerons par l'identification de gènes ou des voies métaboliques cibles ; et fondamental d'autre part, car *Buchnera* montre un parcours évolutif original de son réseau de régulation de la transcription des gènes influencé par les conditions de vie intracellulaire et la réduction drastique de son génome.

Cette question de la régulation de la transcription des gènes chez *Buchnera* a été abordée au BF2I dans les années 2000 (Calevro *et al.,* 2004). C'est à dire parallèlement à la révolution technologique dans la biologie moléculaire qui par les nouveaux types de données biologiques qu'elle a introduits, a provoqué une poussé d'intérêt vers la biologie systémique. Bien que ces études systémiques aient beaucoup gagné en puissance avec le séquençage massif des génomes en ce nouveau début de décennie, nous sommes encore très loin d'avoir compris tous les mécanismes de la régulation de l'expression des gènes dans la cellule, et encore moins la façon dont les acteurs de la régulation interagissent, pour permettre l'adaptabilité et l'homéostasie des organismes en réponse à la variabilité de leur environnement. Plusieurs grands verrous des études des réseaux de régulation, comme par exemple la taille des systèmes et la quantité de données expérimentales nécessaires à leur reconstruction ou encore la description et la comparaison de leurs propriétés topologiques, restent encore à être levés. Les différentes méthodes d'étude des systèmes de régulation se regroupent en deux grands types d'approches : les approches ascendantes et les approches descendantes. La première est la plus traditionnelle, elle consiste à collectionner le plus de détails possible sur les éléments du système pour ensuite les relier dans un modèle intégrateur. Elle permet donc d'obtenir des connaissances avec un grand niveau de détail biochimique et de confiance (Schmidt and Baliga 2007). L'approche descendante quant à elle, permet de créer des réseaux de régulation *de novo*, à partir de mesures globales et simultanées de certains paramètres cellulaires (*e.g.* le niveau de transcription de chaque gène dans une condition donnée). Cette approche peut s'avérer particulièrement intéressante de par sa rapidité et son adaptation aux cas d'organismes dont le réseau de régulation n'a jamais été étudié auparavant. Bien sûr, le niveau de détail et de confiance de ces études n'est pas aussi haut que celui des approches ascendantes.

Ainsi, analyser la régulation de la transcription des gènes de *Buchnera* pour comprendre son réseau de régulation génétique, en combinant des méthodes ascendantes et descendantes, a été le fil directeur de cette thèse.

Après une première partie introductive présentant le modèle biologique, la machinerie transcriptionnelle des procaryotes et les méthodes d'inférences de réseaux de régulation génique, ce manuscrit présente les données et les méthodes que nous avons utilisées et/ou développées dans le cadre de ce travail (notamment une méthode bayésienne de prédiction d'opéron, et une méthodes d'inférence de réseaux de gènes à partir de données d'expression). La partie Résultats est décomposée en quatre parties majeures. La première partie dresse l'inventaire de la machinerie transcrip-

tionnelle de *Buchnera APS*. La deuxième partie a consisté à étudier l'architecture génomique de *Buchnera APS*, *i.e.* l'organisation et l'évolution de sa carte opéronique, l'agencement des fragments synthéniques et non-synthéniques et également les forces d'évolution ayant amené à l'agencement des gènes de *Buchnera* le long de son chromosome. La troisième partie a analysé les propriétés structurelles séquence-dépendantes du chromosome de *Buchnera APS* qui ont ensuite été mises en relation avec le profil de périodicité de l'expression des gènes le long du chromosome. Les résultats obtenus à l'issue de cette approche ascendante (*bottom-up*), nous ont amenés à construire un modèle de réseau de la régulation transcriptionnelle chez *Buchnera APS*. Enfin, la quatrième partie, descendante (*top-down*), a consisté à développer une méthode d'inférence de réseau de régulation à partir de données d'expression, appelée IGOIM. Enfin, une discussion générale est proposée à la fin de ce manuscrit.

Un travail expérimental a également été effectué avec le support particulier de Féderica Calevro, dans le but de confirmer certaines de nos prédictions.

PARTIE I

Introduction

1 La symbiose

1.1 Généralités sur la symbiose

Tous les organismes vivants interagissent de façon durable (c'est à dire sur une durée importante par rapport à leur cycle de vie) avec des organismes d'une ou de plusieurs autres espèces. Ces interactions sont extrêmement diversifiées allant par exemple du parasitisme[1], au commensalisme[2] ou au mutualisme[3]. Le terme de symbiose a été défini dans cette acception pour la première fois en 1879 par le lichénologue Anton de Bary : « La symbiose est une association permanente entre au moins deux organismes spécifiquement distincts, pour au moins une partie de leur cycle de vie » (De Bary 1879). Sachant que beaucoup des relations d'associations entre organismes ne sont pas statiques, des transitions d'un type à l'autre peuvent se faire fréquemment au cours de l'évolution. Ces transitions peuvent être induites par des changements de l'environnement ou par des changements entrainés pas l'association elle-même. Par exemple, il est probable que beaucoup des associations mutualistes actuelles aient débuté comme des associations de parasitisme (Nardon *et al.,* 1993; Douglas 2008).

Dans une relation symbiotique, on appelle l'« hôte » le plus grand des organismes, celui qui héberge, les autres étant appelés les « symbiotes ». Différents types de classifications de symbioses ont été proposés dans la littérature utilisant la localisation des symbiotes, la persistance de l'association ou l'interdépendance des organismes impliqués dans la relation (Nardon *et al.,* 2002). Par exemple, trois types majeurs de symbiose ont été définis en fonction de la localisation du symbiote par rapport à

1 Une association dans laquelle un des symbiotes tire des bénéfices de l'autre au détriment de ce dernier.

2 Une association dans laquelle un des symbiotes en profite, tandis que l'autre n'en est influencé ni de manière négative ni de manière positive.

3 Une association dans laquelle les deux symbiotes tirent des bénéfices réciproques de l'association.

l'hôte : l'exosymbiose, les partenaires restent externes l'un par rapport à l'autre ; l'endosymbiose, le symbiote est contenu à l'intérieur de l'hôte, mais il reste extracellulaire (à l'instar des symbiotes du tube digestif de la vache ou des insectes, ou encore des symbiotes luminescents de certains mollusques) ; l'endocytobiose, le symbiote est contenu dans les cellules de l'hôte (par exemple, les symbiotes nutritionnels des insectes vivant dans des niches écologiques pauvres, ou les symbiotes effectuant des distorsions de la sexualité, comme les *Wolbachia*). Provorov (2005) définissent trois types de symbiose en fonction de la dépendance des bactéries symbiotiques par rapport à leur hôte : facultatives, écologiquement obligatoires et génétiquement obligatoires. Les symbioses facultatives des bactéries avec les animaux ou les plantes ne sont pas accompagnées de changements qualitatifs du potentiel d'adaptation des bactéries (*e.g.* des entérobactéries nonpathologiques et les rhizobactéries). Les interactions écologiquement obligatoires permettent aux bactéries d'occuper temporairement une niche chez leur hôte, pour se soustraire à la compétition avec les bactéries à forme de vie libre. Dans ce cas, les symbiotes peuvent utiliser des gènes spéciaux qu'ils ont acquis dans le contexte symbiotique (*e.g.* la symbiose entre la bactérie *Rhizobium* et les légumineuses). Enfin, les symbioses génétiquement obligatoires concernent les micro-organismes ayant subi des pertes de systèmes essentiels à l'autonomie et ne peuvent donc pas survivre en dehors de l'association symbiotique (*e.g. Ricketsia*, *Buchnera*, *Wolbachia*).

Les associations symbiotiques ont un impact majeur sur l'évolution des espèces, constituant un important moteur de l'évolution en tant que facteur de développement de formes de vie avec un plus haut niveau d'organisation et de nouvelles relations écologiques (Provorov *et al.,* 2008). L'endocytobiose plus particulièrement a eu un impact majeur dans l'apparition de nouveaux niveaux d'organisation, étant à l'origine des organites comme les plastes et les mitochondries (Margulis and Fester 1991).

Une autre raison pour laquelle l'endocytobiose est considérée comme une force d'évolution très importante est le fait qu'elle entraine une reconstruction de l'individu au niveau duquel la sélection naturelle s'exerce. Car en plus de la sélection parmi les individus de la même espèce (symbiote ou hôte), transformée par le cadre symbiotique, il y aussi une sélection jointe qui s'ajoute, s'exerçant sur l'entité chimérique hôte-symbiote, que Nardon a appelée le symbiocosme (Nardon and Grenier 1993). Cet aspect de sélection « jointe » ouvre sur des questions fondamentales comme celle de la notion d'individu et de l'identité biologique.

L'UMR BF2I est spécialisée dans l'étude des endocytobioses et notamment celles impliquant des insectes ravageurs ou auxiliaires et leurs

bactéries intracellulaires. La suite de cette analyse se portera donc sur ces symbioses particulières d'intérêt agronomique.

1.2 Les endocytobioses à bactéries chez les insectes

Les insectes semblent faire partie des êtres vivants les plus tolérants aux organismes étrangers et ont révélé des types extrêmement diversifiés d'associations avec des micro-organismes (Buchner 1965). Grâce à ces associations, les insectes ont notamment pu occuper les niches les plus variées et s'adapter aux conditions les plus extrêmes (homoptères phloémophages strictes, punaises hématophages, etc.). Lorsqu'une population d'insecte acquiert un symbiote et parvient à coloniser une niche, cette population peut s'isoler et conduire à la formation d'une nouvelle espèce (Schewemmler 1991). Plus de 10% des espèces d'insectes connues doivent ainsi leur viabilité et leur reproduction à l'endocytobiose (Moran and Baumann 2000). Dans la majorité des cas, les symbiotes sont des bactéries. Ces bactéries sont présentes dans la lignée germinale et sont donc héritées maternellement, permettant une transmission efficace des « gènes symbiotiques » à travers les générations d'hôtes.

Dans le cas des endocytobioses, l'association est souvent obligatoire, ni l'insecte ni la bactérie ne sont capables de survivre en dehors de l'association. Ainsi, les bactéries ne sont pas cultivables *in vitro* et l'élimination de la bactérie compromet la viabilité et le développement de l'insecte. Il existe globalement deux sortes d'endocytobiotes bactériens : les endocytobiotes nutritionnels et les perturbateurs de sexualité. Les endocytobiotes nutritionnels sont généralement localisés dans des cellules particulières de l'hôte (cf. Partie I, § 2.4), les bactériocytes. Les perturbateurs de la sexualité, à l'instar de *Wolbachia* sont généralement non localisés spécifiquement et même s'ils ne sont pas cultivables, ils ne sont souvent pas obligatoires pour l'hôte (Werren *et al.,* 1995).

1.3 Les endocytobioses à bactériocytes

Les endocytobioses à bactériocytes sont notamment caractéristiques pour 3 groupes d'insectes : l'ordre des blattes, l'ordre des Homoptères et la famille des Curculionidae chez les Coléoptères (Ishikawa 2003). Chez ces insectes, les symbiotes sont logés dans des cellules spécialisées, appelées bactériocytes lorsque les symbiotes sont des bactéries, ou mycetocytes, si les symbiotes sont des champignons. Le principal rôle des bactéries endocytobiotiques est essentiellement nutritionnel, en fournissant l'hôte avec des composés absents de son milieu nutritionnel (*e.g.* acides aminés, vitamines)

(Zientz *et al.,* 2004). Les bactériocytes peuvent être assemblés dans un organe, le bactériome, cet organe ayant des localisations différentes en fonction de l'insecte hôte (Buchner 1965). Globalement, les bactéries trouvées dans les bactériocytes de ces espèces d'insectes ne sont pas monophylétiques (Baumann *et al.,* 1993), néanmoins un grand nombre d'entre elles proviennent du groupe des γ-3 Protéobactéries. Ce clade bactérien ancestral possédait vraisemblablement des propriétés génétiques, métaboliques et écologiques très adaptées pour établir des symbioses avec un grand nombre d'espèces d'insectes (Charles *et al.,* 2001).

2 Le tandem *Buchnera* – puceron

Parmi les endocytobioses à bactériocytes, l'association des pucerons avec leurs bactéries symbiotiques figure parmi les plus étudiées. Cette association a été caractérisée pour la première fois par Paul Buchner dans son livre « Endosymbiosis of animals with plant microorganisms" (Buchner 1965). À l'époque, Buchner avait popularisé le concept de symbiose pour la communauté scientifique et avait déjà formulé l'hypothèse nutritionnelle pour les symbiotes intracellulaires obligatoires des insectes « les endosymbioses doivent d'une certaine façon avoir une cause écologique, leur existence est généralement liée à un milieu nutritionnel déséquilibré ». En effet, *Buchnera aphidicola* (ce nom a été proposé par l'équipe de Paul Baumann (Munson *et al.,* 1991a; Munson *et al.,* 1991b)) est associée à la plupart des pucerons se nourrissant de la sève phloémienne. Elle n'est pas présente dans les Phylloxeridae (pucerons primitifs capables d'ingérer du contenu cellulaire indifférencié), mais elle est présente chez tous les Aphidoidea (phloémophages stricts), à l'exception de la tribu des Cerataphidini, dans laquelle *Buchnera* a été remplacée par une levure symbiotique spécifique (Fukatsu and Ishikawa 1992). *Buchnera* est étroitement associée à la physiologie du puceron. C'est en partie à cause de *Buchnera* que les pucerons ont pu se développer avec tant de succès (Douglas 2006a). *Buchnera* fournit au puceron les acides aminés essentiels que le puceron ne peut synthétiser ni trouver dans la sève phloémienne (**Figure 1**).

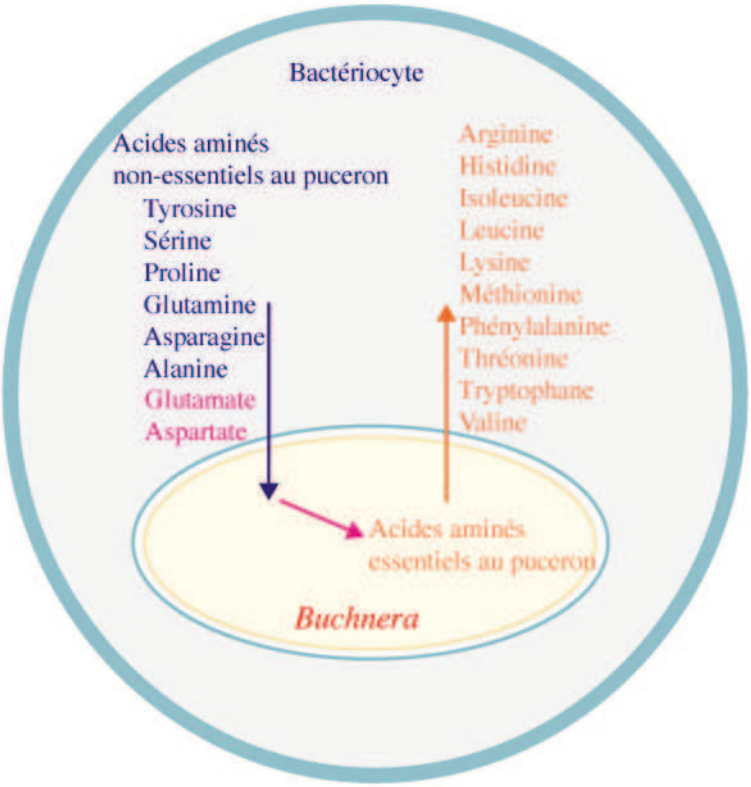


Figure 1. Modèle d'interdépendance des voies de biosynthèse des acides aminés entre *Buchnera* et son hôte. *Buchnera* est capable de synthétiser 10 acides aminés essentiels au puceron (en orange). En revanche, l'hôte fournit à la bactérie les autres acides aminés qu'elle n'est pas capable de synthétiser et qui sont non-essentiels (trouvés dans l'alimentation) pour le puceron (en violet). Les deux précurseurs majeur de la biosynthèse des acides aminés essentiels fournit par l'hôte à *Buchnera* sont marqués en rose (d'après Zientz et al. (2001)), la tyrosine (non alimentation) est synthétisée par le puceron dans le bactériocyte à partir de la phénylalanine.

Les pucerons se nourrissent exclusivement de la sève phloémienne. Ils entrainent des importants dégâts sur les plantes d'une part par le détournement de la sève qui est nécessaire aux plantes, et d'autre part autre part à cause de la toxicité de la salive qu'ils injectent dans la plante lors de la prise alimentaire et des virus dont ils sont souvent les vecteurs (Dedryver 2007). On estime qu'en zone tempérée une espèce végétale sur quatre est attaquée par les pucerons. En pratique, il s'agit de toutes les espèces végétales d'intérêt agricole (Dixon 1998).

Le nom de *Buchnera aphidicola* a été utilisé pour toutes les lignées des symbiotes des pucerons, néanmoins le nom d'espèce *aphidicola* devient de moins en moins utilisé car ces symbiotes semblent trop diversi-

fiés pour être considérés comme une seule espèce (Ishikawa 2003). Souvent, seul le nom de genre, *Buchnera*, est utilisé pour faire référence aux symbiotes des pucerons et le nom du puceron est utilisé pour distinguer les souches de *Buchnera* (Munson *et al.,* 1991b). Ainsi, par la suite nous allons utiliser *Buchnera APS* ou BAp pour faire référence au *Buchnera* du puceron du pois (*Acyrthosiphon pisum*), *Buchnera Sg* ou BSg pour les *Buchnera* du puceron du blé (*Schizaphis graminum*), *Buchnera Bp* ou BBp pour les *Buchnera* du puceron du pistachier (*Baizongia pistaciae*) et *Buchnera Cc* ou BCc, pour les *Buchnera* du puceron du cèdre (*Cinara cedri*), *Buchnera BMb* ou BMb pour les *Buchnera* du puceron du pêcher (*Myzus persicae*). De façon plus globale, nous utiliserons le terme *Buchnera* pour parler des résultats des études génomiques ou physiologiques obtenus presque exclusivement sur les 5 souches de *Buchnera* précitées. Dans le cas contraire, la souche de *Buchnera* sera précisée.

Le positionnement phylogénétique de *Buchnera* reste encore ambigu à cause du fort biais compositionnel en AT de ses gènes et de sa grande vitesse d'évolution. Néanmoins, plusieurs travaux s'accordent pour l'assignation de *Buchnera* dans le groupe des γ3-Protéobactéries (Munson *et al.,* 1991a; Munson *et al.,* 1991b; Charles *et al.,* 2001; Herbeck *et al.,* 2005). De ce fait, *Buchnera* est très proche des Enterobacteriacées et donc d'*Escherichia coli*. Comme *Buchnera* est transmise de manière exclusivement verticale entre les générations de pucerons (cf. Partie I, §2.4), les événements de spéciation des hôtes correspondent aux événements de divergence des endocytobiotes associés entraînant ainsi une parfaite congruence entre les arbres phylogénétiques des deux groupes d'espèces (**Figure 2**). Afin de tracer les événements de spéciation dans le groupe *Buchnera* et les taux d'évolution entre les lignées (Moran 1996; Moran *et al.,* 2009), les dates de divergence ont été estimées en utilisant les fossiles des pucerons et les taux d'évolution des gènes de *Buchnera* (Moran *et al.,* 1993) (**Figure 2**). La date de l'association entre le puceron et *Buchnera* a ainsi été estimée à environ 150 M (Moran *et al.,* 1993; Von Dohlen and Moran 2000). Durant ce processus de coévolution, les deux partenaires ont perdu leur autonomie : les pucerons ne peuvent se reproduire sans leurs endocytobiotes et *Buchnera* ne peut être cultivée en dehors du puceron.

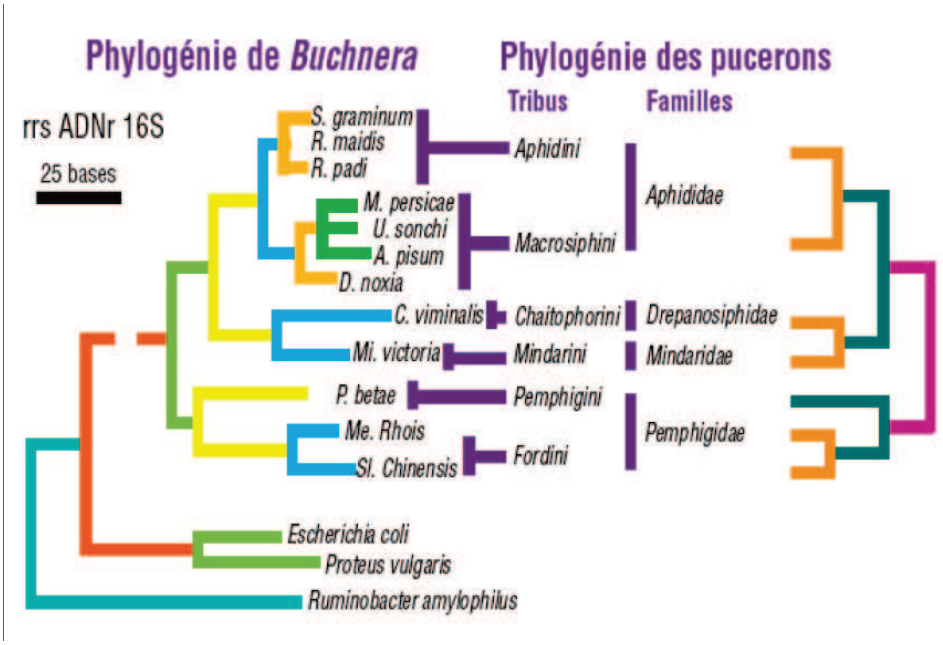


Figure 2. Arbres phylogéniques de *Buchnera* et des pucerons. L'arbre des *Buchnera* a été établi grâce à l'analyse des gènes codant les ARN ribosomal 16S par (Baumann *et al.,* 1995), alors que l'arbre des pucerons a été construit grâce à l'analyse des caractères morphologiques de fossiles de pucerons. Le parallélisme des deux arbres, en dépit de l'utilisation de méthodes distinctes confirment une co-spéciation des pucerons et des *Buchnera* (d'après (Rahbé *et al.,* 2007)).

2.1 Caractéristiques biologiques du puceron

Acyrthosiphon pisum appartient à la famille des Aphididae et à l'ordre des Hémiptères. Son cycle de reproduction annuel implique une alternance de générations parthénogénétiques (reproduction globale sans accouplement) et une génération sexuée, aboutissant à la production d'une forme de résistance, l'œuf d'hiver (**Figure 3**).

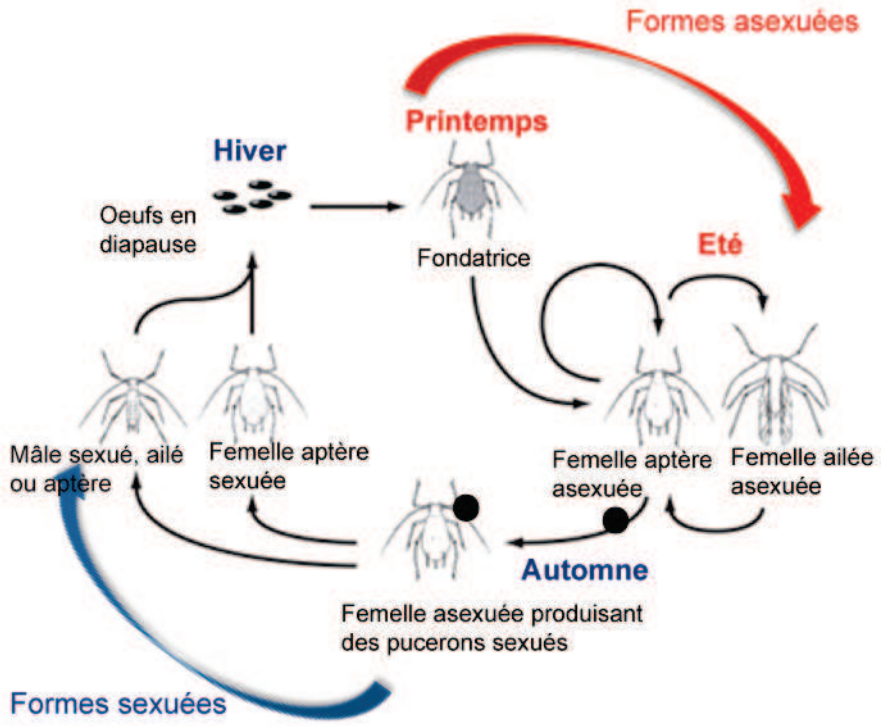


Figure 3. Schéma du cyle de vie du puceron du pois (d'après Simon et al. (2007) et Shingleton et al. (2003)). Les œufs, en arrêt de développement, éclosent au printemps pour donner des femelles parthénogénétiques. Ces femelles vont se reproduire par parthénogénèse durant la belle saison. Au début de l'automne les femelles parthénogénétiques produisent l'unique génération sexuée. Les mâles et les femelles de cette génération s'accouplent et pondent des œufs.

C'est grâce à la spécialisation de leurs pièces buccales que les pucerons peuvent se nourrir exclusivement de la sève phloémienne des plantes. C'est une longue formation flexible, le stylet, que le puceron insère dans les tissus de la plante jusqu'à atteindre les cellules phloémiennes. Riche en carbohydrates et tout particulièrement en saccharose et pauvre en plusieurs composés azotés, la sève phloémienne est un milieu hautement déséquilibré et variable, sa composition variant d'une plante à l'autre, ainsi qu'entre les parties d'une même plante. Comme pour la plupart des insectes, les besoins nutritionnels des pucerons varient durant leur développement.

Les cycles de parthénogenèse, lorsque les mères aptères produisent des larves vivantes, représentent les stades de développement et de reproduction les plus actifs des pucerons. Pendant cette période de télesco-

page de générations (*i.e.* les petits pucerons, futures femelles asexuées vivipares, contiennent déjà des embryons, qui eux mêmes contiennent des embryons, **Figure 4**), deux populations distinctes de bactéries coexistent dans les pucerons : les *Buchnera* maternelles et les *Buchnera* embryonnaires. La population embryonnaire constitue environ 75% des *Buchnera* du puceron (Humphreys and Douglas 1997). Le taux de croissance de *Buchnera* est maximal durant le développement embryonnaire du puceron, juste après sa contamination (Wilkinson *et al.,* 2003), alors que le nombre de *Buchnera* de l'adulte est presque parallèle au développement et décroît dans les vieux pucerons (Baumann *et al.,* 1995), suggérant ainsi un faible taux de croissance de la bactérie dans les bacteriocytes maternels.

2.2 Caractéristiques biologiques de *Buchnera*

Buchnera est l'une des bactéries intracellulaires obligatoires des insectes les plus étudiées (Baumann *et al.,* 1995). Trente-cinq ans après l'apparition du livre de Buchner (Buchner 1965), le premier génome de la *Buchnera* du puceron du pois *Acyrthosiphon pisum* (BAp) a été séquencé par l'équipe de Hajime Ishikawa à l'institut Ricken au Japon (Shigenobu *et al.,* 2000). Trois autres génomes de *Buchnera* provenant d'autres espèces de pucerons ont été publiés depuis : BSg du puceron des céréales *Schizaphis graminum* (Tamas *et al.,* 2002), BBp du puceron du pistachier *Baizongia pistaciae* (Van Ham *et al.,* 2003) et BCc du puceron du cèdre *Cinara cedri* (Perez-Brocal *et al.,* 2006). Récemment 7 autres génomes de BAp ont été séquencés grâce à la technique Solexa (Moran *et al.,* 2009), correspondant à différentes populations d'*A. pisum* dans le cadre d'une étude de dynamique de population et d'évolution moléculaire de la bactérie.

2.3 Interactions physiologiques entre les deux partenaires

L'approvisionnement des pucerons par *Buchnera* a été le sujet de multiples travaux de recherche (Baumann *et al.,* 1995). Des études expérimentales combinant des milieux artificiels de composition contrôlée et des pucerons traités aux antibiotiques (pucerons aposymbiotiques), ainsi qu'utilisant l'information génomique, se sont concentrées sur les composés azotés, et spécialement sur les acides aminés essentiels absents dans la sève phloémienne. Douglas *et al*. ont montré que la synthèse de la méthionine est assurée par BMp chez le puceron du pêcher, *Myzus persicae* (Douglas 1988), ainsi que la synthèse du tryptophane chez *Buchnera APS* (Douglas and Prosser 1992). Sur le même modèle Sasaki et Ishikawa (1995) ont montré que *Buchnera* est impliquée dans le transfert de l'azote pour la biosynthèse

des acides aminés essentiels depuis le glutamate ou la glutamine. Febvay *et al*. ont enfin montré que la biosynthèse de la thréonine, de l'isoleucine, de la leucine, de la valine et de la phénylalanine est réalisée par BAp (Febvay *et al.,* 1995; Liadouze *et al.,* 1996) montrant ainsi la spécialisation de *Buchnera* à produire les acides aminés nécessaires au puceron (Febvay *et al.,* 1999).

Globalement, on considère que 10 acides aminés essentiels sont synthétisés et fournis à l'hôte par *Buchnera* (Zientz *et al.,* 2001) (**Figure 1**). D'autre part, *Buchnera* profite des acides aminés non essentiels trouvés dans l'alimentation du puceron ou synthétisés par ce dernier.

Pour l'instant, la contribution métabolique de *Buchnera* à l'adaptation du puceron aux variations du milieu nutritionnel est connue seulement du point de vue qualitatif (Douglas 2006b). Dans la littérature, *Buchnera* est souvent considérée comme une bactérie en train de dégénérer, peu sensible aux stress extérieurs, plutôt incapable de réguler sa réponse transcriptionnelle et post-transcriptionnelle aux demandes variables de son hôte (Moran and Degnan 2006). De nouvelles idées sur la fonction symbiotique de la bactérie ont été obtenues grâce aux analyses systémiques devenues possibles depuis la publication de plusieurs génomes de *Buchnera* et de génomes d'autres endosymbiotes ainsi que grâce au développement de nouveaux outils d'analyse de réseaux (Cottret 2009).

2.4 Localisation de *Buchnera* dans le puceron et transmission entre les générations de pucerons

Buchnera est localisée à l'intérieur de l'abdomen du puceron à proximité des ovaires, dans des cellules spécialisées appelées bactériocytes (**Figure 4**). Les bactériocytes (60 à 80 par puceron) sont des cellules géantes, polyploïdes formant un organe appelé le bacteriome (**Figure 4**). Les *Buchnera* sont situées dans le cytoplasme des bactériocytes (**Figure 4**), entourées par une membrane d'origine eucaryote, le symbiosome. Dans les embryons, une fine couche de cellules syncitiales entourent le bacteriome, alors que dans les vieux pucerons, la dégénérescence nutritionnelle et reproductive est accompagnée par la dégénérescence de la structure des bactériomes (Baumann *et al.,* 1995).

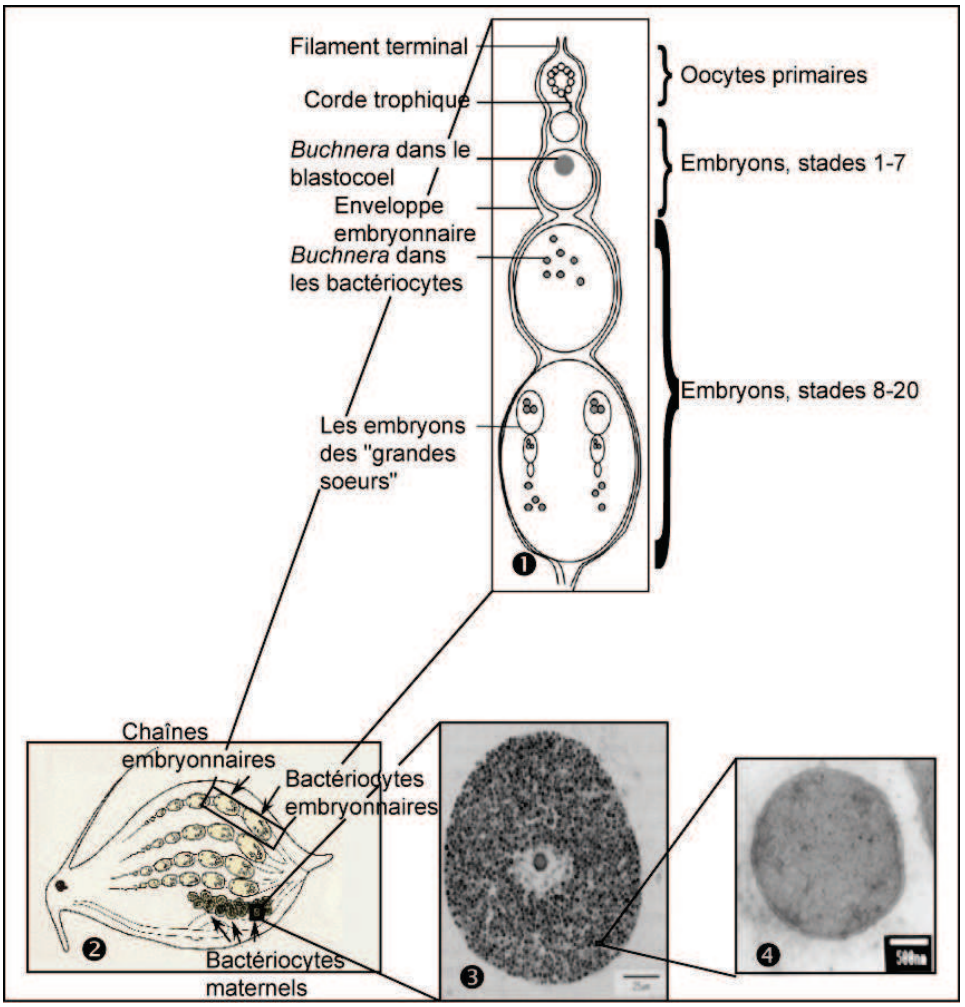


Figure 4. Répartition des bactériocytes à l'intérieur du puceron et localisation des *Buchnera* au sein des bactériocytes. Le deuxième cadre représente schématiquement la répartition des deux types de population de *Buchnera* dans la cavité abdominale du puceron : la population de *Buchnera* maternelles contenue dans le bactériome et la population des *Buchnera* embryonnaires au sein des chaînes embryonnaires (par T. Sasaki, http://buchnera.gsc.riken.go.jp). Le troisième cadre est une photo d'un bactériocyte maternel, on peut observer le noyau de la cellule eucaryote, situé au centre de la cellule et les *Buchnera* distribuées tout autour (par T. Fukatsu, http://buchnera.gsc.riken.go.jp). Le quatrième cadre est une photo d'une cellule de *Buchnera*, (par M. Morioka, http://buchnera.gsc.riken.go.jp). Enfin, le premier cadre représente schématiquement une ovariole du puceron (d'après (Bermingham and Wilkinson 2009)). On peut remarquer sur ce schéma le télescopage des générations, après le huitième stade de développement les embryons contenant eux-même des embryons.

La transmission de *Buchnera* d'une génération d'hôte à l'autre se fait verticalement et nécessite une phase d'infection à chaque génération. L'infection de l'embryon par des bactéries maternelles se produit durant la phase blastoderme, grâce à une ouverture au pôle postérieur de l'embryon dans les morphes vivipares, bien que les œufs soient contaminés également dans les morphes ovipares (Buchner 1965; Braendle *et al.,* 2003; Wilkinson *et al.,* 2003). Une conséquence directe de cette phase d'infection est que la population symbiotique passe par des goulots d'étranglement successifs, *i.e.*, seule une petite proportion de la population maternelle de symbiotes est transmise à la descendance. Wilkinson *et al*. (2003) ont montré que toutes les bactéries infectant un embryon proviennent d'un seul bactériocyte, les *Buchnera* entrant dans l'embryon via un conduit membranaire spécifique.

Ainsi, les populations de *Buchnera* ont des petits effectifs, subissent des forts goulots d'étranglement et sont isolées dans les bactériocytes empêchant tout échange de matériel génétique entre les individus. Cette dynamique de population particulière augmente drastiquement la probabilité de fixation des mutations légèrement délétères sans possibilité de correction, ce phénomène porte le nom de cliquet de Muller (Moran 1996). En effet, *Buchnera* est soumise à la dérive génétique et les conditions intracellulaires lui impose un biais compositionnel vers les bases A et T pour des raisons qui restent mal comprises : dérapage de la polymérase sur des fragments polyA selon Moran et al. (Tamas *et al.,* 2008) ou économique (à cause de la centralité de l'ATP chez *Buchnera*) selon Danchin (Rocha and Danchin 2002). Les taux d'évolution accélérés imposés par le cliquet de Muller sont visualisables au niveau des sites non synonymes chez *Buchnera* (Moran 1996). Cette accélération importante du taux d'évolution impose une augmentation du taux de mutation notamment à cause de l'absence de système de réparation et de la modification de la pression de sélection sur les gènes liée au milieu de vie intracellulaire de la bactérie (Moran 1996; Brynnel *et al.,* 1998; Clark *et al.,* 1999; Rispe and Moran 2000; Funk *et al.,* 2001; Moran *et al.,* 2009).

2.5 Les endosymbiotes secondaires et le remplacement de *Buchnera*

Comme la plupart des eucaryotes, les pucerons peuvent être considérés comme des micro écosystèmes impliquant des consortia de virus et de bactéries, interagissant dans des communautés. Une partie de cette diversité se trouve dans la flore du tractus digestif (Harada *et al.,* 1996), bien que les endosymbiotes intracellulaires hérités soient aussi très diversifiés chez le

puceron (Oliver *et al.,* 2006). Ainsi, les pucerons possèdent en plus des *Buchnera*, d'autres bactéries symbiotiques intracellulaires, appelées endosymbiotes secondaires. Ces endosymbiotes secondaires n'affectent pas en général toutes les espèces de pucerons, ni même tous les individus d'une même espèce. Ils sont transmis verticalement, bien que des transferts horizontaux puissent se produire entre des populations, et même des espèces de pucerons. Ces bactéries sont très diverses et incluent les genres *Spiroplasma*, *Wolbachia* et *Rickettsia* ainsi que plusieurs Entérobactéries (*Regiella insecticola, Hamiltonella defensa, Serratia symbiotica*). Plusieurs rôles fonctionnels des ces endosymbiotes secondaires ont été identifiés : la résistance thermique (Montllor *et al.,* 2002), l'adaptation de l'hôte à la plante (Tsuchida *et al.,* 2004), la résistance au parasitisme et la complémentation du rôle nutritionnel de *Buchnera* (Perez-Brocal *et al.,* 2006). *Hamiltonella defensa* peut envahir quelques bacteriocytes réservés généralement aux *Buchnera* (Moran *et al.,* 2005c) et contient des bactériophages actifs appelés APSE-1 et 2, codant des toxines qui pourraient être impliquées dans l'invasion des cellules de l'hôte ou leur protection face aux parasites (Moran *et al.,* 2005a). Comprendre la symbiose puceron-*Buchnera* et son évolution nécessite donc la considération de ces endosymbiotes secondaires, d'autant plus que des résultats expérimentaux suggèrent que *Buchnera* pourrait même être remplacée par ceux-ci dans certaines souches de pucerons en laboratoire (Koga *et al.,* 2003).

3 Le génome de Buchnera APS

3.1 Taille et évolution du génome de *Buchnera APS*

Les génomes des *Buchnera*, avec des tailles allant de 420 à 650 kb, sont parmi les plus petits génomes bactériens séquencés à ce jour (Gil *et al.,* 2002). Le génome de BAp est composé d'un chromosome circulaire d'environ 640kb et de deux plasmides, pLeu et pTrp, contenant les gènes nécessaires pour la biosynthèse de la leucine et deux gènes codant pour deux enzymes intervenant dans la voie de biosynthèse du tryptophane (Bracho *et al.,* 1994; Lai *et al.,* 1994).

Le génome de *Buchnera* contient environ 70% de bases AT (90% dans les régions intergéniques, **Tableau 1** ci-dessous). Des importants biais inter et intra-brins permettent de localiser l'origine et le terminus de la réplication.

Tableau 1. Composition des chromosomes de *Buchnera APS*, *Buchnera Bp*, *Buchnera Cc* et *E. coli*. *Buchnera Sg*, très proche de *Buchnera APS* n'a pas été reportée dans cette comparaison.

	BAp	BBp	BCc	Eco
AT total	73.69	74.66	79.90	49.21
AT codant	72.37	72.80	78.27	48.11
AT non-codant	84.77	83.63	91.57	57.93
Proportion du non-codant	10.65	17.19	12.30	11.16
Taille du génome	640681	615980	416380	4639675

3.1.1 Le chromosome de *Buchnera APS*

Les cellules de *Buchnera APS* sont hautement polyploïdes (Komaki and Ishikawa 1999). Elles ont un volume jusqu'à 15 fois plus large que celles d'*E. coli* et contiennent jusqu'à 10 fois plus d'ADN (Komaki and Ishikawa 1999). Les divisions incomplètes dues à des réplications et des ségrégations inefficaces pourraient être une des causes de la polyploïdie de *Buchnera APS* (Tamas *et al.,* 2002). Le nombre de copies de chromosome par *Buchnera* varie de 50 à 200 en fonction de l'âge du puceron (Komaki and Ishikawa 1999, 2000). Les nombres relatifs de plasmides et de copies de chromosome dans *Buchnera* ont probablement été à la base des transferts des gènes des voies de biosynthèse de la leucine et du tryptophane entre les plasmides et le chromosome durant l'évolution de *Buchnera* (Latorre *et al.,* 2005). Ce nombre relatif pourrait aussi être impliqué dans la régulation de l'expression des gènes, mais à ce jour aucune preuve n'a été apportée dans ce sens.

Le chromosome de *Buchnera* est très riche en bases A et T. Ce biais augmente la probabilité d'apparition des codons stop et Charles et al. (1999) ont montré que les séquences de *Buchnera* sont plus courtes par rapport aux gènes *d'E. coli*. De même, la composition en acides aminés des protéines de *Buchnera* est enrichie en acides aminés codés par des codons AT-riches (Charles *et al.,* 2006).

On observe un léger biais de répartition des gènes sur les deux brins du chromosome de *Buchnera* : 56% des gènes de BAp sont codés sur le brin direct. Lorsqu'on étudie la distribution des gènes essentiels, ce biais devient plus fort : 60% des gènes essentiels se trouvent sur le brin direct. Le lien entre l'expression des gènes et leur localisation sur le chromosome, leur organisation en opéron ou leur essentialité a été étudiée chez BAp (Viñuelas *et al.,* 2007). L'essentialité et l'organisation en opéron semblent avoir un impact significatif sur l'expression. Par contre, la localisation des gènes ne semble pas influencer leur niveau de transcription, ce qui est plutôt cohérent avec une bactérie à croissance lente (Viñuelas *et al.,* 2007). Lorsque l'organisation des unités de transcription est déduite par homologie avec *E. coli* aucun des régulateurs spécifiques ne semble avoir été conservé. Néanmoins, des résultats expérimentaux attestent une corrélation significative entre le niveau de transcription (basal ou différentiel) et l'organisation en unités de transcription putatives, suggérant un rôle fonctionnel de ce type de structures et leur conservation chez BAp (Reymond *et al.,* 2006; Viñuelas *et al.,* 2007). Ce travail et cette relation entre

l'expression des gènes et l'organisation du chromosome seront très largement repris dans cette thèse.

3.1.2 Les plasmides de *Buchnera APS*

Un grand nombre de *Buchnera* analysées à ce jour possèdent deux plasmides, pLeu et pTrp (Gil *et al.,* 2006). Chez *Buchnera APS* pLeu mesure 7786pb et contient 7 gènes : les 4 gènes de l'opéron de biosynthèse de la leucine (*leuABCD*), les gènes de la réplicase (*repA1*, *repA2*) et un gène codant pour une protéine membranaire (*yqhA*) (Silva *et al.,* 1998). Son plasmide tryptophane mesure 7258pb et contient 3 gènes (*trpE* et *trpG* répliqué et *trpG2*).

L'analyse des séquences des deux plasmides suggère des mécanismes de réplication différents (Latorre *et al.,* 2005; Gil *et al.,* 2006). Néanmoins des analyses plus amples sont nécessaires afin de confirmer cette hypothèse. Les hypothèses sur l'origine de ces plasmides convergent pour l'origine chromosomique (Gil *et al.,* 2006). L'histoire évolutive de ces deux plasmides est particulièrement dynamique, incluant de la recombinaison, une circulation des gènes entre le chromosome et le plasmide et une réorganisation originale pour chacune des quatre souches de *Buchnera* (Latorre *et al.,* 2005). Une hypothèse sur l'intérêt du transfert de ces gènes sur des plasmides serait la surproduction efficace, s'affranchissant de la régulation génomique (Latorre *et al.,* 2005), de ces deux acides aminés essentiels au puceron. Il a été démontré chez *Buchnera SG* que l'amplification des gènes localisés sur le plasmide pLeu et pTrp, est de 23.5 et respectivement 14.5 fois par rapport aux gènes situés sur le chromosome (Plague *et al.,* 2003). Néanmoins, ce rapport entre le nombre de gènes plasmidiques et le nombre de copies de gènes chromosomiques varie entre les espèces de puceron (Plague *et al.,* 2003). De plus, Birke et al. (2004) ont montré que le niveau d'amplification de pTrp ne semble pas corrélé avec les besoins nutritionnels du puceron. J. Viñuelas a montré au cours de sa thèse dans l'UMR BF2I que l'amplification du plasmide leucine est un mécanisme de régulation de la production de cet acide aminé en condition de demande variable du puceron en leucine (Viñuelas 2008).

3.2 Le biais intra-brin de composition en bases A et T et l'usage de code chez *Buchnera APS*

Les procaryotes montrent des compositions en bases G et C très variables, comprises entre 25% et 77% (Heddi *et al.,* 1998), reflétant les forces muta-

tionnelles et la pression de sélection associées aux processus faisant intervenir l'ADN (réplication, transcription, recombinaison, réparation et même traduction). Ce biais reflète la composition globale de la molécule d'ADN. Nous nous intéressons ici, au biais de composition à l'intérieur d'un même brin d'ADN. Le biais compositionnel brin-spécifique est surtout attribué à l'asymétrie de la réplication (Lobry 1996; Frank *et al.,* 2002; Lobry and Sueoka 2002; Rocha 2004). Il est plus fort au niveau des troisièmes positions de codon ainsi que dans les régions non-codantes (connues comme étant soumises à une moindre pression de sélection), par conséquent ce biais serait plus la conséquence d'un biais mutationnel que d'une sélection naturelle (Lobry and Sueoka 2002; Rocha 2004).

En règle générale, le brin direct est enrichi en G et en T, par rapport à C et A. Un des processus à l'origine de ce biais est la désamination. Les désaminations les plus fréquentes sont celles de la cytosine en uracile et de la cytosine méthylée en thymine. Ces réactions se produisent spontanément dans les conditions physiologiques normales. Le taux de la déamination est majoré par le stress oxydatif et les radiations ionisantes et il est ainsi 100 fois plus grand pour l'ADN simple brin dans la cellule que pour l'ADN double brin (Frederico *et al.,* 1990). Sachant que la réplication des deux brins de l'ADN se fait simultanément dans la cellule, et que l'élongation d'un brin d'ADN ne peut se faire par la polymérase que dans le sens 5'-3', l'un des deux brins doit être synthétisé de façon discontinue. Il s'agit du brin tardif ou indirect, qui est répliqué sous la forme de fragments d'Okazaki. Le brin avancé ou le brin direct, synthétisé de manière continue, est présent plus longtemps sous la forme de simple brin. Par conséquent le brin direct est beaucoup plus sujet à ce type de stress. Si la désamination n'est pas corrigée avant la réplication suivante, il y a une transition de la paire originale C:G vers T:A. Frank et Lobry considèrent la désamination de l'ADN simple brin comme la principale source du biais intra-brin (Frank and Lobry 1999). Néanmoins, le biais intra-brin AT est moins important que le biais intra-brin GC ce qui suggère qu'il doit y avoir d'autres causes (Rocha 2004).

Parmi les *Buchnera*, BBp et BCc montrent le biais compositionnel le plus important, ceci serait une conséquence d'une plus grande fréquence de désamination dans ces deux espèces à cause de la perte de plusieurs enzymes impliquées dans l'initiation et la ré-initiation de la réplication (Klasson and Andersson 2006). Ce biais est significativement différent par rapport aux autres espèces de *Buchnera* (Klasson and Andersson 2006). Une différence entre BBp et les autres *Buchnera* séquencées est la conservation chez BBp du gène codant l'endonucléase MutH (sous-unité du système de réparation des mésappariements). Néanmoins, les auteurs n'ont pas

pu faire le lien entre le biais compositionnel plus élevé et la présence de ce gène car le système de réparation des mésappariements ne semble pas fonctionnel chez *Buchnera* (les *Buchnera* ne possèdent pas de voie de méthylation), de plus le gène *mutH* n'est pas conservé chez toutes les γ-protéobactéries. Enfin, chez les γ-protéobactéries libres la présence/absence de ce gène ne semble pas être corrélée avec le biais compositionnel. Quelques autres gènes impliqués dans la réparation ont subi un décalage de phase de lecture, *i.e.* le gène *ung* impliqué dans la réparation de la désamination de la cytosine. Mais là encore, le lien ne peut pas être établi car ce gène a subi des décalages de phase de lecture chez BSg aussi. BBp a aussi perdu *priA*, *dnaT*, *topA*, *himA*, *himD* et *fis*. Ces sont des gènes intervenant dans la réplication et l'organisation de l'ADN mais aussi dans le redémarrage de la réplication lors des arrêts de la fourche de réplication. Les mutants *priA* de *E. coli* sont viables, néanmoins ils sont déficients en recombinaison et ils présentent une sensibilité accrue face aux enzymes endommageant l'ADN. Le mécanisme de réplication et de redémarrage de la réplication est sérieusement affecté, du coup il est possible que l'ADN de BBp passe plus longtemps à l'état simple brin lors des plus longs arrêts de la fourche de réplication. C'est d'ailleurs peut-être pour cette raison qu'il existe une sélection vers des génomes de petite taille chez les *Buchnera*, car ainsi il y aurait moins d'arrêts de la fourche de réplication par unité de génomique répliquée (Klasson and Andersson 2006).

Dans la plupart des bactéries, la sélection sur le niveau d'expression des gènes se reflète à travers l'usage des codons des séquences des gènes les plus exprimés. Ceux-ci utilisent les codons correspondant aux ARNt les plus abondants. L'usage de codons chez *Buchnera* est dominé par le fort biais en bases A et T. En effet, les codons majoritaires finissent systématiquement par A ou par T chez *Buchnera* (Rispe *et al.,* 2004). Néanmoins, il existe une corrélation entre l'abondance des ARNt isoaccepteurs et l'usage de codons soutenant l'hypothèse d'une pression de sélection s'exerçant sur l'expressivité des gènes (Charles *et al.,* 2006). Dans cette même étude, un usage de codon a de plus été observé sur des codons rares grâce à l'ensemble des gènes les plus exprimés déterminés expérimentalement chez BAp. Ces gènes privilégient les codons finissant par C à ceux finissant par G, lorsqu'on les compare à l'ensemble des gènes les plus faiblement exprimés. Ce choix favorise les codons optimaux du point de vue thermodynamique (Grantham *et al.,* 1981) pour certains acides aminés dans les gènes les plus exprimés chez *Buchnera* suggérant l'existence d'une sélection qualitative de la traduction au niveau de certains sites importants fonctionnellement pour les protéines essentielles à *Buchne-*

ra (Charles *et al.,* 2006). Les travaux récents de Toft et Fares (2009) ont confirmé cette hypothèse.

3.3 Dynamique d'évolution du génome de *Buchnera*

Le processus de réduction du génome a été amplement étudié dans la communauté scientifique. Plusieurs travaux ont déterminé indépendamment, mais d'une façon similaire, le contenu en gènes du dernier ancêtre entre *Buchnera* et *E. coli* (Shigenobu *et al.,* 2000; Moran and Mira 2001; Silva *et al.,* 2001). Van Ham *et al*. (2003) et Silva *et al*. (2003) ont estimé le contenu du dernier ancêtre commun symbiotique entre les quatre souches de *Buchnera* séquencées, afin de retracer le processus de réduction de chaque lignée (Latorre *et al.,* 2005). Différents modèles explicatifs de réduction génomique ont été proposés pour *Buchnera*. Moran *et al.* (2001) avancent l'hypothèse que la réduction a commencé par un processus très intense de pertes de larges portions du génome (durant lequel certains gènes essentiels aurait été perdus). Ce processus a été suivi par une érosion individuelle des gènes qui se poursuit encore actuellement. Une deuxième théorie, s'opposant à cette première a été proposée par Silva *et al.* (2001). Ils expliquent la réduction du génome par un seul type de mécanisme, des événements de désintégration des gènes dispersés dans tout le génome. Des études récentes soutiennent plutôt ce deuxième modèle (Dagan *et al.,* 2006; Delmotte *et al.,* 2006; van Hoek and Hogeweg 2007).

De par la quasi identité des génomes de BAp et de BSg et étant donné le manque de séquences d'insertions ainsi que d'autres éléments indispensables à la recombinaison homologue, Tamas *et al*. (2002) ont conclut que cette période dynamique de réduction aurait été suivie par une stase génomique depuis environ 50M d'années car *Buchnera* aurait atteint l'équilibre entre l'ensemble de gènes conservés et les besoins fonctionnels de la bactérie et de son hôte. Néanmoins, le processus de réduction génomique semble se poursuivre dans certaines lignées, notamment chez BCc qui a été séquencée récemment (Perez-Brocal *et al.,* 2006). Des travaux très récents argumentent la poursuite de ce processus d'érosion dans toutes les lignées de *Buchnera* à cause de la déficience des appareils de réplication et de réparation (Moran *et al.,* 2009).

Ces hypothèses sont également cohérentes avec les scénarios d'évolution de la taille des génomes décrits par Lawrence (2001). Un génome ne peut conserver qu'une quantité d'information finie (Lawrence 2001), la taille du génome étant ainsi une fonction dépendante des pressions de sélection (et donc de l'environnement et de la taille de la population) et du taux de recombinaison. Les *Buchnera* n'étant plus capables de

recombiner avec un environnement qui semble stable du moins chez BAp, BSg et BBp permettant la stabilité de leur génome. A l'inverse, dans l'environnement de BCc les forces de sélection sont encore en train d'évoluer à cause de la présence d'endosymbiotes secondaires ayant récupéré des fonctions essentielles, et expliquant ainsi la poursuite de la réduction du génome de *Buchnera*.

3.4 Le répertoire des gènes de *Buchnera*

3.4.1 Les gènes du métabolisme de *Buchnera*

Les premières recherches génomiques sur *Buchnera* ont été effectuées par l'équipe de Paul Baumann sur BSg (Baumann *et al.,* 1995). Les chercheurs avaient séquencé de larges fragments d'ADN et ainsi, ils avaient pu décrire les voies de biosynthèse de plusieurs acides aminés essentiels et souligner la conservation des gènes et la synthénie avec le génome d'*E. coli*. Ils avaient aussi émis l'hypothèse de la réduction du génome. L'analyse complète du génome de BAp par Shigenobu et al. (2000) a confirmé la réduction du répertoire de gènes et a permis une première analyse complète du métabolisme de la bactérie.

La partie la plus importante du métabolisme de *Buchnera* est concentrée autour du métabolisme du sucre (glucose ou mannitol) à travers la glycolyse et les voies du pentose phosphate. Cependant, le cycle de l'acide tricarboxylique (TCA) qui permet l'oxydation du pyruvate en dioxyde de carbone, reste incomplet chez *Buchnera*. Les flux de carbone et d'azote sont orientés majoritairement vers la synthèse des acides aminés essentiels. En effet, *Buchnera* a retenu dans son répertoire les gènes des enzymes nécessaires à la synthèse de la chaîne carbonée de l'arginine, la thréonine, la lysine, la valine, l'isoleucine, la leucine, le tryptophane, la phénylalanine et l'histidine. Ces 9 acides aminés essentiels (la méthionine n'est pas reportée dans cette liste faute de preuve expérimentale) sont synthétisés chez *Buchnera* grâce aux voies de biosynthèse classiques connues chez les bactéries, en utilisant comme précurseurs certains acides aminés non-essentiels, ainsi que l'α-kétoglutarate, le 3-phosphoglycérate, l'oxaloacétate, le pyruvate, le phosphoénol pyruvate, la 4-erythrose-phosphate et/ou le ribose-5-phosphate importée depuis les cellules hôtes ou synthétisés à partir du catabolisme du sucre. *Buchnera* a conservé l'ensemble complet des gènes de la voie de réduction du soufre et la capacité de synthétiser la cystéine, vraisemblablement en utilisant comme précurseurs la sérine et l'adésosine-5-

phosphosulfate (APS). Chez *Buchnera*, la biosynthèse de la méthionine peut se faire seulement à partir de l'homocystéine fournie par l'hôte.

Buchnera respire de façon aérobie et a perdu les gènes nécessaires à la fermentation et à la respiration anaérobie. Elle est capable de synthétiser l'ATP à partir de l'ADP grâce à un système F0F1 d'ATP-synthase fonctionnel. Actuellement, la question de la biosynthèse par *Buchnera* de ses propres nucléotides reste non résolue car certains enzymes nécessaires chez *E. coli* à l'interconversion des nucléotides entre leur formes desoxy ou mono-, di- et tri-phosphatées, ne sont pas retrouvés chez *Buchnera*. Une hypothèse serait que les enzymes de *Buchnera* auraient une plus large spécificité et que l'importation directe de certains précurseurs à partir de la cellule hôte pourraient quand même permettre une synthèse des nucléotides propres à *Buchnera* (Zientz *et al.,* 2004).

Buchnera a retenu l'ensemble des gènes nécessaires à la synthèse de l'acide lipoïque, du FAD (à partir de la riboflavine néosynthétisée) et du NAD (à partir du nicotinate importé de la cellule hôte), ainsi que l'ensemble des gènes pour la synthèse de la Biotine. Parmi les autres voies des cofacteurs et des groupes prosthétiques, les gènes codant pour le folate, le pantothénate, le coenzyme A, l'ubiquinone, la thiamine, la pyridoxine et le protohème ont été perdus. Ainsi *Buchnera* est dépendante de son hôte pour la synthèse de ces derniers composés (Shigenobu *et al.,* 2000).

Buchnera a perdu les gènes pour la gluconéogénèse, l'entier appareil enzymatique de biosynthèse des composantes de la surface cellulaire (lipopolysaccharides et phospholipides comprises) ainsi que les gènes régulateurs et les gènes impliqués dans la défense de la cellule (Shigenobu *et al.,* 2000).

3.4.2 Les gènes de l'appareil du métabolisme de l'ADN chez *Buchnera*

L'analyse génomique de *Buchnera* a mis en évidence la bonne conservation de l'appareil de réplication nécessaire à l'initiation et le processus même de réplication : les gènes codant les sous-unités de l'ADN polymérase III (*dnaE*, *dnaN*, *dnaQ*, *dnaX*, *holA*, *holB*) sont conservés ; de même que ceux codant l'hélicase (*dnaB*), la primase (*dnaG*), les deux sous-unités de la gyrase (*gyrA* et *gyrB*), la ligase (*lig*), ou encore le gène codant la protéine de liaison à l'ADN simple brin (*ssb*) (Gil *et al.,* 2004; Perez-Brocal *et al.,* 2006). Par contre, les protéines nécessaires à la finalisation de la réplication, comme la protéine Tus, ou à la séparation des chromosomes (ParC, ParE, TopA) ne sont pas codées dans les génomes de *Buchnera*, ou du moins pas systématiquement chez les quatre *Buchnera* séquencées.

La perte des systèmes de réparation et de recombinaison est une caractéristique des bactéries à génome réduit. *Buchnera* a une capacité fortement limitée de réparation et de recombinaison, étant dépourvue du gène essentiel à la recombinaison, *recA*, ainsi que du système de réparation *uvrABC* (Moran and Mira 2001). Elle manque aussi de gènes de restriction et de méthylation.

Le système de traduction est complet, bien que minimaliste avec 32 ARNt, 55 protéines ribosomales, 12 facteurs de traduction et 21 ARNt amino-acyl transférases (Charles *et al.,* 2006). On a constaté que *Buchnera* possède une seule copie pour chacun des gènes des ARNr, alors que chez *E. coli*, on en compte sept. C'est une propriété plutôt caractéristiques des organismes à faible taux de croissance.

3.4.3 Les gènes du transport et les gènes du flagelle chez *Buchnera*

Bien que la fonction symbiotique implique un intense échange de composés, actuellement nous disposons d'un très modeste ensemble de transporteurs annotés chez *Buchnera* par rapport aux bactéries de forme libre. Pour le transport de petites molécules *Buchnera* a conservé quelques gènes des transporteurs ABC (un des systèmes de transport au travers de la membrane cellulaire des plus présents chez de nombreuses espèces bactériennes (Tomii and Kanehisa 1998)), un système de transport du sucre PTS (Phosphoenolpyruvate-carbohydrate-phosphotransférase) pour le mannitol et le glucose, un système porine, ainsi que quelques transporteurs non-spécifiques.

On note chez *Buchnera* la conservation de la majorité des gènes codant pour les composants protéiques du flagelle. Les gènes codant pour le filament et la partie terminale du flagelle, *fliC* et *fliD*, ont été perdus ce qui indique que *Buchnera* est une bactérie immobile (Shigenobu *et al.,* 2000). Parmi les 16 gènes du flagelle, 11 font partie chez *Buchnera APS* des gènes fortement exprimés et à évolution accélérée (Viñuelas *et al.,* 2007). Il a été proposé que ces gènes sont essentiels dans le cadre de l'association symbiotique ayant acquis une fonction de transport de protéines (Shigenobu *et al.,* 2000; Viñuelas *et al.,* 2007; Toft and Fares 2008) comme cela a déjà été observé chez *Salmonella typhimurium* et *Yersinia enterocolitica* utilisant l'appareil flagellaire pour la sécrétion de protéines et de facteurs de virulence lors des phases d'infection (Kubori *et al.,* 1998; Young *et al.,* 1999). Majander et al. (2005) ont montré chez les mutants *fliC* et *fliD* d'*E. coli*, la capacité de la bactérie à utiliser son système flagellaire comme transporteur de peptides. Enfin, Maezawa et al. (2006) ont

montré chez *Buchnera APS* la présence de nombreux corps basaux du flagelle incorporés dans la membrane cellulaire.

Les gènes du flagelle pourraient donc tenir un rôle capital dans le fonctionnement et le maintien de la symbiose *Buchnera*/puceron par leur rôle de transporteur.

Il se peut aussi que le répertoire réduit de gènes de *Buchnera* puisse être complété par des gènes de l'hôte qui auraient été transférés de la bactérie vers le puceron après l'intégration symbiotique, comme dans le cas des mitochondries. Néanmoins, à l'opposé des mitochondries qui ont évolué à partir d'un hôte monocellulaire, *Buchnera* a peu de contacts avec la lignée germinale de son hôte, excepté des courtes périodes après l'infection de l'œuf. Les premières études réalisées en ce sens montrent qu'a priori très peu de gènes sont susceptibles d'avoir été transférées de *Buchnera* vers son hôte (IAGC 2010).

4 La régulation de l'expression des gènes chez les procaryotes

La transcription, c'est à dire le processus au cours duquel un ARN messager (ARNm) est synthétisé par l'ARN polymérase à partir d'une matrice ADN, se déroule en plusieurs étapes : (1) la pré-initiation, lorsque l'holoenzyme reconnaît le promoteur du gène et s'associe à l'ADN, dont les brins sont séparés, permettant ainsi l'assemblage du complexe de transcription ; (2) l'initiation, correspondant à la biosynthèse des 10 premières bases du transcrit environ et à la fin de laquelle le facteur σ est libéré ; (3) l'élongation, correspondant à la poursuite de la synthèse par la polymérase du transcrit ; (4) la terminaison qui peut se faire de deux façons, soit grâce à la présence d'un terminateur (séquence formant une structure d'épingle à cheveux, terminaison Rho-indépendante), soit grâce à l'intervention d'un facteur de terminaison Rho (ou terminaison Rho-dépendante). On considère que la fréquence à laquelle le processus de transcription a lieu est régulé majoritairement durant les étapes précoces de préinitiation et d'initiation pour des raisons d'économie et d'optimisation de l'énergie cellulaire. Néanmoins, cela ne signifie pas que la régulation du niveau de transcrits

présents dans la cellule se fait exclusivement par la régulation de l'initiation, car d'autres mécanismes peuvent par exemple réguler les concentrations des ARNm et leurs stabilité, comme nous allons le voir par la suite.

La séquence du promoteur, les protéines régulatrices et les sites opérateurs vont « décider » si le gène est accessible à la transcription et avec quelle force (fréquence) il sera transcrit.

4.1 La régulation de l'initiation de la transcription chez les procaryotes

4.1.1 L'ARN polymérase, les facteurs σ et leurs promoteurs

L'ARN polymérase constitue la pièce centrale du processus de transcription. L'apoenzyme est formée de 5 sous-unités : $\beta\beta'\alpha_2\omega$ (**Figure 5**). Le site catalytique est contenu dans les sous-unités β et β'. Le domaine N terminal des sous-unités α (αNTD) est responsable de l'assemblage des sous-unités β, alors que le domaine C terminal des sous-unités α (αCTD) est un domaine de liaison à l'ADN. Enfin la sous-unité ω n'a pas un rôle direct dans la transcription, mais un rôle indirect de chaperonne des sous-unités β, elle ne participe donc pas à l'assemblage final (Browning and Busby 2004).

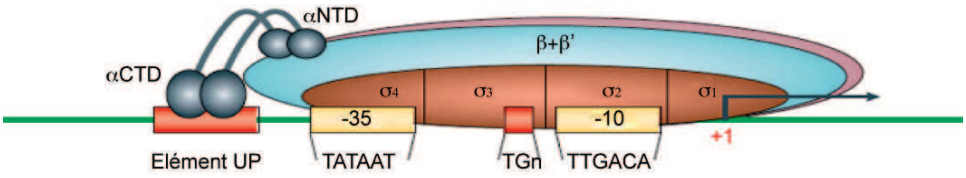


Figure 5. Schéma de l'ARN polymérase et de ses interactions avec les différentes parties du promoteur (d'après Browning et al. (2004)). Le double brin d'ADN est représenté en vert, avec les deux parties des promoteurs σ^{70}, la partie -35 et la partie -10 (pb en amont du site d'initiation de la transcription) soulignées en jaune. Deux autres éléments importants des promoteurs, découverts plus récemment, sont la partie -10 étendue – l'élément TGn et l'élément UP, tous les deux soulignés en rouge. Les sous-unités constituant l'ARN polymérase sont représentées avec différentes couleurs : les sous-unités β en cyan et rose, la sous-unité α avec ces deux domaines αCTD et αNTD, est colorée en gris, et les différents domaines de la sous-unité σ sont colorés en marron, la sous-unité ω de la polymérase n'apparaissant pas sur ce schéma. Comme on peut le constater sur le schéma, le deuxième domaine de la sous-unité σ interagit avec l'élément -10 des promoteurs, le troisième – avec l'élément TGn des promoteurs et le quatrième

domaine de la sous-unité σ interagit avec l'élément -35 des promoteurs, l'élément UP étant reconnu par le domaine αCTD de la sous-unité α.

L'apoenzyme est capable de synthétiser de l'ARN mais pas d'initier la transcription. Pour ce faire, elle doit d'abord se lier à un facteur σ et former l'holoenzyme. Le facteur σ est une sous-unité essentielle de l'ARN polymérase, permettant l'association de la polymérase à l'ADN juste en amont du site d'initiation de la transcription, en orientant la transcription (le sens de la lecture) et en aidant la séparation des brins d'ADN (Browning and Busby 2004). Il existe plusieurs types de facteurs σ, chacun permet de reconnaître un seul type de promoteurs (**Tableau 2**), ainsi en changeant de facteur σ dans la cellule, il est possible de changer l'ensemble des gènes transcrits et ainsi, de « programme » d'expression. En dépit de la différence de séquences et des distances entre les deux parties des différents promoteurs σ, les deux composantes des promoteurs restent centrées approximativement autour des positions -35 et -10 par rapport au site d'initiation de la transcription (**Tableau 2**). Deux autre élément importants, constituant les promoteurs σ^{70} ont récemment été découverts : l'élément -10 étendu (TGn) et l'élément UP. L'élément TGn (**Figure 5**) est un motif de 3 à 4 pb situé en amont de l'élément -10 (Sanderson *et al.,* 2003), 20% des promoteurs d'*E. coli* possédant cet élément (Burr *et al.,* 2000). L'élément UP (**Figure 5**) est un motif très AT-riche, situé à 20 pb environ en amont de l'élément -35 (Ross *et al.,* 2001) et pouvant augmenter la force du promoteur jusqu'à 100 fois (Estrem *et al.,* 1998). Le facteur σ^{70} est le seul facteur σ bactérien constitutif, dirigeant l'activité de l'ARN polymérase dans les conditions normales de croissance. Les autres facteurs σ permettent aux bactéries de répondre à divers stress auxquels elles doivent faire face. Chez *E. coli*, 7 facteurs σ ont été décrits : σ^{70}, le facteur constitutif ; σ^{38}, le facteur de phase stationnaire, impliqué dans la réponse aux stress oxydatif, thermique, acide et osmotique ; σ^{32}, responsable de la transcription des protéines du choc thermique ; σ^{28}, associé à la transcription des gènes du flagelle ; σ^{24}, impliqué dans le maintient des enveloppes cellulaires lors du stress thermique ; σ^{19}, responsable du captage et du transport du fer et σ^{54} nécessaire à la transcription des gènes impliqués dans le métabolisme de l'azote (**Tableau 2**).

Chez *E. coli*, un mécanisme de mise a disposition rapide des facteurs σ, qui doivent intervenir lors des changements brusques de conditions de vie a été décrit et est réalisé grâce aux facteurs anti-σ. Ces facteurs anti-σ permettent de maintenir inactifs les facteurs σ alternatifs, présents dans la cellule *e.g.* FlgM, les facteurs anti-σ de σ^{28}. Un autre mécanisme peut être illustré par le facteur σ^{32} qui en condition normale de croissance est as-

socié de façon stable à son récepteur DnaK (hsp70). L'augmentation de la température va modifier la conformation de DnaK et diminuer fortement son affinité pour le σ^{32}. Le σ^{32} se trouve ainsi libéré et la cellule peut transcrire les gènes du régulon de choc thermique (Chattopadhyay and Roy 2002).

Tableau 2. Facteurs σ décrits chez *E. coli*.

Facteur σ	Promoteur	Fonction
σ^{70}	TTGACA-(N_{16-18})-TATAAT	Constitutif
σ^{54}	CTGGNA-(N_6)-TTGCA	Métabolisme de l'azote
σ^{38}	TTGACA-(N_{16})-CTATACT (Lacour *et al.,* 2003)	Phase stationnaire
σ^{32}	CTTGA-(N_{14})-GNCCCCATNT (Wang and deHaseth 2003)	Choc thermique (stress cytoplasmique)
σ^{28}	CTAAA-(N_{15})-GCCGATAA	Flagellaire
σ^{24}	GGACTT-(N_{15-19})-GTCWAA (Rhodius *et al.,* 2006)	Choc thermique (stress périplasmique)
σ^{19}	GAAAAT-(N_{15})-TGTCCT (Enz *et al.,* 2003)	Transport du fer

σ^{70} se fixe a deux hexamères en position -35 et -10, dont les séquences consensus sont TTGACA et TATAAT (Browning and Busby 2004). La plupart des promoteurs σ^{70} conservent huit nucléotides des 12 de la séquence consensus, néanmoins 10% des promoteurs d'*E. coli* ne conservent que 50% de la séquence consensus et sont pourtant fonctionnels (Mendoza-Vargas *et al.,* 2009). Des études récentes ont mis en évidence deux nouveaux éléments constitutifs des promoteurs avec lesquels chacun des quatre domaines du facteur σ^{70} interagit (Murakami *et al.,* 2002a; Murakami *et al.,* 2002b). Il s'agit de l'élément -10 étendu, de 3 à 4 pb, situé juste en amont de l'élément -10 (TGn) et reconnu par le domaine 3 du facteur σ, et de l'élément UP, situé 20pb environ en amont de l'élément -35 et qui est reconnu par le domaine C-terminal de la sous-unité α. Les éléments -10 et -35 des promoteurs interagissent avec les domaines 2 et 4 respectivement du facteur σ^{70} (**Figure 5**).

La polymérase est peu disponible dans la cellule, de plus la majorité est utilisée pour la synthèse des ARN non-codants (essentiellement des ARNr). Environ 90% des ARN synthétisés dans *E. coli* cultivée en milieu riche sont des ARN stables (ARNr et ARNt) (Dennis *et al.,* 2009). Le reste de polymérase disponible se distribue de façon non homogène entre les gènes restants en fonction de : (1) la séquence ADN de leur promoteur, (2) du facteur σ reconnaissant leur promoteur, (3) des petits ligands, (4) des facteurs de transcription et de la distribution de leurs sites de fixation à l'ADN et (5) de la structure du chromosome (Browning and Busby 2004).

Les facteurs σ, leurs promoteurs et leurs facteurs anti-σ, représentent un mécanisme essentiel de la transcription, mais en même temps une capacité rapide de régulation et d'adaptation du programme d'expression des gènes en fonction des conditions de vie.

4.1.2 Les petits ligands

Les petits ligands sont un autre moyen de répondre au niveau de l'expression des gènes aux changements environnementaux rapides. Le cas le plus connu est l'alarmone, ou le guanosine 5',3' biphosphate (ppGpp). L'alarmone apparaît en cas de pénurie d'acides aminés (Chatterji and Ojha 2001). En se fixant l'ARN polymérase durant l'initiation de la transcription, l'alarmone déstabilise le complexe ouvert de transcription et empêche la transcription des gènes (Barker *et al.,* 2001), spécialement des gènes codant pour des produits nécessaires à la machinerie cellulaire de traduction (les gènes des ARNr, et la majorité des gènes des ARNt) (Srivatsan and Wang 2008). Il a également été découvert que ppGpp fonctionne avec le co-facteur DksA chez *E. coli* ainsi que chez d'autres bactéries (Paul *et al.,* 2004).

Récemment, une autre alarmone a été découverte, il s'agit de l'ARN 6S qui agit comme inhibiteur compétitif de la transcription en se liant au quatrième domaine de la sous-unité σ^{70} de l'ARN polymérase, ce qui empêche cette région de se lier à l'élément -35 des promoteurs (Cavanagh *et al.,* 2008).

4.1.3 Les facteurs de transcription et les sites opérateurs

Plus de 300 facteurs de transcription ont été décrits chez *E. coli* (Gama-Castro *et al.,* 2008). Parmi ceux-ci, plus de 50% ne régulent la transcription qu'au niveau d'un seul promoteur (ils sont spécifiques à un seul gène). Un problème majeur des études *in silico* portant sur les sites de fixation à l'ADN des facteurs de transcription est la prédiction de leur fonctionnalité. En effet, celle-ci peut difficilement être déduite seulement à partir de la séquence ADN de ces sites, les facteurs de transcription tolérant très souvent des mutations. D'autre part, l'inférence de cette fonctionnalité à d'autres organismes même phylogénétiquement proches est encore plus problématique, car les sites de fixation à l'ADN des facteurs de transcription sont connus pour évoluer à des très grandes vitesses (Rodionov *et al.,* 2004).

Les sites de fixation ont une taille comprise entre 12 et 30 pb et peuvent être constitués de répétitions ou peuvent être des palindromes, in-

diquant alors une activité multimérique des facteurs de transcription. Il a été remarqué que lorsqu'un facteur de transcription régule la transcription d'une unité de transcription, plusieurs sites de fixation vont généralement lui être associé pour la régulation du même promoteur (van Hijum *et al.,* 2009).

Le contexte génomique des sites de fixation des facteurs de transcription est très important pour leur fonctionnalité. Par exemple, la composition des séquences avoisinant le site de fixation va avoir une influence sur l'affinité du facteur de transcription pour son site. Ainsi, l'affinité des séquences avoisinantes va influencer la vitesse à laquelle le FT trouvera sa cible par diffusion le long du chromosome. Aussi, la composition des séquences avoisinantes peut affecter la demi-vie de l'association facteur de transcription - site de fixation, si le facteur de transcription a une affinité plutôt importante pour les séquences avoisinante, il diffuse plus rapidement (Mayo *et al.,* 2006).

Un facteur de transcription se distingue en fonction de son action sur la transcription du gène en activateur s'il augmente ce niveau, en répresseur, s'il le diminue. Enfin un facteur de transcription est dual, s'il est capable d'activer ou de réprimer le gène en fonction des facteurs de transcription secondaires interagissants. La position des sites de fixation des facteurs de transcription par rapport au promoteur du gène régulé peut parfois nous informer sur le type de régulation (activation ou inhibition) du facteur de transcription. Ainsi, les sites des répresseurs peuvent se situer très en amont du site d'initiation de la transcription (même si la plupart de ces sites sont situés entre les positions -60 et +60) (van Hijum *et al.,* 2009). A l'inverse, les sites des activateurs se situent le plus souvent en amont et dans un voisinage proche du promoteur (van Hijum *et al.,* 2009). Une autre contrainte concernant les sites de fixation fonctionnels est que le facteur de transcription doit se lier à l'ADN du même côté que l'ARN polymérase, donc à des distances multiples de 10.5 bases (le pas de l'hélice d'ADN).

4.1.4 Les activateurs de la transcription chez les procaryotes

Browning et al. (2004) distinguent 3 types d'activation (**Figure 6**) :

I. L'activateur possède un site de fixation situé en amont de la position -35, et il recrute la polymérase au niveau du promoteur en interagissant directement avec l'αCTD ;
II. L'activateur se lie à un site opérateur chevauchant la partie -35 du promoteur et interagit en même temps avec le domaine 4 du facteur σ, ceci permet aussi de recruter la polymérase au niveau du promoteur ;

III. L'activateur se lie en général au voisinage du promoteur, et son association à l'ADN entraîne en changement de conformation du promoteur le rendant plus accessible à la polymérase. Un exemple de changement de conformation est le changement de la distance apparente entre les éléments -35 et -10 du promoteur, lorsque cet espace est trop grand, le promoteur est difficilement reconnu par le facteur σ, en tordant la séquence entre -35 et -10 la distance diminue et le promoteur est plus facilement reconnu ;

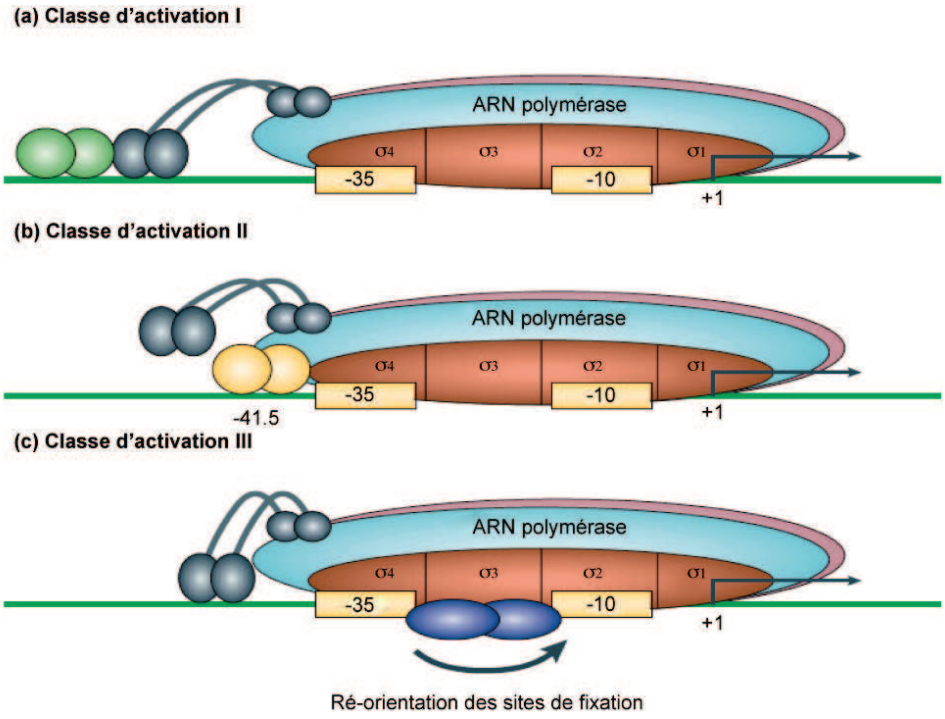


Figure 6. Trois classes d'activateurs définies en fonction de leur mécanisme de fonctionnement au niveau du promoteur. Les activateurs sont toujours représentés sur ce schéma sous forme de dimères (en vert, jaune et bleu) car la plupart des activateurs connus fonctionnent sous cette forme. A) La classe d'activation I. L'activateur se fixe sur un site en amont du promoteur et recrute l'ARN polymérase en interagissant avec le domaine C-terminal de sa sous-unité α (αCTD). B) La classe d'activation II. L'activateur se fixe sur un site adjacent à l'élément -35 du promoteur et interagit avec le domaine quatre du facteur σ. C) La classe d'activation III. L'activateur se fixe sur un site très proche voire chevauchant le promoteur, en modifiant la distance apparente entre les éléments -10 et -35 et permettant ainsi à l'ARN polymérase de reconnaître le promoteur (d'après Browning et al. (2004)).

Nous devons enfin citer un type supplémentaire d'activation de la transcription d'un gène, qui n'est pas médié par un facteur de transcription, mais indirectement par l'ARN polymérase qui modifie localement le niveau de surenroulement de certains promoteurs, évènement qui modifie le niveau de transcription de certains gènes (Dorman 2008) un exemple est donnée dans la **Figure 7**.

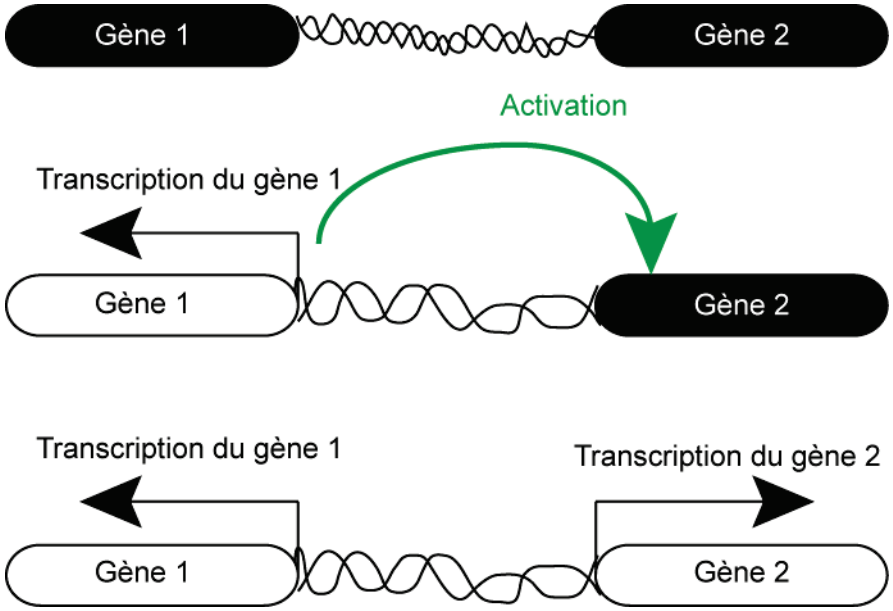


Figure 7. Activation des promoteurs voisins par modulation locale du surenroulement de l'ADN. Lorsque le gène 1 est transcrit, cela conduit à l'augmentation du niveau de surenroulement négatif de la région intergénique en amont, contenant le promoteur du gène 2, et ainsi à l'initiation de la transcription du gène 2 (Dorman (2008)).

4.1.5 Les inhibiteurs de la transcription chez les procaryotes

Trois types de répression de la transcription ont été décrits (Browning and Busby 2004) (**Figure 8**) :

I. Le répresseur se lie à l'ADN sur une région chevauchante au promoteur, empêchant l'association de la polymérase avec le promoteur, ou empêchant le déroulement de l'initiation de la transcription ;
II. Plusieurs molécules du répresseur se lient à l'ADN entrainant la formation d'une boucle d'ADN, qui à son tour empêche l'initiation de la transcription ;

III. Le répresseur module l'activité d'un activateur, en se liant sur un site chevauchant le site de liaison de l'activateur ;

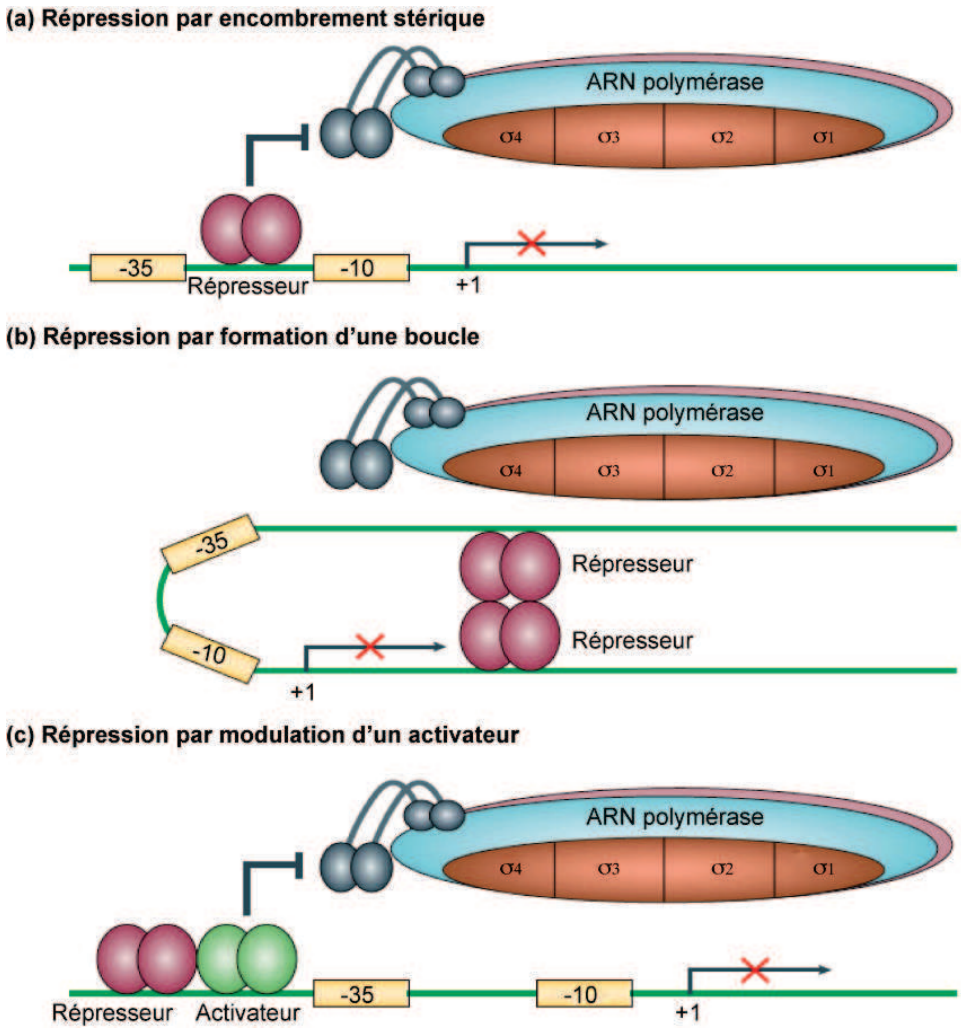


Figure 8. Trois classes de répresseurs définies par Browning et al. (2004). A) La répression par encombrement stérique. Le répresseur se fixe sur un site chevauchant le promoteur, empêchant ainsi l'ARN polymérase de reconnaître le promoteur. B) La répression par formation de boucle. Le répresseur se fixe sur des sites distants et les interactions entre les répresseurs induisent la formation d'une boucle piégeant le promoteur. C) La répression par la modulation de l'activité d'un activateur. Le répresseur se fixe sur un site proche du site de l'activateur, interagit avec ce dernier et empêche donc l'activateur de recruter l'ARN polymérase au niveau du promoteur (d'après Browning et al. (2004)).

Un quatrième type de répression est l'interférence de la transcription définie par Shearwin et al. (2005) comme l'influence négative d'un

processus de transcription, directement, en *cis* sur un deuxième processus de transcription. L'interférence de la transcription est asymétrique et résulte de l'association de deux promoteurs, un promoteur fort (ou agressif) qui inhibe la transcription au niveau du deuxième promoteur, le promoteur faible (ou sensible). Les orientations respectives des deux promoteurs peuvent être de trois types : convergente, tandem ou divergente. Plusieurs mécanismes d'interférence ont été décrits : (i) la compétition des promoteurs, l'occupation du promoteur fort par l'ARN polymérase prévient l'occupation du deuxième promoteur ; (ii) « sitting duck » dans le cas d'une paire de promoteurs convergents ou en tandem, la transition lente du complexe ouvert formé par l'ARN polymérase au niveau du promoteur faible vers le complexe d'élongation, permet au complexe d'élongation du gène à promoteur fort de déloger le complexe du gène à promoteur faible et d'annuler sa transcription ; (iii) l'occlusion du promoteur qui peut arriver dans le cas de deux promoteurs convergents dont les transcrits se superposent, dans ce cas la circulation des ARN polymérases lors de la transcription à partir du promoteur fort empêche la fixation de la polymérase sur le promoteur faible ; (iv) la collision des complexes d'élongation dans le cas des promoteurs convergents ;(v) le « roadblock », le complexe ouvert au niveau d'un promoteur bloque le passage de la polymérase en élongation à partir d'un autre promoteur (Shearwin *et al.,* 2005).

4.2 La régulation de la transcription via les opérons

L'organisation en opérons[4] est l'un des mécanismes fondamentaux de la régulation de l'expression des gènes chez les procaryotes, permettant une co-régulation et une co-expression synchrone d'un groupe de gènes. Les opérons sont des cas particuliers des unités de transcription[5] d'un génome, ce sont des unités de transcription polycistroniques.

Plusieurs hypothèses ont été proposées pour expliquer la pression de sélection positive vers la conservation des opérons dans les génomes bactériens. Une hypothèse naturelle est évidemment que ces structures permettent une (co-)régulation économique et fiable (Salgado *et al.,* 2000).

[4] On appelle *opéron* un groupe de gènes transcrits dans un unique ARNm (polycistronique) à partir d'un même promoteur.

[5] L'*unité de transcription* est une séquence d'ADN exprimée par l'intermédiaire de la synthèse d'une seule molécule d'ARNm. Elle peut contenir un seul gène – unité de transcription *monocistronique*, ou plusieurs gènes – unité de transcription *polycistronique*.

Une autre hypothèse avancée est que l'organisation en opéron serait très avantageuse dans le transfert horizontal (Lawrence and Roth 1996). Lors d'un transfert, l'ensemble des gènes de l'opéron est transféré (et donc potentiellement un nouveau phénotype), ce qui augmenterait les chances de conservation de ces gènes dans le génome d'accueil.

Les opérons sont des structures très dynamiques. Les transferts horizontaux ainsi que les réarrangements génomiques sont des processus gouvernant le cycle de vie des opérons. L'analyse de Price et al. (2006) a mis en évidence le fait que l'apparition des opérons est un processus fréquent.

La caractérisation expérimentale exhaustive des unités de transcription est la méthode la plus fiable, mais elle reste hors d'atteinte en termes de temps de travail et de coût, même pour les génomes de petite taille. Par conséquent la seule solution envisageable pour reconstruire la carte opéronique d'un organisme est la prédiction *in silico* des unités de transcription.

Toutes les méthodes de prédiction des unités de transcription ont un cadre universel de travail. L'ensemble du génome est divisé en paires de gènes adjacents. Les paires de gènes adjacents n'ayant pas la même direction de transcription sont éliminées, car parmi tous les opérons connus à ce jour, aucun cas ayant des gènes avec différentes directions de transcription n'a été observé. Les autres paires de gènes adjacents et ayant la même direction de transcription sont étiquetées soit comme paires de gènes opéroniques ou comme paires de gènes non-opéroniques. Les différentes méthodes de prédiction se distinguent par les propriétés des paires de gènes adjacents et l'approche mathématique qu'elles vont employer pour faire la distinction entre les paires de gènes opéroniques et non-opéroniques.

Un des travaux pionniers de l'étude et de la comparaison des propriétés des paires de gènes adjacents opéroniques et non-opéroniques, a été réalisé chez *E. coli* par Salgado *et al*. (2000). Les auteurs ont montré que certaines propriétés comme la distance intergénique et les fonctions communes de ces paires de gènes peuvent être utilisées pour une prédiction fiable (environ 80% de précision) des unités de transcription. Néanmoins, l'estimation de précision a été faite sur l'ensemble même qui a servi à l'entraînement du classificateur, ce qui conduit à une estimation trop optimiste de cette précision (Romero and Karp 2004). Ce problème est très fréquent dans les divers travaux de prédiction d'unités de transcription, car les seuls organismes pour lesquels on dispose de l'ensemble des opérons confirmés expérimentalement sont *E. coli* et *B. subtillis*.

Les propriétés qui ont été utilisées dans la littérature jusqu'à présent pour la prédiction des unités de transcription sont les suivantes :

- la distance intergénique (Salgado *et al.,* 2000; Moreno-Hagelsieb and Collado-Vides 2002; Romero and Karp 2004) ;
- l'annotation fonctionnelle des gènes (Zheng *et al.,* 2002; Strong *et al.,* 2003; Chen *et al.,* 2004; Romero and Karp 2004) ;
- les données de régulation : les promoteurs et d'autres motifs intervenant dans la transcription (Yada *et al.,* 1999; Craven *et al.,* 2000; Tjaden *et al.,* 2002) ;
- les clusters de gènes conservés à travers différents génomes (Ermolaeva *et al.,* 2001) ;
- les groupes de gènes reliés par leur fonction métabolique (Ogata *et al.,* 2000; Zheng *et al.,* 2002) ;
- la co-expression des gènes dans les données d'expression de puces à ADN (Sabatti *et al.,* 2002; Bockhorst *et al.,* 2003a; Bockhorst *et al.,* 2003b; De Hoon *et al.,* 2004) ;

Les méthodes utilisant des données comme les clusters de gènes conservés, les fonctions ou encore les données de régulation ne sont applicables qu'aux gènes pour lesquels on dispose de ces informations (gènes orthologues, promoteurs, sites de fixation, etc.) et non à la totalité du génome.

La distance intergénique

Parmi l'ensemble des propriétés citées ci-dessus, la distance intergénique est celle qui individuellement permet la meilleure qualité de prédiction. Chez *E. coli*, utiliser la distance intergénique comme critère de classification des paires de gènes opéroniques ou non-opéroniques permet une précision de 82% (Salgado *et al.,* 2000; Moreno-Hagelsieb and Collado-Vides 2002). L'addition de nouvelles propriétés comme les fonctions des gènes (Gene Ontology - GO ou MultiFun) permet d'augmenter les performances, mais d'une manière moins importante (Romero and Karp 2004). Cette puissance de prédiction lorsqu'on utilise les distances intergéniques est due à un phénomène récurrent chez tous les organismes chez lesquels des opérons ont été caractérisés : les distances entre les gènes appartenant à un même opéron ont une tendance prononcée à être plus courtes, quelquefois, il y a même superposition de une ou quatre bases entre les séquences codantes au sein d'un opéron.

Plusieurs explications (qui ne sont pas mutuellement exclusives) à ce phénomène sont possibles. Une première est que le nombre d'opérons ayant des éléments régulateurs internes et donc nécessitant des régions in-

tergéniques plus grandes est peu important chez *E. coli*. Une autre raison possible pourrait être la protection des ARNm de la dégradation, les régions non-traduites des ARNm ne sont en effet pas protégées par les complexes de traduction et sont donc plus sujet à la dégradation (Schneider *et al.,* 1978). L'hypothèse de la protection des ARNm est appuyée par les distributions très similaires des distances entre les gènes intraopéroniques chez les différentes bactéries (Moreno-Hagelsieb and Collado-Vides 2002).

La distribution des distances intergéniques n'est pas un critère complètement discriminant des paires de gènes opéroniques et non-opéroniques puisque leurs distributions respectives sont chevauchantes chez *E. coli*. Par conséquent la distance intergénique, bien que le meilleur prédicteur lorsqu'on l'utilise comme unique critère, n'est pas suffisante et d'autres critères doivent être pris en compte pour mieux discerner les paires de gènes opéroniques des paires non-opéroniques.

Les méthodes de prédiction d'opérons utilisant la distance intergénique peuvent être divisées en deux classes : les méthodes supervisées (nécessitant un ensemble d'apprentissage qui est *E. coli* dans la majorité des cas) et les méthodes non-supervisées. Ces dernières partent de l'hypothèse que le biais de distribution des distances intergéniques entre les gènes adjacents ayant la même direction de transcription vers les valeurs plus faibles est entièrement due à la présence des opérons. En conséquence les propriétés des paires de gènes adjacents non-opéroniques mais ayant la même direction de transcription peuvent être déduites de celles des paires de gènes adjacents mais ayant des directions de transcription opposées (Price *et al.,* 2005). Néanmoins, il existe des travaux allant plutôt à l'encontre de cette hypothèse : la population des régions intergéniques des paires de gènes à direction de transcription opposée n'est pas homogène. Généralement elle se divise en deux populations : les régions intergéniques des paires convergentes (voir définition plus bas), qui sont plus courtes, et les régions intergéniques des paires divergentes, plus longues (Rogozin *et al.,* 2002).

La conservation de la proximité des gènes à travers les génomes

Les génomes microbiens subissent des multiples réarrangements même entre les souches d'une même espèce bactérienne, les gènes d'une paire de gènes étant adjacents dans une souche mais pas chez l'autre. Néanmoins, certains gènes ont tendance à être colocalisés systématiquement dans le même ordre chez des espèces éloignées. Dans l'hypothèse d'une sélection positive sur les opérons, ces regroupements de gènes ordonnés conservés par l'évolution dans des génomes phylogénétiquement distants correspondraient aux opérons ancestraux. Ermolaeva *et al*. (2001)

ont développé de telles méthodes phylogénétiques qui ne s'appliquent que pour des opérons ancestraux conservés dans des groupes assez diversifiés.

Données d'expression de gènes

Certaines méthodes de prédiction d'opéron utilisent la corrélation des niveaux d'expression des gènes. On part de l'hypothèse que les gènes appartenant à un même opéron ont des niveaux de transcription très similaires. Cette similarité permet de discerner les paires de gènes adjacents opéroniques de celles non-opéroniques (Sabatti *et al.,* 2002). Une condition nécessaire pour réaliser cette inférence est que lors des expériences (généralement des puces à ADN) beaucoup de gènes soient perturbés afin de récupérer une information significative pour de nombreux opérons. Une critique possible serait que deux gènes adjacents peuvent avoir leur niveau d'expression très corrélé sans pour autant appartenir à un opéron. Price et al. (2006) ont montré que le niveau d'expression des gènes adjacents n'appartenant pas à la même UT est significativement plus corrélé que le niveau d'expression de deux gènes pris aux hasard. Ainsi, la colocalisation des gènes sur le chromosome donne déjà lieu à une co-expression. Néanmoins, dans la même étude Price et al. montrent également que la corrélation due au fait que deux gènes appartiennent à une même UT est significativement supérieure à la corrélation due à la simple colocalisation des gènes (Price *et al.,* 2006).

La fonction des gènes

Comme nous l'avons mentionné plus haut, une hypothèse forte sur l'origine des opérons constitue leur implication dans une même fonction cellulaire ou un même processus biologique. Selon cette hypothèse, la fonction des gènes devrait donc être un critère pertinent pour distinguer les paires de gènes opéroniques des paires de gènes non-opéroniques. Une difficulté évidente survient néanmoins lorsqu'on veut utiliser cette propriété dans le cadre de la prédiction d'opéron. Si l'on utilise la fonction métabolique, on est limité à une fraction du génome. Ce problème reste présent même lorsqu'on utilise des annotations très généralistes comme les GO. De plus, un autre problème apparaît dans ce cas, les GO sont déclinés selon un schéma fonctionnel et non sur des fonctions physiologiques. Enfin, les gènes appartenant à un opéron ne sont pas toujours reliés fonctionnellement, surtout dans le cas de nouveaux opérons (Price *et al.,* 2006).

4.2.1 Les cartes opéroniques de *Buchnera APS*

Nous avons trouvé dans la littérature trois prédictions des unités de transcription de BAp : BioCyc[6], DOOR[7] et MicrobesOnline[8]. Il s'agit de prédictions automatiques provenant de l'application de méthodes de prédiction d'opéron sur l'ensemble des génomes bactériens séquencés et annotés. MicrobesOnline contient des prédictions réalisées avec la méthode non-supervisée proposée par Price *et al*. (2005). Cette méthode utilise (1) la distance intergénique ; (2) la fréquence de la co-occurrence de la paire dans un intervalle de 5kb dans d'autres génomes ; (3) la similarité fonctionnelle (COG) ; (4) la similarité de l'indice d'adaptation des codons (CAI). La méthode de la base DOOR a été développée par Dam *et al*. (2007), elle utilise : (1) la distance intergénique ; (2) la conservation du voisinage du gène ; (3) la distance phylogénétique ; (4) un score de similarité des termes GO ; (5) le ratio des longueurs de gènes ; (6) l'information des motifs de séquence. Enfin, la méthode sous-jacente aux prédictions de BioCyc, développée par Romero et Karp (2004), utilise : (1) la distance intergénique ; (2) la similarité des fonctions ; (3) la fonction métabolique des gènes ; (4) les interactions de type protéine-protéine ; (5) le CAI. Pour cette dernière méthode *E. coli* a été utilisée comme ensemble d'entraînement. Toutes ces méthodes utilisent la distance intergénique et la similarité des fonctions des paires de gènes, auxquelles sont greffés ensuite d'autres critères qui vont engendrer des différences de prédiction. Ces différences d'annotation peuvent justement être enrichissantes.

4.3 La topologie du chromosome bactérien – régulation par la structure de l'ADN

Tout changement topologique de l'ADN est susceptible d'affecter toutes les transactions de l'ADN. Les changements topologiques doivent donc être considérés comme des facteurs de contrôle de l'expression des gènes. Le terme topologie fait dans ce cas référence au repliement de l'ADN dans

6 http://biocyc.org/gene-search.shtml

7 http://csbl1.bmb.uga.edu/OperonDB/displayNCoperon.php?id=70&page=1

8 http://www.microbesonline.org/operons/

l'espace. Les propriétés topologiques majeures ayant un rôle dans le contrôle de l'expression génétiques sont le surenroulement, la courbure, la flexibilité et la stabilité. La courbure, la flexibilité et la stabilité de l'ADN, contrairement au surenroulement, dépendent en partie de la structure primaire de l'ADN. Certaines de ces propriétés topologiques ont une composante extrinsèque, contrainte par des protéines interagissant avec le nucléoïde[9] : la polymérase, les facteurs de transcription et un ensemble hétérogène de protéines impliquées dans la réparation, la protection, la réplication et la structuration du chromosome.

4.3.1 Les toporégulateurs

Le nucléoïde est hautement associé à des protéines, le rapport protéines/ADN étant d'environ 5/1 chez les bactéries (Murphy and Zimmerman 1997a, b). Avec le concours de ces protéines, les bactéries concilient une haute condensation d'ADN (chez *E. coli* le chromosome est condensé 1000 fois) et une intense activité du chromosome qui semble entièrement accessible à la transcription. Nous appellerons cet ensemble très hétérogène de protéines, les toporégulateurs, car par leur association à l'ADN ces protéines régulent les propriétés topologiques du chromosome et ainsi les autres processus impliquant des transactions d'ADN.

4.3.1.1 Les « Nucleoid associated proteins » (NAP)

Le chromosome d'*E. coli* est associé à approximativement une quinzaine de protéines appelées NAP (Nucleoid Associated Proteins). Azam et al. (1999) ont décrit chez *E. coli* 12 NAP : CbpA (curved DNA-binding protein A) ; CbpB ou Rob (curved DNA-binding protein B) ; Dps (DNA-binding protein from starved cells) ; Hfq (host factor for phage Qb) ; StpA (suppressor of td2 phenotype A) ; Lrp (leucine responsive regulatory protein) ; IciA (inhibitor of chromosome initiation A) ; H-NS (histone-like nucleoid structuring protein) ; HU (heat-unstable nucleoid protein) ; IHF (integration host factor) ; DnaA (DNA-binding protein A) et FIS (factor for inversion stimulation).

[9] Le nucléoïde est une structure dans laquelle l'ADN d'une cellule procaryote est concentré ; elle a un contour irrégulier et elle n'est pas délimitée par une membrane (Berthet, 2005).

Les NAP représentent la majorité des protéines s'associant au nucléoïde. Ces protéines participent non seulement à la compaction de l'ADN, mais interviennent aussi dans les processus comme la réplication, la recombinaison et la transcription. L'association plutôt aspécifique et la grande concentration des NAP suggèrent un contrôle d'ordre global de la transcription.

Les NAP sont des protéines basiques, ayant une petite masse moléculaire (Azam and Ishihama 1999). La grande majorité de toutes les NAP s'associe non-spécifiquement à l'ADN. Certaines des NAP sont capables néanmoins de se lier de façon spécifique. En fonction de l'affinité à l'ADN des NAP reconnaissant des séquences spécifiques, on peut les ordonner de la façon suivante : IHF > Lrp > CbpB > Fis > DnaA ; pour les NAP qui n'interagissent pas spécifiquement avec le nucléoïde le classement est le suivant : HU > H-NS > StpA > CbpA > IciA > Hfq > Dps (Azam and Ishihama 1999). On peut aussi classer les NAP en fonction de leur abondance : HU > FIS > H-NS > IHF > StpA (**Tableau 3**).

Si on considère qu'une cellule en phase exponentielle contient environ trois chromosomes, alors on estime que 20% des chromosomes sont liés par les NAP (Johnson *et al.,* 2005).

La composition de la population des NAP dans la cellule n'est pas fixe. Ceci est dû au fait que les NAP sont sous l'influence d'une régulation complexe grâce au réseau dans lequel elles sont interconnectées. Chaque condition de croissance implique, par exemple, un profil d'expression spécifique de chacune des NAP (Azam and Ishihama 1999; Dorman 2004).

Bien que la population des NAP soit estimée à environ 15 protéines différentes, la grande majorité des travaux porte sur l'étude de 4 NAP particulières HNS, Fis, HU et IHF (Dorman 2004). Une description de chacune de ces NAP est donnée par la suite.

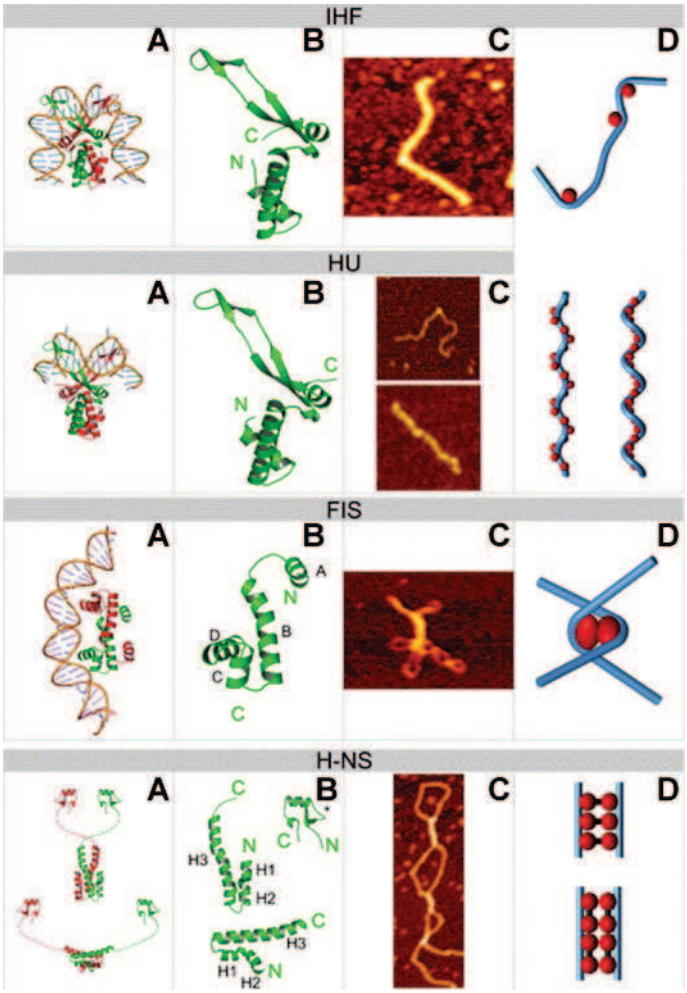


Figure 9. Structure et propriétés architecturelles des protéines associées au nucléoïde (NAP). IHF. A) Structure du complexe IHF-ADN (PDB : 1IHF, (Rice *et al.,* 1996)). B) Structure de l'unité monomérique de IHF. C) Image du complexe IHF-ADN (Dame *et al.,* 2005). HU. A) Structure d'un complexe HU-ADN (PDB : 1P51, (Swinger *et al.,* 2003). B) Structure de la sous-unité monomérique de HU. C) Image d'un complexe HU-ADN, HU se lie à l'ADN à des faibles concentrations (en haut, 200x200nm) et induit la formation de filaments rigides à des hautes concentrations (van Noort *et al.,* 2004) (en bas, 70x70nm). D) Modèle de la compaction de l'ADN par la liaison multiple de HU et de IHF. FIS. A) Modèle du complexe FIS-ADN (Pan *et al.,* 1996). B) Structure de la sous-unité monomérique de FIS. C) Image du complexe FIS-ADN (Schneider *et al.,* 2001). D) Modèle de formation de nœuds par des interactions FIS-FIS. H-NS. A) Modèle du dimère H-NS basé sur la structure du domaine de liaison à l'ADN (partie C-terminale) de H-NS chez *E. coli* (PDB : 1HNR-, résidus 80-136) (Shindo *et al.,* 1995)), et deux structures différentes du domaine de dimérisation (partie N-terminale) de H-NS (PDB :1N18, résidus 1-46, chez *E. coli* (Bloch *et al.,* 2003) et 1LR1, résidus 1-57, chez *Salmo-*

nella typhimurium (Esposito *et al.,* 2002). B) Structure de constituants monomériques du dimère de H-NS. C) Image des boucles d'ADN formées suite à l'établissement des ponts entre les différentes parties du chromosome par H-NS, taille de l'image : 110x340nm. D) Modèle des ponts ADN établis par H-NS. (d'après Luijterburg et al. (2006)).

FIS, Factor for Inversion Stimulation

FIS est présente chez les γ-protéobactéries séquencées et une forme tronquée de FIS a été retrouvée chez des β-protéobactéries (Johnson *et al.,* 2005).

Contrairement à HU, H-NS et IHF, FIS n'est pas distribuée uniformément le long des chromosomes, mais plutôt concentrée sur certains points (Azam and Ishihama 1999). Sa concentration décroît d'environ deux ordres de magnitude lors du passage de la phase exponentielle à la phase stationnaire chez *E. coli* (Blot *et al.,* 2006).

Cette protéine contient un motif hélice-tour-hélice dans sa partie C-terminale (**Figure 9**). Les acides aminés essentiels à son bon fonctionnement ont été déterminés par de nombreuses études à l'aide de mutations ponctuelles (Koch *et al.,* 1991; Osuna *et al.,* 1991; Safo *et al.,* 1997; Yang *et al.,* 1998; Cheng *et al.,* 2000).

Les mutants *fis* sont viables, manifestant des effets pléiotropiques : une croissance plus lente, une incapacité de vivre dans des conditions d'anaérobie. Ils possèdent une membrane aux propriétés altérées et montrent une motilité réduite, ainsi qu'une forme cellulaire allongée et filamenteuse (Osuna *et al.,* 1995; Xu and Johnson 1995; Membrillo-Hernández *et al.,* 1999).

La comparaison de plusieurs sites de fixation de FIS à l'ADN a permis d'établir une séquence consensus palindromique de 15 pb (**Tableau 3**). Cette séquence consensus est dégénérée. Plusieurs autres séquences des sites de fixation de FIS déterminées expérimentalement diffèrent par une ou deux bases apparaissant comme fortement conservées dans la séquence consensus. Enfin, bien qu'on parle souvent du site de fixation de FIS à l'ADN, elle s'associe principalement de manière non spécifique, avec une affinité importante (Pan *et al.,* 1996) (**Tableau** 3). Lors de son interaction avec l'ADN, FIS courbe l'hélice et le rayon de courbure dépend de la séquence à laquelle FIS se fixe et de la composition et des propriétés structurelles des séquences avoisinantes du site de liaison (McLeod *et al.,* 1999).

Tableau 3. Caractéristiques des 4 NAP les plus étudiées chez les bactéries. Le Kd indique l'affinité d'interaction entre la protéine et l'ADN, plus cette valeur est faible et plus l'interaction est forte.

	FIS	HU	IHF	H-NS
Forme moléculaires fonctionnelle	Homodimère	hétérodimère	hétérodimère	Dimer-oligomère
% de chromosome occupé en phase exponentielle	6	8	4	1
% de chromosome occupé en phase stationnaire précoce	<1	6	11	1
% de chromosome occupé en phase stationnaire tardive	<1	6	12	1
Kd	Spécifique : 2 à 5 nM Non spécifique : < 20 nM	De 400 nM à 30 μM	Sp : 0.3 à 20 nM	1 à 3 μM
Induction des courbures de l'ADN	Jusqu'à 90°	oui	Jusqu'à 140°	oui
Site de fixation	GNTYAWWWWWTRANC (Pan *et al.,* 1996; Hengen *et al.,* 1997; Ussery *et al.,* 2001)	-	WATCAANNNNTTR (Goodrich *et al.,* 1990)	TCGWTWAAWW (Lang *et al.,* 2007)

La liaison non-spécifique et la capacité de courber l'ADN rendent FIS importante pour la condensation de l'ADN. Néanmoins, FIS ne semble pas capable d'introduire des supertours négatifs. Par contre FIS est un des principaux régulateurs des gènes codant les protéines capables d'agir directement sur le niveau de surenroulement de l'ADN comme les topoisomérases (Schneider *et al.,* 1999). D'une manière plus globale et indirecte, FIS régule un grand nombre de systèmes cellulaires en couplant la physiologie de la cellule avec la topologie de l'ADN (Schneider *et al.,* 1999). Il assure le maintien du niveau de surenroulement de promoteurs, optimal pour leur reconnaissance et donc pour l'expression de leur gène, même quand le niveau total de surenroulement du chromosome n'est pas favorable à la transcription (Schneider 1997; Schneider *et al.,* 1999).

FIS régule directement la transcription de certains gènes, en interagissant avec les unités α et σ de l'ARN polymérase (Travers and Muskhelishvili 2005b). Les gènes se trouvant sous son influence directe, sont des gènes dont le niveau d'expression varie en fonction de la croissance : les gènes codant les ARN stables (ARNr et ARNt), les gènes codant les protéines ribosomales, ainsi que les gènes impliqués dans la régulation du métabolisme et de la croissance.

IHF, Integration Host Factor

IHF est fonctionnel sous forme d'hétérodimère, IhfA-IhfB, codés par les gènes *himA*, *himD* ou *ihfA* et ihfB*),* retrouvés chez certaines protéobactéries (**Figure 9**) (Johnson *et al.,* 2005). IHF a une structure similaire à HU et une identité d'environ 45% a été mesurée entre les séquences de HU et de IHF, indiquant qu'il s'agit de gènes paralogues (Drlica and Rouviere-Yaniv 1987).

Les mutants *ihfA-ihfB* ont des modifications légères de phénotype. Certaines études suggèrent que IHF soit essentiel à la réponse à l'azote limité sous des conditions d'anaérobie (Nash 1996). IHF est également nécessaire dans des conditions d'inanition (comme HU) (Nystrom 1995).

IHF se lie à des séquences de 30-35 pb (Goodrich *et al.,* 1990). Les sites de fixation sont asymétriques et contiennent un élément central consensus : WATCAANNNNTTR. La sous-unité α interagit avec la partie WATCAA et la sous-unité β, avec TTR. Néanmoins, il semblerait que la plupart des molécules de IHF se lient de façon non-spécifique au chromosome. Ces interactions non-spécifiques sont importantes pour la compaction du chromosome (Ali *et al.,* 2001). IHF est impliqué dans l'initiation de la réplication, en déformant l'ADN de façon à faciliter l'assemblement du complexe de réplication (Ryan *et al.,* 2002).

HU, Heat Unstable nucleoid protein

HU est la NAP la plus abondante chez *E. coli* (**Tableau 3**). Des homologues de HU sont retrouvés dans toutes les grandes divisions des bactéries (Johnson *et al.,* 2005). La protéine est très abondante en phase exponentielle et bien que sa concentration diminue lorsque la bactérie atteint la phase stationnaire, HU reste encore l'une des NAP les plus abondantes, avec IHF, son paralogue. Il y a environ 45% d'identité entre les sous-unités de HU et de IHF. La différence entre HU et IFH étant que la dernière est capable de se lier à l'ADN de façon beaucoup plus spécifique.

Bien que chez *E. coli* HU soit majoritairement présente sous forme hétéro-dimérique (HU-αβ), chez les autres bactéries, un seul gène code pour HU-α et la protéine est fonctionnelle sous la forme homodimérique HU-α2. C'est d'ailleurs la forme la plus abondante durant la phase exponentielle, alors que la forme hétérodimérique devient prédominante lors du passage à la phase stationnaire (Claret and Rouviere-Yaniv 1997). La forme hétérodimériques semble essentielle à la survie dans des conditions d'inanition. HU-β2 se lie très faiblement à l'ADN « linéaire » comparé à HU-α2 et HU-αβ (Pinson *et al.,* 1999).

HU semble avoir plus d'affinité pour l'ADN non-courbé que pour l'ADN courbé (Azam and Ishihama 1999). Elle se lie à l'ADN sous forme de dimères, de façon non-spécifique et en courbant l'ADN. Son association avec l'ADN dépend de la concentration en sels et est stimulée par l'augmentation du surenroulement négatif (Kobryn *et al.,* 1999). HU se lie à l'ADN surenroulé, à l'ADN simple brin, à l'ARN simple et double brin, ainsi qu'aux complexes ARN-ADN, à l'ADN cruciforme et à l'ADN ayant des bouts 3' simple brin (Drlica and Rouviere-Yaniv 1987). Ainsi, HU lie notamment les produits intermédiaires de la recombinaison ou de la réparation de l'ADN. Enfin HU participe à la formation des microcercles[10] d'ADN.

Cette protéine ne semble pas établir des interactions avec d'autres protéines (Dorman and Deighan 2003).

Les mutants *ΔhupAΔhupB* sont viables, néanmoins leur phénotype est le plus atteint (parmi les mutants NAP). Chez les mutants *ΔhupAΔhupB*, IHF semble pouvoir assurer certaines fonctions de HU permettant ainsi la viabilité de ces mutants. Par contre, les doubles mutants *ΔhupABΔihfAB* ne sont pas viables. Les cellules déficientes en HU sont sensibles au froid et déficientes en réparation de l'ADN par la recombinaison (Li and Waters 1998). Enfin les populations des cellules *ΔhupAΔhupB* contiennent un certain nombre de cellules filamenteuses et/ou anuclées. Les mutants *ΔhupABΔihfAB* d'*E. coli* sont difficiles à construire et leur développement est compromis (Johnson *et al.,* 2005). Des mutants *ΔhupAΔmukB* n'ont jamais pu être construits.

HU intervient dans la transcription en jouant sur la flexibilité et en courbant l'ADN. Cette NAP permet de lever l'effet de H-NS quand c'est nécessaire. Ainsi, le ratio HU/H-NS est un facteur important pour la régula-

10 Complexes circulaires d'ADN, de quelques centaines de bases, stabilisées par des protéines.

tion générale de l'expression des gènes et doit rester dans un certain intervalle (Dame and Goosen 2002). Ce ratio a été estimé à 2.5 durant la phase exponentielle et à 1 durant la phase stationnaire (Azam and Ishihama 1999). Il semblerait que HU puisse aussi réguler la traduction étant donnée sa capacité à s'associer à l'ARN. Par ailleurs, HU est impliquée à plusieurs niveaux dans la réplication de l'ADN : l'initiation de la réplication, la séparation des chromosomes et la division cellulaire (Johnson *et al.,* 2005). Ces fonctions sont confirmées par les phénotypes des mutants *ΔhupAΔhupB*.

Une autre fonction majeure de HU est la structuration du chromosome (Paull *et al.,* 1994). HU est capable de moduler la flexibilité de l'ADN (Flashner and Gralla 1988). Elle est également capable de maintenir le niveau de surenroulement négatif du chromosome en collaboration avec la topoisomérase I (elle participe à la formation de supertours négatifs) (Rouvière-Yaniv *et al.,* 1979). La déficience de HU est compensée par l'activité de GyrB. En conclusion, il apparaît que HU facilite l'activité de la gyrase et diminue celle de la topoisomérase I.

H-NS, Histone like Nucleoid Structuring protein

H-NS est un régulateur transcriptionnel global oligomérique participant à la structuration du nucléoïde. Des homologues de H-NS sont retrouvés dans de nombreuses bactéries Gram négatives. Chez *E. coli*, H-NS possède un paralogue, StpA, une chaperonne des ARN. H-NS est exprimé à peu près à taux constant, indépendamment de la phase de croissane (**Tableau 3**), avec néanmoins une légère surexpression en début de phase stationnaire (Dorman 2004).

H-NS a deux domaines fonctionnels, le premier N-terminal, permet l'oligomérisation alors que le domaine C-terminal est responsable de l'interaction avec l'ADN, reliés par une région de liaison (**Figure 9**) (Dorman *et al.,* 1999). L'unité fonctionnelle basique est le dimère.

H-NS agit généralement seul, ou parfois en coopération avec d'autres facteurs de transcription (Dorman 2004). Toutes ses cibles répondent aux changements brusques de l'environnement (*e.g.* les changements de la pression osmotique) (Dorman and Deighan 2003). La liaison de H-NS à l'ADN *in vitro* s'explique en partie par les caractéristiques de l'ADN. La participation des propriétés de l'ADN au mécanisme de liaison de H-NS n'est pas suffisante pour expliquer la précision avec laquelle H-NS réprime ses cibles, car H-NS semble pouvoir se lier à l'ADN de 2 façons : structurellement – reconnaissant l'ADN courbé intrinsèquement, celui-ci mobilise la majorité des H-NS de la cellule, ou avec un grand degré de spécificité – reconnaissant des sites de fixation spécifiques.

Une séquence consensus du site de liaison à l'ADN du HNS a été déterminée, **Tableau 3** (Lang *et al.,* 2007). Le dinucléotide central T-A a un rôle probablement très important car il confère une instabilité thermique et une flexibilité torsionnelle et axiale du site de fixation. H-NS se lie préférentiellement à l'ADN courbé, riche en bases A et T, de façon non-spécifique (par rapport à la séquence primaire de l'ADN) et à l'habilité de contraindre les surenroulements *in vitro*.

Lors de son association non-spécifique avec l'ADN, H-NS est capable de former des boucles d'ADN piégeant la polymérase, empêchant ainsi la transcription. Pour lever la répression de H-NS, l'intervention d'autres protéines est nécessaire (HU, FIS, ou d'autres facteurs de transcription spécifiques (Dorman and Deighan 2003)), ou encore l'intervention de certains signaux externes, comme la température et l'osmolarité modificateur de la topologie de l'ADN.

La régulation du gène *hns* est faite par les protéines H-NS, StpA et FIS. Son promoteur est aussi stimulé par le choc thermique (froid) par l'intermédiaire de la protéine CspA. L'expression de *hns* serait aussi régulée post-transcriptionnellement par un ARN antisense, DsrA et une protéine se liant à l'ARN, Hfq. L'expression de *hns* augmente aussi si la pression hydrostatique augmente. Enfin, l'expression de *hns* est corrélée au cycle cellulaire de la façon suivante, elle augmente lorsque la synthèse de l'ADN augmente (Dorman 2004).

H-NS, comme d'autres NAP semble impliquée également dans la recombinaison (Kawula and Orndorff 1991). C'est un facteur protéique intervenant de façon globale dans la transcription (Blot *et al.,* 2006), sa mutation entraînant une croissance lente et une sensibilité accrue aux changements de température et de pression osmotique. H-NS participe donc à la régulation thermique et osmotique de la transcription (probablement à travers le changement de la conformation de l'ADN dû à ces facteurs).

H-NS est capable de contraindre les supertours négatifs mais pas de moduler le niveau de surenroulement, contrairement à HU (Yasuzawa *et al.,* 1992). H-NS n'apparaît pas comme essentielle à la structuration et à la maintenance du chromosome (Johnson *et al.,* 2005).

4.3.2 Les topoisomérases

En plus des protéines NAP, les topoisomérases et les protéines SMC sont essentielles à l'organisation du nucléoïde.

De par leur rôle fondamental dans la cellule, les topoisomérases sont présentes dans tous les organismes, avec de grandes différences structurelles. Chez les bactéries, la fonction fondamentale des topoisomérases

est de maintenir un niveau de surenroulement négatif optimal à l'expression des gènes en fonction des conditions de vie, et donc au final à la viabilité de la cellule (Zechiedrich *et al.,* 2000).

Chez *E. coli* le surenroulement est maintenu grâce à 4 topoisomérases : la topoisomérase I (*topA*) et la topoisomérase III (*topB*), capables de relâcher les supertours négatifs mais pas les supertours positifs ; l'ADN gyrase (*gyrA* et *gyrB*), la seule topoisomérase capable d'introduire des supertours négatifs, grâce à l'utilisation de l'ATP et la topoisomérase IV (*parE* et *parC*), une autre topoisomérase ATP-dépendante qui a comme principal rôle la décaténation des chromatides à la fin de la réplication (Zechiedrich *et al.,* 2000; Champoux 2001).

La relaxation du chromosome active *gyrA* et *gyrB* (gyrase) et réprime *topA* (topoisomérase I).

La dépendance à l'ATP de la gyrase la lie au ratio cellulaire [ADP]/[ATP] et donc à l'état énergétique de la cellule, le niveau énergétique cellulaire est donc un des principaux facteurs liant le métabolisme au niveau de surenroulement du chromosome.

4.3.3 Le complexe SMC

Chez *E. coli* les protéines MukBEF ont été identifiées comme formant un complexe de maintenance structurale du chromosome. Ces protéines interviennent dans le niveau de surenroulement et la condensation de l'ADN (Lindow *et al.,* 2002). Contrairement aux NAP, les protéines SMC ont un poids moléculaire beaucoup plus important (Melby *et al.,* 1998). Leurs interactions avec l'ADN sollicitent de l'ATP et sont stabilisées par d'autres cofacteurs (Hopfner *et al.,* 2000; Strunnikov 2006). Les mutants *mukBEF* montrent une décondensation du nucléoïde et un changement du surenroulement, ainsi qu'une baisse de viabilité à des températures supérieures à 30°C (Lindow *et al.,* 2002).

4.3.4 Le surenroulement de l'ADN comme régulateur transcriptionnel

Le surenroulement négatif favorise la compaction du nucléoide et est essentiel à la survie des cellules bactériennes. Crozat et al. (2005) ont démontré que le niveau global de surenroulement est un paramètre soumis à la sélection et que par conséquent tous les acteurs impliqués dans son maintien le sont aussi.

Le lien étroit entre le niveau de surenroulement et la transcription est donné par la polymérase, qui va reconnaître plus ou moins bien le promoteur des gènes en fonction du surenroulement local de leur promoteur. Comme nous l'avons vu, le surenroulement varie sous l'effet des topoisomérases et d'autres protéines comme les NAP constituant des barrières à la diffusion ou l'ARN polymérase (**Figure 10**). En fonction de l'architecture de leur promoteur certains gènes vont s'exprimer de façon optimale à un faible niveau de surenroulement négatif alors que d'autres s'exprimeront à un haut niveau de surenroulement. De plus, le niveau de surenroulement n'intervient pas seulement dans la modulation de l'initiation de la transcription mais aussi au niveau de l'élongation en provoquant l'apparition de structures spéciales empêchant l'élongation (par exemple, l'excès de surenroulement négatif lié au manque de la topoisomérase I, inhibe la croissance de la bactérie en permettant à la polymérase de former des R-loops et en bloquant ainsi l'élongation de la chaîne d'ARN (Travers and Muskhelishvili 2005b)).

Cheung *et al.* (2003) ont dépisté un cluster de gènes sensibles au stress osmotique chez *E. coli*, dont une partie importante est sensible au changement du surenroulement du chromosome. C'est ce qui a suggéré dans un premier temps que le surenroulement du chromosome reflète les changements environnementaux. Par conséquent, le changement du niveau de surenroulement devrait induire le changement du niveau d'expression d'un large spectre de gènes car le changement de la pression osmotique affecte différentes classes de gènes.

Peter *et al*. (2004) ont mesuré le niveau d'expression des gènes d'*E. coli K12* en conditions basales et après relaxation du chromosome, ce qui leur a permis de définir deux classes de gènes : les gènes induits par la relaxation (106), et les gènes inhibés par la relaxation (200). Ainsi, 7% des gènes semblent être affectés par le surenroulement du chromosome. Pour pouvoir dépister l'effet de la relaxation du chromosome, les auteurs ont fait l'hypothèse que la stabilité des ARNm est la même pour tous les ARNm. Ce qui est encore plus remarquable est que les promoteurs des gènes induits par la relaxation sont significativement AT-riches, ainsi que leurs séquences codantes. Les séquences des gènes inhibés par la relaxation montrant plutôt des séquences GC-riches. Tous les gènes sensibles au surenroulement sont dispersés le long du chromosome et appartiennent à différentes classes fonctionnelles.

Parmi les promoteurs les plus activés par le surenroulement négatif, on compte les promoteurs des ARNt et des ARNr (Travers and Muskhelishvili 2005b). Dans les promoteurs stables deux éléments participent à la réponse de ces promoteurs à la superhélicité : (1) la région GC-

riche entre le site d'initiation de la transcription et le -10 et (2) l'élément UP en amont du -35.

Blot et al. (2006) ont également montré que l'organisation du profil de transcription est étroitement reliée au niveau de surenroulement du génome. Une grande proportion des gènes sensibles au surenroulement du chromosome sont des gènes codant pour des produits impliqués dans des voies anaboliques, ce qui est cohérent avec le couplage entre le métabolisme et le profil de transcription à travers la topologie de l'ADN. La connexion entre les deux étant faite par le rapport de concentration [ATP]/[ADP], quand la valeur de ce rapport est grande (>1), cela favorise l'augmentation du surenroulement négatif grâce à la gyrase et subséquemment la biosynthèse (Blot *et al.,* 2006) alors que la relaxation de l'ADN va plutôt stimuler la production de l'ATP en activant le catabolisme. Enfin, les voies de synthèse de novo des nucléotides impliquent également des gènes dont la transcription est sensible au surenroulement.

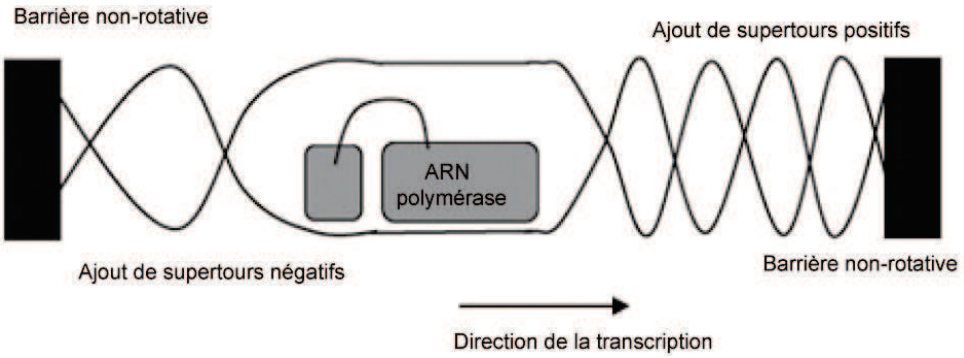


Figure 10. Impact local de l'ARN polymérase durant la transcription sur le niveau de surenroulement de l'ADN. Des protéines s'associant à l'ADN constituent des barrières topologiques pouvant être considérées comme des barrières non-rotatives. L'ARN polymérase ne peut pas tourner autour de l'ADN, son mouvement lors de la transcription entraîne la création de supertours positifs devant, pouvant être relâchés par la gyrase, et des supertours négatifs derrière, pouvant être relâchés par la topoisomérase I (d'après Dorman (Dorman 2008)).

FIS agit sur le niveau de surenroulement par deux biais, en se liant à l'ADN et donc contraignant le niveau de surenroulement ou en régulant l'activité de la gyrase (Schneider *et al.,* 1999).

4.3.5 Les domaines topologiques

Il existe de plus en plus de preuves que le nucléoïde est constitué de domaines topologiques (Travers and Muskhelishvili 2005a), un domaine topologique étant défini comme un territoire du chromosome typologiquement indépendant. La taille de ces domaines a été estimée initialement à environ 50-100 kb (Sinden and Pettijohn 1981). Postow *et al.* (2004) ont ré-estimé cette dimension à environ 10 kb, de plus, ils ont aussi conclu que les frontières entre ces domaines topologiques sont dynamiques ce qui implique que l'existence de barrières interdomaines est seulement transitoire. La longueur de 10 kb coïncide avec la longueur moyenne des supertours de l'ADN observés avec la microscopie électronique (Travers and Muskhelishvili 2005a). L'efficacité et la fréquence de ces barrières dépend de l'activité de transcription (car la transcription intensive au niveau de certains promoteurs peut entraîner aussi la formation de barrières topologiques), des NAP, mais également de l'activité des topoisomérases. Valens *et al*. (2004) affirment que la durée de vie des domaines topologiques, ou plutôt des barrières inter-domaines, est bien inférieure au temps de réplication. Ce premier niveau d'organisation du nucléoïde est nécessaire d'un côté pour la stabilité du chromosome, une modification de structure ou de surenroulement dans un des domaines n'ayant pas d'effet sur l'ensemble du chromosome car les barrières empêchent la libre diffusion des supertours, et d'un autre côté, cette organisation participe à la compaction du chromosome (Johnson *et al.,* 2005).

Les constituants des barrières entre les domaines topologiques restent une question ouverte. Les principaux acteurs semblent être les NAP. Plus précisément, Hardy et Cozarelli (2005) ont montré que ce sont FIS et H-NS qui interviennent dans la composition des barrières, ce qui ne semble pas être le cas de IHF et HU. Ceci serait dû à leur capacité de contraindre des boucles d'ADN par liaison non-spécifique (Dame *et al.,* 2001). Dans leurs études, Hardy et Cozarelli ont également montré l'implication dans les barrières topologiques de trois autres protéines : une phosphoglucomutase impliquée dans la glycolyse, Pgm ; DksA, une protéine interagissant avec plusieurs facteurs intervenant dans la réplication et avec MukB ; l'enzyme TktA. Les protéines impliquées dans la ségrégation des chromosomes en fin de réplication constituent de bons candidats pour les barrières topologiques.

Le nucléoïde est situé plus au centre de la cellule, alors que la polymérase et les ribosomes sont situés à sa périphérie (Lewis *et al.,* 2000). Par conséquent l'expression des gènes est dépendante de l'accessibilité de l'ADN. Une fois qu'une région est rendue accessible, les régions avoisinantes le deviennent aussi. La transcription des régions voisines est rendue plus probable par la proximité de la machinerie de la transcription. Des

structures comme les domaines d'environ 10kb du chromosome, et vraisemblablement d'autres types de structures, participent ainsi à la mise en place d'une accessibilité dynamique des régions du chromosome. Le repliement du chromosome entraîne le rapprochement de régions distantes conduisant à des corrélations à moyenne et longue distance entre les niveaux de transcription des gènes (Jeong *et al.,* 2004; Carpentier *et al.,* 2005).

La faible concentration de MukBEF dans la cellule indique que ces protéines ne peuvent pas être les constituants de barrières topologiques participant à la formation des domaines topologiques du chromosome (Kido *et al.,* 1996).

Cette organisation à l'échelle approximative de 10kb est un premier niveau d'organisation. Une structuration à plus large échelle a été observée chez *E. coli* par Valens et al. (2004). Ils ont montré que des recombinaisons et donc des interactions entre des portions du chromosome sont confinées à l'intérieur de régions de chromosome, ce qui leur a permis de définir 4 macrodomaines : une région centrée autour de l'origine de réplication d'environ 1Mb, une autre d'une taille similaire est centrée autour du terminus, et deux régions encadrant la région du terminus et séparées de la région de l'origine par des régions non-structurées. Le repliement de ces macrodomaines entraîne des interactions entre des régions d'ADN internes aux domaines. Les régions d'ADN appartenant à des macrodomaines distincts ne semblent pas interagir.

4.3.6 Les propriétés structurelles séquence-dépendantes du chromosome

L'ADN possède des propriétés structurelles inhérentes influençant de façon considérable de nombreux processus biologiques. Il a été montré que la courbure (inhérente ou induite par les NAP) de l'ADN influence l'interaction entre l'ARN polymérase et les protéines avec lesquelles elle interagit (Pedersen *et al.,* 2000). Certaines de ces propriétés structurelles inhérentes à l'ADN ont été démontrées comme fortement corrélées avec la séquence de l'ADN (Brukner *et al.,* 1990; Bolshoy *et al.,* 1991; Hassan and Calladine 1996; Olson *et al.,* 1998; Sinden *et al.,* 1998). Des expériences ont été mises en place afin de mesurer cette relation entre la séquence d'ADN et ses propriétés. Ces travaux ont permis la construction de modèles de prédiction des propriétés structurelles des séquences d'ADN mesurant la flexibilité de l'ADN, la courbure de l'ADN ou la stabilité de l'hélice. L'intérêt s'est porté sur ces propriétés car c'est en fonction de la flexibilité, de la courbure et de la stabilité que l'interaction entre l'ADN et des fac-

teurs protéiques sera possible ou pas. Tous ces modèles sont des tables reliant des paramètres structuraux à des di- ou tri-nucléotides.

Hassan et Calladine (1996) ont démontré qu'il existe une corrélation entre la flexibilité de l'ADN au niveau des dinucléotides et l'angle entre les plans des bases aromatiques des nucléotides voisins (angle de torsion). Cet angle de torsion peut donc être utilisé comme indice de flexibilité local de l'ADN.

La stabilité de la double hélice est donnée en partie par les liaisons hydrogènes mais aussi par l'énergie d'empilement des bases. Des valeurs d'énergie d'empilement ont été déterminées par des calculs de mécanique quantique pour chaque dinucléotide (kcal/mol) (Ornstein *et al.,* 1978).

La courbure de l'ADN peut être intrinsèque, liée à la séquence primaire de l'ADN ou extrinsèque, induite par des protéines se liant à l'ADN. En raison de son implication dans le contrôle de l'accessibilité des promoteurs, la courbure de l'ADN a un important rôle dans le contrôle de la transcription, de par son implication dans les interactions ADN-protéines, des fortes courbures intrinsèques sont nécessaires et présentent à des points de contrôle comme le site d'initiation de la réplication, la réparation et la compaction de l'ADN. Bolshoy et al. (1991) ont proposé un algorithme déduit à partir de données de mobilité des dinucléotides en gel d'acrylamide pour prédire la courbure intrinsèque de l'ADN.

Un des évènements clef de l'initiation de la transcription ou de la réplication est l'ouverture de l'hélice d'ADN. Sachant que les molécules d'ADN sont négativement surenroulées *in vivo*, cela implique l'existence d'un stress de torsion à certains endroits de l'hélice, susceptibles de déstabiliser le double brin. Des légers changements dans la stabilité locale d'association des deux brins peuvent avoir des conséquences considérables quant à la réaction d'ouverture généralement médiée par divers facteurs protéiques. Si la transcription au niveau d'un promoteur dépend étroitement de la possibilité et de la fréquence d'ouverture de l'hélice, alors les déstabilisations locales du double brin, apparaissant suite aux stress de déroulement/enroulement de l'ADN, auront une grande influence sur le taux de transcription au niveau de ce promoteur. Ce type de déstabilisation du double brin de l'ADN est appelé « Stress-induced DNA duplex destabilization » (SIDD) (Wang and Benham 2008). Le SIDD dépend non seulement de la séquence de l'ADN mais aussi des séquences avoisinantes, ainsi que du niveau de surenroulement de l'ADN (Benham 1996). Un exemple bien connu de la régulation de la transcription par le SIDD est le promoteur du gène *ilvP*, chez *E. coli*. Une région instable est située à environ 100pb en amont de son promoteur. Lorsque IHF se lie à cette région, il induit la déstabilisation de la région en aval, et donc une ouverture facilitée de l'hélice

au niveau du promoteur de *ilvP* (Sheridan *et al.,* 1998). Benham et collaborateurs ont travaillé sur la caractérisation de la stabilité des molécules d'ADN et sur la recherche des régions sensibles à la séparation des brins d'ADN (Benham and Bi 2004). Ils ont ainsi développé un outil permettant de calculer le SIDD en fonction du niveau de surenroulement. Leur méthode propose deux paramètres, la probabilité que l'hélice d'ADN soit ouverte au niveau de la paire de bases, et l'énergie incrémentale libre des états dans lesquels cette paire de bases est toujours ouverte.

4.4 La régulation post-transcriptionnelle

D'autres moyens de régulation agissent lors des étapes ultérieures à l'initiation de la transcription. Il existe par exemple des facteurs protéiques interagissant de façon spécifique avec la polymérase et régulant l'élongation comme GreA (Stepanova *et al.,* 2007). L'élongation peut également être régulée par l'interférence de transcription qui peut se produire dans le cas de deux promoteurs adjacents convergents, tandem ou chevauchants (Shearwin *et al.,* 2005).

Les riboswitches sont des domaines de régulation situés dans les régions UTR des ARNm, susceptibles d'être liés par des métabolites et ainsi peuvent réguler l'expression des gènes.

Chronologiquement, le processus d'expression des gènes est régulé d'abord au niveau de la transcription comme nous l'avons vu, ensuite au niveau de la traduction et post-traduction, à ces niveaux se rajoutant d'autres mécanismes intermédiaires comme la régulation de la stabilité des ARNm. Un des mécanismes de régulation post-transcriptionnelle largement utilisé par les bactéries est l'atténuation. L'exemple classique de ce type de régulation est l'opéron du tryptophane. Avant le premier gène structurel de l'opéron, *trpE*, on trouve une séquence de tête contrôlant la transcription. Cette séquence de tête contient un atténuateur qui est un terminateur Rho-indépendant et code également pour un court peptide riche en Trp jouant le rôle de senseur de l'environnement. En fonction de l'environnement (présence du tryptophane ou pas) la transcription sera arrêtée au niveau du terminateur de la transcription où se poursuivra pour transcrire la totalité de l'unité de transcription.

La stabilité des ARNm n'est pas constante, leur demi-vie étant comprise entre 1 min et 10 min, la majorité des transcrits ayant une demi-vie située entre 2 et 20 min (Selinger *et al.,* 2003). Un des mécanismes de régulation de la stabilité des ARNm employé par les procaryotes sont les petits ARN non-codants. Enfin, il a aussi été noté l'implication des petits ARN non-codants dans la régulation de la traduction des ARNm. Il s'agit

d'ARN d'environ 100pb dont les séquences codantes ne se chevauchent pas avec d'autre séquences codantes (comme c'est la cas des miARN chez les eucaryotes). Chez *E. coli* ont été décrits environ 60 ARN de ce type, qui sont donc susceptibles de réguler plus de 1% des ARNm de la bactérie. Bien qu'ayant les mêmes fonctions biologiques que les miARN chez les eucaryotes, ces ARN ne sont pas maturés et n'agissent pas de la même manière (Gottesman 2005).

4.5 L'évolution des systèmes de régulation chez les procaryotes

Les réseaux de régulation de la transcription sont connus pour évoluer très vite, et c'est pour cette raison que toute étude évolutive comparative de ces objets ne peut se faire qu'entre des organismes très proches phylogénétiquement (Baumbach *et al.,* 2009). Les sites de fixation des facteurs de transcription évoluent très rapidement, de plus les facteurs de transcription orthologues régulent rarement des gènes orthologues (Price *et al.,* 2007) et même leurs types de régulation (activation ou inhibition) ne sont pas toujours les mêmes. Ainsi le réseau de régulation évolue par des changements de séquence des facteurs de transcription et des changements de sites de fixation. Des changements mineurs sur ces structures entraînent des remaniements du réseau. Il semblerait que les répresseurs évoluent de façon plus corrélée avec leurs cibles de régulation par rapport aux activateurs, un répresseur étant perdu qu'une fois que ses cibles sont perdues, alors qu'un activateur peut être perdu avant la disparition de ses cibles (Hershberg *et al.,* 2005; Hershberg and Margalit 2006).

Une corrélation négative entre la spécificité des sites de fixation des facteurs de transcription et leur pléiotropie a été observée (Lozada-Chávez *et al.,* 2008). Les facteurs de transcription pléiotropiques sont plus fortement conservés (van Hijum *et al.,* 2009). La dégénérescence des sites de fixation n'est pas la même pour chaque position du site. La dégénérescence peut s'expliquer soit par la nécessité d'une faible affinité d'association, nécessaire pour sa fonctionnalité, il s'agit donc dans ce cas d'une optimisation, soit par l'équilibre entre une faible sélection et la mutation et dans ce cas, il n'y pas d'optimisation car la seule nécessité est que l'affinité soit au-delà d'un seuil. Le deuxième scénario s'explique par le fait que la vitesse d'apparition d'une nouvelle mutation dans un site de fixation est plus importante que la vitesse d'élimination de ces mutations qui diminuent légèrement l'affinité du facteur de transcription à son site.

L'évolution des réseaux de régulation peut être approchée de plusieurs manières. On peut par exemple étudier l'ensemble des facteurs de

transcription et/ou des gènes cibles dans les génomes des bactéries séquencées, la conservation des interactions, la conservation des motifs structuraux des réseaux ou encore l'évolution des paramètres de structure des réseaux (*e.g.* densité, connectivité, etc.). Les facteurs de transcription sont moins conservés que leur gènes cibles (Babu *et al.,* 2007), le nombre de facteurs de transcription augmentant avec la taille du génome, et plus précisément il augmente proportionnellement avec le carré du nombre de gènes (Nimwegen 2006).

Dans une autre étude comparant le réseau de régulation d'*E. coli* à celui de 175 autres réseaux de régulation bactériens, Babu et al. (2006) ont montré qu'à une échelle évolutive faible à modérée, notamment dans les protéobactéries, l'évolution des réseaux de régulation bactériens reflète les mêmes relations entre les organismes que leur arbre phylogénétique. Par contre, au delà de cette échelle d'autres forces semblent contribuer à l'évolution des réseaux de régulation, puisque la classification des organismes éloignés en fonction de leur réseau de régulation n'est plus aussi corrélée à l'arbre phylogénétique, l'environnement de vie semble être l'une de ces forces.

5 Reconstruction de réseaux de la régulation de la transcription

Le séquençage à haut débit à marqué le début des années 90 (grands projets d'analyse génomique d'*E. coli*, de la levure et de l'humain) alors que les années 2000 ont vu l'avènement de la génomique fonctionnelle à large échelle (analyses transcriptomiques, tilling, banques de mutants). Le passage du séquençage à haut débit à la génomique fonctionnelle à large échelle[11] a marqué une nouvelle époque de la recherche en biologie moléculaire et préparé le terrain pour le développement de la biologie systémique dans ce domaine[12].

La biologie systémique ouvre la voie vers une approche globale d'étude des systèmes biologiques, et marque aussi le passage d'une modélisation « procédurale »[13] à une modélisation « déclarative »[14].

11 La notion de génomique fonctionnelle à large échelle se réfère au développement et aux applications des approches expérimentales globales afin de valider la fonction des gènes. Elle est caractérisée par des méthodologies expérimentales à large échelle ou à haut débit combinées avec l'analyse statistique des résultats (Hieter and Boguski, 1997).

12 Dès les années 60, les écologistes utilisaient déjà des approches systémiques pour comprendre l'organisation et le fonctionnement des écosystèmes (Frontier and Pichod-Viale, 1991).

13 Une méthode *procédurale* se concentre sur la séquence des étapes qui mène des données jusqu'aux conclusions (*e.g.* on classifie les gènes d'après leurs profils d'expression, ensuite, à l'intérieur de chaque classe on cherche des motifs surexprimés dans les promoteurs des gènes appartenant à cette classe) (Friedman, 2004).

14 Dans une approche *déclarative*, on commence par construire un modèle, pour faire référence à l'exemple donné plus haut, un modèle qui lie les promoteurs et les données d'expression. Ainsi, on

Avec une approche globale, on n'est cependant pas toujours capable de construire des modèles qui tiennent compte des mécanismes d'interaction entre tous les composants du système. On est par conséquent systématiquement amené à considérer de fortes simplifications et on parle dans ce cas de modèles "haut-niveau". Le niveau de détail d'un modèle est très coûteux en données. Ainsi, plus le système est grand, plus l'abstraction du modèle devra être importante.

La représentation la plus classique des systèmes cellulaires est un graphe (orienté ou pas), qu'on appelle aussi « réseau », mais le réseau cellulaire complet est d'une complexité telle, qu'il sort de la capacité d'analyse actuelle. Ainsi, en biologie des organismes, les réseaux d'interaction moléculaire constituent les principaux objets d'études de la biologie systémique.

Une première simplification consiste à ne prendre en compte que certains types de composants. On peut citer l'exemple de réseau de Paul Brazhnik (2002), dans lequel ne sont considérés que les gènes, les protéines et les métabolites. Cette simplification fait implicitement l'hypothèse que ce sont les molécules les plus importantes dans la cellule. Ce type de réseau macromoléculaire reste très complexe, des simplifications plus courantes sont généralement réalisées avec les réseaux de gènes, d'interaction protéine-protéine ou encore les réseaux métaboliques, qui ne contiennent qu'un seul type de molécules. En fonction des composants introduits dans l'étude et des données employées, un modèle ne peut retenir que certains types de processus.

Nous allons par la suite utiliser le terme de réseau de gènes, pour un objet qui dans la littérature est parfois appelé réseau de régulation génétique ou encore réseau de régulation transcriptionnelle. Ces termes font référence à un même type d'objet avec parfois des nuances (*e.g.*, certains chercheurs utilisent la notion de réseaux de régulation transcriptionnelle pour les réseaux dans lesquels seules les interactions du type facteur de transcription - promoteur du gène sont incluses, ce sont ce que Gardner appelle des modèles physiques car ils caractérisent de réelles interactions physiques entre les facteurs de transcription et les sites opérateurs des gènes qu'ils régulent (Gardner and Faith 2010). Les réseaux des gènes font donc référence à une très large gamme de modèles étudiant un des processus primordiaux de la cellule : la régulation de l'expression coordonnée des gènes produisant les protéines cellulaires. Les réseaux des gènes sont com-

---

décrit explicitement les hypothèses que l'on fait et les prédictions que l'on pourra faire avec le modèle construit (Friedman, 2004).

plémentaires des réseaux métaboliques, ces derniers décrivant la transformation et le transport des composés chimiques de la cellule médiés par les enzymes et les transporteurs.

Le processus d'expression d'un gène est la séquence de toutes les étapes (événements biochimiques) : du gène jusqu'à la protéine fonctionnelle, accomplissant une fonction spécifique dans le système. Ce processus peut être régulé à chaque étape : (1) l'initiation de la transcription ; (2) l'élongation de l'ARNm ; (3) la stabilité/dégradation de l'ARNm ; (4) l'initiation de la traduction ; (5) l'élongation du peptide ; (6) la dégradation de la protéine ; (7) le repliement/conformation de la protéine. Le raccourcissement ou le rallongement de cette liste en fonction du niveau de description du processus d'expression jugé pertinent et nécessaire amène à des modèles de plus haut ou de plus bas niveau. Comme nous venons de le dire, l'expression des gènes est un processus cellulaire qui fait intervenir des espèces moléculaires telles que l'ADN, les ARNm, les protéines et certains métabolites. Souvent les réseaux de gènes sont focalisés sur les sous-systèmes constitués seulement par les gènes et leurs ARNm. Bien que ce ne soit qu'une facette de la régulation, il s'agit déjà d'un grand nombre de processus complexes, impliqués dans la synthèse, la maturation et la dégradation de l'ARNm. Ainsi, on dit qu'il y a interaction entre deux gènes, si certains des produits de ces gènes interagissent entre eux (protéine-ADN, protéine-ARN, etc.). On constatera par la suite que les relations gène-gène des réseaux des gènes peuvent s'appuyer ou non sur une réalité physique, en fonction du modèle choisi. Ce qui est vrai pour la plupart des réseaux de gènes est que les attributs utilisés pour caractériser les gènes (les composants du système) sont le plus souvent les profils d'expression (le niveau d'ARNm du gène dans différentes conditions expérimentales, ou à différents instants dans une seule condition expérimentale).

L'étude des réseaux de gènes a pris une grande ampleur durant la dernière décennie. Néanmoins, on s'intéressait à ces objets avant même le séquençage des génomes entiers. Dès les années 60-70, Jacob et Monod avaient par exemple, tenté de décrire le mécanisme de régulation génétique de la synthèse des protéines. L'approche réductionniste de cette époque se concentrait sur les "parties" du système, comme certaines voies de synthèse ; l'étude partait d'un gène ou d'une protéine d'intérêt et s'étendait au fur et à mesure que l'on identifiait d'autres molécules interagissant avec la première. C'est-à-dire qu'on avançait de proche en proche, et un certain composant ne pouvait être étudié s'il n'interagissait pas avec un autre déjà connu et caractérisé. En plus de ne s'intéresser qu'à un nombre réduit de composants, on n'avait pas beaucoup d'informations globales concernant la communication entre les différentes "parties" du système. Cette façon

d'aborder les systèmes de régulation est appelée *approche ascendante* (ou *bottom-up*) et elle conduit à la construction de modèles mécanistiques, avec un haut niveau de détails moléculaires, mais largement incomplet quant à l'ensemble des constituants du système et donc également incomplet par rapport à l'ensemble des interactions.

Aujourd'hui les études des réseaux de gènes débutent majoritairement par des approches dites *descendantes* ou (*top-down*). Ces approches conduisent quant à elles à la construction directe d'un modèle global qui nécessite ensuite des études déductives afin de comprendre les mécanismes qui se mettent en place lors de la régulation.

Le nombre de modèles de réseaux de gènes a explosé dans la littérature de ces dernières années. Des classifications de types de modèles peuvent être faites en utilisant des critères tels que le modèle mathématique utilisé pour déduire et/ou le modèle mathématique utilisé pour représenter les interactions gène/gène, ou encore le type de données à partir desquelles le réseau a été construit. Nous allons utiliser la classification proposée par Schlitt et al. (2007), utilisant le type d'information contenu dans le réseau :

- les modèles de type « liste » collectionnent tous les composants du système de régulation considéré (gènes, annotation, facteurs de transcription, promoteurs) ;
- les modèles topologiques collectionnent les composants accompagnés d'interactions caractérisées en général de façon qualitative ;
- les modèles dynamiques sont les modèles les plus détaillés, et les plus proches du système réel car très souvent, il y a une réelle correspondance entre les composants physiques et les processus biochimiques du système avec ceux du modèle ; les interactions sont caractérisées quantitativement ; ces modèles sont utilisés pour la simulation et la prédiction.

Nous avons choisi cette classification, car elle peut être vue comme les étapes incontournables par lesquelles il est nécessaire de passer lorsqu'on construit un réseau de gènes par une approche descendante. À cause de la complexité des systèmes et des données toujours insuffisantes, on ne peut pas construire un modèle dynamique directement à partir des données. On suit plutôt un processus itératif : on commence par identifier les principaux acteurs, on identifie ensuite les interactions entre les constituants pour enfin caractériser les interactions grâce à un modèle dynamique. Le modèle dynamique permet de faire des simulations, des prédictions et donc de formuler des hypothèses, ainsi, il permet d'améliorer en retour les autres modèles (**Figure 11**).

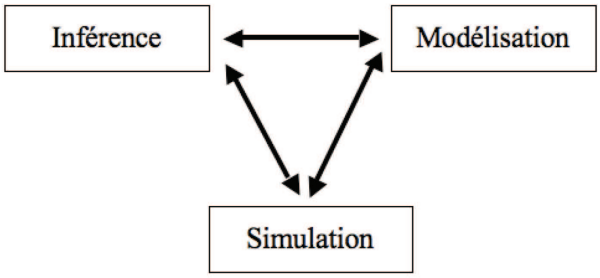


Figure 11. Etapes d'étude des réseaux de gènes. L'étude des réseaux de gènes se fait à travers un processus itératif passant par trois types d'analyse : l'inférence – construction d'un modèle brut censé représenter la structure des interactions du réseau ; la modélisation – à l'aide de la structure des interactions et de données cinétiques, on construit un modèle dynamique ; la simulation – l'utilisation des modèles dynamiques pour tester des hypothèses et pour analyser de nouvelles méthodes d'inférence.

Nous avons vu donc que l'étude et la construction des réseaux de gènes peuvent être abordées selon deux approches complémentaires : l'approche ascendante et l'approche descendante, le but ultime étant de déceler toutes les interactions causales entre les gènes et les produits de ces gènes régissant les régimes de transcription de la cellule. Si par l'approche ascendante on risque de perdre des parties du réseau faute d'éléments disponibles, avec l'approche descendante on risque d'introduire du bruit ou de fausses interactions. Des approches couplant et confrontant les résultats de ces deux types de construction de réseau semblent ainsi intéressantes à aborder. Le couplage signifie que le réseau doit simultanément être abordé conjointement par les deux approches, l'approche descendante permettant par exemple de pointer le type de composants cellulaires à étudier, alors que l'approche ascendante peut orienter le choix sur le type d'interactions moléculaires à introduire dans le modèle. Une autre façon de coupler les approches qui est actuellement beaucoup employée, consiste à utiliser les résultats obtenus avec l'approche ascendante, qui normalement sont considérés plus fiables, pour valider les résultats obtenus avec l'approche descendante.

5.1 Inférence des réseaux de gènes

Dans cette partie, nous avons concentré nos efforts vers une analyse des méthodes d'inférence de réseaux de gènes à partir des données d'expression afin de choisir une méthodologie adaptée et applicable à notre modèle biologique, *Buchnera*. Une étude préalable des méthodes de prédiction des ré-

seaux de gènes nous a rapidement orientés vers les méthodes d'inférence s'appuyant sur les modèles probabilistes graphiques.

5.1.1 Les réseaux de gènes

Les réseaux de gènes sont les modèles abstraits permettant de capturer le plus d'informations sur les réseaux cellulaires complets (Brazhnik *et al.,* 2002). C'est une propriété d'autant plus importante qu'il s'agit d'acquérir de l'information sur des processus qui ne sont pas mesurés physiquement. Ainsi, on mesure l'expression des gènes qui résultent de processus complexes d'interaction protéine-protéine, et protéine-ADN.

Lorsqu'on étudie un réseau de gènes, on peut dans un premier temps construire un modèle simple, qualitatif et descriptif, capturant seulement sa structure. Mais on peut également construire un modèle plus complet, avec des paramètres caractérisant sa dynamique, et utilisable en prédiction.

Elowitz *et al.* (2002) ont montré que l'expression des gènes est un processus stochastique provoquant donc des fluctuations importantes des paramètres qui caractérisent les réseaux de gènes, ce qui limite de façon drastique les possibilités d'inférence et de prédiction de leur dynamique. D'un autre côté, Dassow et al. (2002) ont montré que les réseaux de gènes sont des systèmes dynamiques très robustes aux fluctuations des paramètres numériques, même d'amplitude importante. On suppose que ce sont les propriétés topologiques de ces réseaux, qui permettraient de contrôler le bruit intrinsèque du système ce qui expliquerait donc la robustesse aux différentes fluctuations.

Sur *Buchnera*, notre approche est seulement descriptive, c'est-à-dire que nous cherchons à détecter les interactions intergéniques, mais pas à les quantifier. Nous nous sommes donc intéressés aux méthodes d'inférence de la structure des réseaux de gènes. Celles-ci sont divisées en deux classes (Gardner and Faith 2010) : (1) les méthodes physiques, construisant des réseaux de gènes dans lesquels les seuls régulateurs possibles sont les facteurs de transcription (interagissant physiquement avec le gène régulé) et (2) les méthodes d'influence, prenant en compte toutes les régulations possibles (à différents niveaux moléculaires de l'expression des gènes). Comme les informations sur les systèmes de régulation de *Buchnera* ne sont pas connues, nous avons adopté une approche probabiliste (les Modèles Graphiques) utilisant les principes de base des statistiques (les tests d'indépendance).

5.1.2 Les modèles probabilistes graphiques

5.1.2.1 Définition et description des modèles probabilistes graphiques

Les modèles probabilistes graphiques réunissent la théorie des graphes et celle des probabilités pour représenter et analyser les propriétés de la fonction de densité de probabilités jointes (DPJ) de grands ensembles de variables aléatoires (Jordan and Weiss 2002). Les MG possèdent deux composants, un composant structurel (le graphe) et un composant paramétrique (la DPJ).

Une loi de probabilité sur un ensemble de variables aléatoires définit un modèle d'indépendance[15]. La finalité du graphe est la représentation du modèle d'indépendance. Le problème est qu'un graphe représente des relations binaires, alors que le modèle d'indépendance suppose des relations ternaires. Ce qui permet de faire le lien entre le graphe et la DPJ est la notion de « séparation »[16]. Cette notion, ainsi que la propriété globale de Markov[17] établissant le lien entre le graphe et la DPJ de façon formelle, sont différentes suivant le cas où le graphe est orienté, comme dans les réseaux bayésiens[18], où la notion de dépendance n'est alors pas symétrique, ou non-orienté, comme dans les réseaux markoviens[19]. Par la suite, nous al-

15 *Définition*. Soit *E* un ensemble de variables aléatoires, un *modèle d'indépendance M*, est une liste de triplets *(A, B, C)* tels que *A* est indépendant de *B* si *C* est connu, avec la notation $A \perp B \mid C$ (Naïm, 2004).

16 *Définition*. Soit $G = (V, E)$ un graphe non-orienté ; pour tout triplet (A, B, C) de sous-ensembles disjoints de *G*, *A* est *séparé* de *B* par *C* dans *G* si et seulement si toute chaîne de nœuds de *A* vers *B* passe par un nœud de *C*.

17 *Définition*. Le graphe *G* et la loi de probabilité *p* vérifient la *propriété globale de Markov* si et seulement pour tout triplet $(A, B, C) \subset V$, *A*, *B* et *C* étant disjoints, *A* et *B* séparés par *C*, dans *G* $\Rightarrow A \perp B \mid C$.

18 *Définition*. Un *réseau bayésien* noté $B = (G, P)$, est défini par (1) un graphe acyclique orienté $G = (V, E)$, où *V* est l'ensemble des nœuds de *G*, et *E* est l'ensemble des arcs de *G* ; (2) un ensemble de variables aléatoires (X_i), associées aux nœuds du graphe : $P(X_1, X_2, ..., X_n) = \prod_{1 \le i \le n} P(X_i \mid C(X_i))$, où $C(X_i)$ est l'ensemble des causes (parents) de X_i dans le graphe *G*.

19 *Définition*. Un graphe vérifiant la propriété globale de Markov pour *p*, un DPJ, est un *réseau markovien* de *p*.

lons utiliser le formalisme des réseaux markoviens, il faut toutefois garder à l'esprit le fait qu'il existe des DPJ dont le modèle d'indépendance ne peut pas être représenté fidèlement par un graphe.

5.1.2.2 Les modèles probabilistes graphiques : une solution du problème de l'inférence de réseaux de gènes à partir des données d'expression

Notre cadre de travail est un réseau G de k gènes. Chaque gène est caractérisé par un seul attribut, le niveau du transcrit du gène, qui est modélisé par une variable aléatoire continue. Dans ce cas, la structure du réseau est caractérisée par le modèle d'indépendance de la DPJ de G, avec l'hypothèse que toute relation de dépendance entre deux gènes indique une interaction fonctionnelle. Le formalisme des réseaux markoviens semblerait être bien approprié pour l'inférence de la structure des réseaux de gènes. En effet, l'approche statistique de l'apprentissage de la structure d'un réseau markovien est basée sur la mise en place des tests d'indépendance conditionnelle[20] en s'appuyant donc sur des principes de base de la statistique (ce qui n'est entièrement valable pour les réseaux bayésiens). Les réseaux markoviens permettent une description synthétique de la DPJ, sous-jacente aux observations (Friedman 2004; Bishop 2006).

5.1.2.3 Les méthodes d'inférence de réseaux de gènes par l'inférence des modèles probabilistes graphiques sous-jacents aux données d'expression

Le but des méthodes d'inférence de la structure des MG est de déduire le graphe à partir des observations identiquement distribuées de la DPJ. Pour que ce graphe reste cohérent avec le modèle d'indépendance et le représente le plus fidèlement possible dans une approche statistique, on doit utiliser la dépendance conditionnelle. Ceci revient à tester pour chaque couple de gènes l'hypothèse d'indépendance, conditionnellement à l'ensemble des

---

20 *Définition*. Soient trois variables aléatoires *A, B* et *C*, on note $A \perp B \mid C$ et on dit que *A* et *B* sont *indépendantes conditionnellement* à *C* si et seulement si $\forall (a,b,c)\ P(A=a, B=b \mid C=c) = P(A=a \mid C=c) P(B=b \mid C=c)$.

autres gènes du réseau. Quelle que soit la mesure utilisée par le test d'indépendance adopté, elle sera une fonction de k variables (avec k le nombre de gènes dans le réseau). Or, le plus grand problème de l'inférence est justement le sous-échantillonnage de la loi jointe des profils d'expression (très peu d'observations par rapport au nombre de gènes étudiés). Ce problème devient de toute évidence encore plus sévère quand on augmente le nombre de gènes à prendre en compte. Pour surpasser ce problème d'échantillonnage, on peut, soit introduire des informations concernant le réseau, soit faire des hypothèses qui simplifieraient le modèle. Nous allons présenter brièvement trois approches de simplification du modèle.

Les MG Gaussiens, (MGG). L'approche utilisant les MGG fait l'hypothèse de normalité des profils d'expression des gènes. Pour tester l'indépendance entre deux gènes, il suffit de tester leur coefficient de corrélation partielle. Pour le calculer, on fait d'abord la régression des profils d'expression de chaque gène en fonction de tout l'ensemble des gènes, (excepté les gènes du couple testé). On calcule ensuite le coefficient de corrélation entre les résidus des régressions multiples de chaque gène. Mathématiquement, cette étape est résolue grâce à l'inversion de la matrice de covariance. Le problème du sous-échantillonnage n'est pas pour autant résolu, car pour pouvoir inverser la matrice de covariance, il faudrait au moins autant d'observations que de gènes (Wille and Buhlmann 2006). Des améliorations de la méthode d'estimation de l'inverse ont été proposées, mais l'ordre du nombre d'observations reste trop grand pour les possibilités actuelles d'expérimentation.

Les graphes de co-expression. Cette approche, tout comme celle des MGG, fait implicitement l'hypothèse de normalité des profils d'expression des gènes, et donc de la linéarité des relations pouvant exister entre les concentrations des transcrits, si la mesure adoptée pour les tests d'indépendance est le coefficient de corrélation. L'inférence des graphes de co-expression emploie pour les tests d'indépendance, des mesures de la dépendance qui sont seulement fonction de deux variables, ce sont des tests d'indépendance simple et non d'indépendance conditionnelle. Les graphes de co-expression sont aussi appelés graphes 0, ou graphes d'ordre 0, car le conditionnement se fait par rapport à 0 autres variables. On s'éloigne ici du formalisme des MG, car les graphes 0 ne donnent pas une bonne idée de la vraie allure du réseau de gènes, étant donné qu'ils ne sont pas fidèles au modèle d'indépendance de la DPJ de G. Ces graphes sont beaucoup plus denses, on a donc beaucoup de faux positifs parmi les liaisons inférées. Cette approche a quand même un grand avantage : la fiabilité de l'estimation ne dépend plus du nombre de gènes. Comme mesure de la dépendance, en plus du coefficient de corrélation, on dispose de l'information mutuelle

(Butte and Kohane 2000) (quantité introduite par la théorie de l'information). Avec ce dernier coefficient aucune hypothèse n'est faite quant à la nature de la distribution du profil d'expression.

Les graphes 0-1. C'est une approche intermédiaire entre les deux précédentes. On utilise les mesures de la dépendance conditionnelle, mais cette fois en conditionnant par rapport à un seul autre gène. L'idée est que pour chaque couple de gènes on fait le test d'indépendance, conditionnellement à un troisième gène, si parmi les *k-2* gènes restants du réseau, il y en a au moins un pour lequel l'hypothèse d'indépendance conditionnelle est acceptée, on infère l'indépendance entre les gènes du couple. Évidemment, les graphes inférés ne représentent pas fidèlement le modèle d'indépendance de la DPJ non plus, mais ils ne sont pas très éloignés, voire même exacts dans certains cas (*e.g.* si les variables aléatoires sont gaussiennes, et si le graphe complet est creux, alors le graphe 0-1 est le graphe complet (Wille and Buhlmann 2006)). Comme pour l'approche précédente, on peut utiliser le coefficient de corrélation ou celui de l'information mutuelle.

Partie II

Matériels et Méthodes

1 Données utilisées au cours de la thèse

1.1 Données génomiques

Dans nos études nous avons utilisé les génomes de 4 *Buchnera* issues des pucerons du pois (BAp), du blé (BSg), du pistachier (BBp) et du cèdre (BCc) :

- *Buchnera aphidicola* str. Aps (*Acyrthosiphon pisum*) - BAp, GenBank: BA000003, Refseq : NC_002528 (Shigenobu *et al.,* 2000) ;
- *Buchnera aphidicola* str. Sg (*Schizaphis graminum*) - BSg, GenBank: AE013218, Refseq : NC_004061 (Tamas *et al.,* 2002) ;
- *Buchnera aphidicola* str. Bp (*Baizongia pistaciae*) - BBp, GenBank: AE016826, Refseq : NC_004545 (Van Ham *et al.,* 2003) ;
- *Buchnera aphidicola* str. Cc (*Cinara cedri*) - BCc, GenBank: CP000263, Refseq : NC_008513 (Perez-Brocal *et al.,* 2006) ;

Les annotations de BAp, BSg et BBp ont été obtenues à partir de la base BuchneraBASE (Prickett *et al.,* 2006).

Nous avons aussi utilisé la séquence d'*Escherichia coli* str. K-12 substr. MG1655, GenBank : U00096, Refseq : NC_000913 (Blattner *et al.,* 1997; Riley *et al.,* 2006). En ce qui concerne son annotation, ainsi que son réseau de régulation et ses unités de transcription, elles ont été récupérées à partir de RegulonDB 6.2 (Gama-Castro *et al.,* 2008).

1.2 Données d'annotation fonctionnelle des gènes

1.2.1 L'annotation Gene Ontology

Les annotations des termes GO ont été récupérées à partir de la base de données UniProtKB-GOA[21]. Une fonction R a été développée afin de transformer une annotation GO, en une annotation GO de niveau choisi (*h*), *i.e.* chaque terme GO de l'annotation initiale est remplacé par tous ses parents du niveau *h*, si le niveau du terme est supérieur à *h*, si le niveau du terme est inférieur à *h*, alors aucune transformation n'est faite et le terme est éliminé de l'annotation de niveau *h*. Il est considéré que les niveaux 3 et 4 sont des bons choix de niveau, n'entraînant d'ailleurs que peu d'éliminations (Al-Shahrour *et al.,* 2004).

1.2.2 Les classes métaboliques des gènes

Nous avons utilisé dans nos analyses les fonctions métaboliques des gènes (il s'agit bien évidemment d'un abus de langage couramment utilisé pour désigner les fonctions métaboliques des protéines qu'ils codent). La classe métabolique des gènes d'*E. coli* est celle utilisée par Seshasayee et al. (2009). Cette classification comportant 3 classes : anabolisme, catabolisme et métabolisme de l'énergie. L'annotation de la classe métabolique des gènes de *Buchnera* a été déduite par orthologie, à partir de la classification des gènes d'*E. coli*.

1.2.3 L'annotation PFAM[22] (Protein families)

PFAM est une annotation fondée sur des collections d'alignements multiples et de profils de Markov cachés qui ont permis de déduire des familles de protéines. Cette annotation fournit des informations, soit sur la structure, soit sur la fonction moléculaire des protéines, soit les deux aspects. L'annotation PFAM des protéines de BAp a été récupérée à partir de

[21] ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/

[22] http://pfam.sanger.ac.uk/

KEGG[23]. Nous avons utilisé cette annotation pour chercher des protéines faisant partie des familles protéiques impliquées dans la régulation transcriptionnelle.

1.2.4 La classification des protéines selon les fonctions définies par Riley et al. (1998)

En plus des annotations de fonction des gènes mentionnées plus haut, nous avons également utilisé l'annotation fonctionnelle des gènes d'*E. coli*, faite par Riley et al. (Riley 1993, 1998). La version initiale de cette classification des protéines associait à chaque gène une seule fonction : celle qui est considérée la plus importante et la plus représentative du gène, c'est d'ailleurs cette annotation que nous avons utilisée. La version ultérieure, associant plus d'une fonction par gène porte le nom MultiFun[24]. L'annotation des gènes de *Buchnera* a été déduite de celle d'*E. coli* par orthologie.

1.3 Données d'expression des gènes de *Buchnera APS*

Trois ensembles de données d'expression obtenues grâce à la puce *Buchnera* conçue dans l'UMR BF2I par Fédérica Calevro (2004) ont été utilisés dans nos travaux.

Les premières données ont été obtenues par Nancie Reymond, durant sa thèse au BF2I. Elles ont permis de mesurer la variation de deux facteurs dans l'environnement du puceron sur la réponse transcriptionnelle de *Buchnera APS*. Ces deux facteurs, la concentration en saccharose et le rapport d'acides aminés essentiels et non-essentiels dans le milieu nutritif du puceron, ont permis d'étudier le rôle de la régulation transcriptionnelle de *Buchnera APS* dans la relation trophique. En effet, les changements du rapport en acides aminés essentiels/non-essentiels permettent de faire varier qualitativement la demande nutritionnelle en acides aminés de son hôte, et la variation de la concentration en saccharose permet de faire varier quantitativement la force de cette demande en jouant sur l'appétence et donc sur le taux de croissance du puceron. Le plan d'expérience a été conçu de façon

[23] ftp://ftp.genome.jp/pub/kegg/genes/organisms/buc/

[24] http://genprotec.mbl.edu/

à étudier l'effet de chacun des facteurs, ainsi que de leurs interactions (Reymond 2004).

Le deuxième jeu de données a été obtenu par José Viñuelas, également durant sa thèse au sein de l'UMR BF2I sous la direction de F. Calevro. Dans ce cas, le niveau de transcription de chaque gène a été mesuré et normalisé grâce à une hybridation de la puce en utilisant de l'ADN génomique (ADNg), permettant ainsi la comparaison des niveaux de transcription entre les gènes d'une même expérience (Viñuelas 2008).

Enfin, le troisième jeu de données a été acquis par John Bermigham et Tom Wilkinson en collaboration avec l'UMR BF2I. Dans cette expérience, ils ont comparé le niveau de transcription des gènes des *Buchnera APS* provenant des embryons de pucerons à différents stades de développement et des *Buchnera* provenant des bactériocytes maternels.

2 Approche expérimentale : validation des unités de transcription de *Buchnera APS* par RT-PCR

Afin de confirmer expérimentalement certains opérons prédits de *Buchnera*, nous avons utilisé le protocole de RT-PCR conçu à cet effet par Charaniya *et al*. (2007).

Les cellules de *Buchnera* ont été purifiées à partir de 900 mg de pucerons selon la procédure décrite par Charles et al. (1999a). L'ADNg total en a été extrait grâce au kit QIAamp DNA Mini Kit (Qiagen, Helden, Germany). L'ADNg a été employé pour la détermination des conditions de PCR et en tant que contrôle positif des réactions de RT-PCR. Les ARN totaux ont été isolés et purifiés avec le kit RNeasy kit (Qiagen), comme cela a été décrit par Calevro et al. (2004). La qualité, la pureté et la concentration des ARN ont été vérifiées par une mesure au spectrophotomètre NanoDrop® ND-1000 et par électrophorèse sur gel d'agarose dénaturant. Les ARN totaux ont alors été traités avec de la DNase - Turbo DNA-free™ DNase (Ambion, Austin, TX, USA).

La réaction de retrotranscription a été réalisée à partir de 1 µg d'ARN, des amorces aléatoires (hexamères aléatoires) et de la SuperScript™ III, selon le protocole du kit SuperScript™ First-Strand Synthesis system kit for the RT-PCR (Invitrogen, Paisley, UK). L'addition de 1 µl de RNase H à la fin de la rétrotranscription et l'incubation durant 20 min à 37°C, ont permis d'éliminer les ARN initialement présents dans la solution. Pour chacune des réactions de RT-PCR, 2 µl du produit de la réaction RT ont été utilisés.

Des paires d'amorces spécifiques ont été construites avec le logiciel Oligo 6 (Molecular Biology Insight, Inc) pour chaque produit que nous avons voulu amplifier. Les réactions PCR ont été réalisées à partir d'ADNc issu des ARN totaux. Le contrôle négatif de la RT-PCR a été fait en utilisant le produit d'une réaction de RT pour laquelle la transcriptase reverse n'a pas été ajoutée et le contrôle positif a été fait en utilisant de l'ADNg au lieu de l'ADNc. Les réactions PCR ont été réalisées avec le kit AccuPrimeTM Taq DNApolymerase high fidelity (Invitrogen), permettant l'amplification de grands fragments, pouvant aller jusqu'à 20kb. Les conditions utilisées pour la PCR sont les suivantes : 30 sec à 94°C suivies de 36 cycles d'amplification-dénaturation (dénaturation 30 sec à 94°C, hybridation 30 sec à 47°C ou 43.5°C, et extension 2 à 5 min à 68°C). Le volume total de chaque réaction a été de 50 µl, dont 10 ont été utilisés pour la mi-

gration sur un gel d'agarose à 1%. La présence de certains produits a été testée en utilisant 26 cycles.

3 Algorithmes utilisés au cours de la thèse

3.1 Le module Bprom pour la prédiction des promoteurs

L'outil Bprom[25] (Bacterial Promoter Prediction Program) de SoftBerry a été utilisé chez *Buchnera* pour prédire les promoteurs des gènes reconnus par le facteur σ^{70} ainsi que leur site d'initiation de la transcription. La méthode de Bprom repose sur une fonction discriminante linéaire, combinant des caractéristiques des promoteurs σ^{70} connus. Bprom a une précision de 80% chez *E. coli*. Sa spécificité a aussi été estimée à 80%, en mélangeant des séquences contenant des promoteurs σ^{70} avec des séquences ne contenant pas ces promoteurs. Dans notre travail sur *Buchnera APS*, les promoteurs ont été recherchés dans les 500 pb en amont de chaque séquence codante, et non sur l'intégralité du génome, afin d'augmenter la spécificité de la prédiction.

3.2 Le logiciel MacVector pour la prédiction des promoteurs σ^{32}

Nous avons utilisé la fonction de recherche de sous-séquences du logiciel MacVector pour rechercher les promoteurs σ^{32} de *Buchnera APS*.

[25] http://linux1.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb

L'avantage de la fonction de recherche de MacVector est que le motif recherche peut contenir deux parties, séparées par une distance variable que l'on peut spécifier. Enfin, on peut aussi paramétrer les positions conservées du motif. Les promoteurs σ^{32} sont connus pour être constitués de deux boîtes : CTTGAAAA et CCCCTNT, séparées par 11 à 15 pb (Gross 1996; Wilcox *et al.,* 2003). Ainsi, pour la recherche des promoteurs σ^{32} dans le génome de BAp, nous avons utilisé trois contraintes : (1) le promoteur se trouve dans les 500 pb en amont du codon start du gène ; (2) chaque boîte montre au minimum 50 % d'identité avec la séquence consensus lui correspondant ; (3) au minimum 50 % du GC de la séquence consensus est conservé.

3.3 Recherche des sites de fixation des facteurs de transcription

Etant donné que nous avons recherché les sites de fixation de facteurs de transcription généralistes (les protéines NAP), qui sont donc dégénérés, nous avons opté pour une recherche simple grâce aux fonctions R *words.pos* de la librairie Seqinr (Charif *et al.,* 2010) et *gregexpr* de la librairie Base. Ces deux fonctions permettent de faire une recherche de mots en utilisant des expression régulières (modèles permettant de représenter un ensemble de mot, *e.g.* « [AT]T » cherchera tous les mots de deux lettres commençant par A ou par T, et finissant avec T). La fonction *words.pos* contrairement à *gregexpr*, détecte les motifs chevauchants.

3.4 Propriétés structurelles séquence-dépendantes du chromosome

Quatre propriétés structurelles du chromosome ont été estimées *in silico* : le SIDD, la courbure, l'angle de torsion et l'énergie d'empilement des bases. Les algorithmes utilisés pour l'estimation de ces quatre paramètres ont été calibrés à partir de données expérimentales et de la composition en bases des séquences utilisées dans les expériences (à l'exception de l'énergie d'empilement des bases) (Ornstein *et al.,* 1978; Shpigelman *et al.,* 1993; Benham 1996; Hassan and Calladine 1996; Benham and Bi 2004). Pour l'estimation des paramètres le long du chromosome d'intérêt, seule sa séquence nucléotidique est nécessaire au calcul grâce à ces méthodes. Les programmes permettant de calculer ces quatre propriétés (la courbure, (Shpigelman *et al.,* 1993), l'énergie d'empilement des bases (Ornstein *et al.,* 1978), l'angle de torsion (Hassan and Calladine 1996) et le SIDD (Benham and Bi 2004)) font partie des modules regroupés dans l'outil inte-

ractif développé par Hallin et al. (2009), dans un outil nommé GeneWiz[26], qui en plus de calculer différentes propriétés des séquences d'ADN permet la visualisation dynamique de ces propriétés sous forme d'atlas.

Pour chaque chromosome et pour chacun des quatre paramètres, nous avons une valeur pour chaque paire de base. Ces valeurs ont été utilisées de plusieurs manières dans nos études.

Pour construire des distributions globales d'une propriété structurelle d'un chromosome bactérien entier (cf. Introduction, §4.3.6), comme Pedersen et al. (2000), nous avons utilisé les valeurs moyennes de la propriété, calculées sur des fenêtres de 300 pb non chevauchantes et couvrant la totalité du chromosome.

Afin d'évaluer les distributions des propriétés structurelles pour une régions génique particulière (cf. Introduction, §4.3.6), nous avons utilisé des valeurs moyennes, calculées sur des fenêtres non chevauchantes, de 20 pb, couvrant un type de régions géniques.

Une analyse par région promotrice des propriétés structurelles a également été faite. Nous appelons région promotrice, une région génomique de taille fixée située directement en 5' d'une séquence codante. Si la taille de cette région en 5' est de x pb (x pb en amont du codon start du gène), nous allons utiliser la notation RP_x pour la désigner. Les RP_{150} nous ont paru un bon choix étant donné que le logiciel Bprom prédit généralement la présence d'un promoteur σ^{70} dans les 120 pb en amont du codon start, ce choix a été d'ailleurs confirmé, par les profils des propriétés structurelles (cf. Résultats, §2.3.3). Si à la place de prendre une région de taille fixée, nous prenons toute la région non-codante en 5' de la séquence codante, nous allons utiliser la notation RP_{ig}. Dans les analyses des régions promotrices des gènes (cf. Résultats, §2.3.3), nous avons calculé une seule valeur par région promotrice et par propriété structurelle, en utilisant la moyenne des valeurs des positions occupées par la région promotrice dans le cas de la courbure, de l'énergie d'empilement et de l'angle de torsion, et le minimum des valeurs de ces positions pour le SIDD. La fonction minimum avait déjà été utilisée pour le SIDD par Wang et al. (2008).

L'analyse spectrale des valeurs des propriétés structurelles en parallèle avec l'analyse spectrale des données d'expression a nécessité des données homogènes en terme d'échelle (mesures effectuées à intervalle régulier le long du chromosome). Dans ce but, nous avons construit des vec-

26 http://www.cbs.dtu.dk/services/gwBrowser/

teurs de valeurs des propriétés structurelles et d'expression en moyennant sur des fenêtres non chevauchantes de 100 pb le long du chromosome. Pour l'expression, il s'agit d'une moyenne pondérée par la proportion du/des gène(s) contenus dans la fenêtre de 100 pb.

3.5 Le programme Helix-Turn-Helix (HTH) pour la recherche de domaines protéiques hélice tour hélice

La recherche du motif hélice-tour-hélice dans le protéome de BAp a été réalisée grâce aux programme HTH (Dodd and Egan 1990; Combet *et al.,* 2000). Ce programme parcourt les séquences des protéines par morceaux en livrant en sortie la position et la séquence en acides aminées du motif HTH potentiel trouvé, avec un score de prédiction et une probabilité de présence associés.

3.6 Le programme CCCPart (C3P) pour l'analyse des réseaux d'interaction

Afin d'étudier les forces d'évolution qui ont modelé l'organisation du chromosome de *Buchnera APS*, nous avons utilisé le programme C3P[27] développé par Boyer et al. (2005). Ce programme implémente un algorithme permettant de trouver les composantes connexes communes à deux graphes. Le concept de composante connexe commune (CCC) à deux graphes est expliqué et visualisé dans la **Figure 12**.

[27] http://www.inrialpes.fr/helix/people/viari/cccpart/

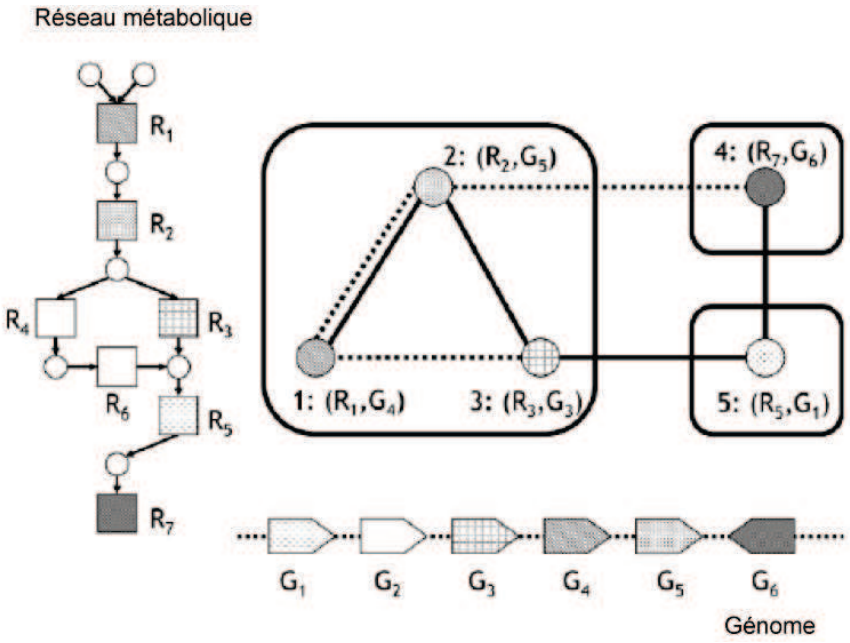


Figure 12. Illustration de la notion de multigraphe et de la notion de composante connexe commune d'un multigraphe. Dans le premier graphe, le réseau métabolique, les gènes sont reliés par des réactions (les réactions que les enzymes codées par les gènes, catalysent). Dans le deuxième graphe, le génome, les gènes sont liés par leur voisinage sur le chromosome. Le multigraphe est alors défini de la façon suivante : un ensemble de nœuds - les gènes, et deux ensembles d'arêtes, (1) les arêtes reliant les gènes dans le premier graphe, représentées en traits continus, dans ce cas les gènes correspondent aux enzymes catalysant les réactions, l'association gène-réaction est marquée sur le multigraphe, *e.g.* le gènes 5 code pour l'enzyme catalysant la réaction 2), (2) les arêtes reliant les gènes dans le deuxième graphe, représentées en traits pointillé. Une composante connexe commune d'un multigraphe est un sous-ensemble maximal de nœuds connectés (chaque sommet du sous-ensemble est atteignable à partir de tout autre sommet du sous-ensemble) dans le premier ensemble d'arêtes, comme dans le deuxième. Dans le multigraphe de cet exemple nous avons trois composantes connexes communes {1,2,3}, {4} et {5}, entourées par des rectangles (d'après Boyer et al. (2005)).

Nous pouvons donc constater sur la **Figure 12**, que les gènes 3, 4 et 5 sont voisins sur le chromosome, et que leurs produits sont impliqués dans des réactions qui se succèdent dans le réseau métabolique. En comparant les gènes colocalisés sur le chromosome et connectés dans certains réseaux d'interaction moléculaire, nous avons cherché à détecter les forces d'évolution contraignant l'arrangement des gènes sur le chromosome (cf. Résultats, §1.2). Avec le programme C3P, nous avons la possibilité de relâ-

cher certaines contraintes, comme par exemple la contrainte de connectivité. Ainsi, au lieu d'exiger que les gènes soient directement connectés sur le chromosome, nous pouvons choisir de contraindre la distance entre ces gènes, (*e.g.* considérer que deux gènes sont connectés s'ils sont séparés par moins de quatre gènes sur le chromosome).

Nous avons utilisé C3P pour calculer les CCC entre le génome (les gènes ne devant pas être séparés sur le chromosome par plus d'un gène) et le réseau métabolique – les métabolons, entre le génome et le réseau d'interaction protéine protéines – les interactons (les gènes ne devant pas être séparés sur le chromosome par plus d'un gène) et entre le génome et le réseau de régulation – les transcriptons (les gènes ne devant pas être séparés sur le chromosome par plus d'un gène) ; ceci chez *E. coli* et chez *Buchnera APS*. Les réseaux de *Buchnera APS* ont été déduits par orthologie à partir des réseaux d'*E. coli*. Enfin, nous avons déterminé les synthons entre *Buchnera APS* et *E. coli* qui sont les CCC de leurs génomes respectifs (les gènes ne devant pas être séparés sur le chromosome par plus de quatre gènes).

4 Algorithmes développés au cours de la thèse

4.1 La méthode DisTer pour la prédiction des unités de transcription chez *Buchnera APS*

Nous avons développé une méthode de prédiction des unités de transcription supervisée, que nous avons appelée DisTer. Il s'agit d'un prédicteur Bayesien naïf utilisant comme données d'entrée la distance intergénique et les terminateurs Rho-indépendant prédits par TransTermHP (Kingsford *et al.,* 2007) et entraîné sur *E. coli*. Comme la plupart des prédicteurs d'opérons, il s'applique sur l'ensemble des paires de gènes adjacents tandem du génome, en classant chacune des paires soit comme appartenant à la même UT (MUT) soit comme appartenant à des UT distinctes (UTD).

DisTer a été choisi parmi les trois modèles que nous avons testés et comparés. Chaque modèle se distingue par l'ensemble de données qu'il utilise et par son calcul de la probabilité qu'une paire de gènes adjacents soit de type MUT (**Tableau 4**).

Pour estimer la propriété a priori qu'une paire de gènes soit du type MUT, nous avons considéré que le nombre de gènes dans une unité de transcription suit une loi géométrique, $P(\mathrm{L_{TU}} = n) = P(\text{paire MUT})^{n-1}(1 - P(\text{paire MUT}))$, ce modèle étant le modèle statistique le plus simple. L'espérance de cette loi géométrique est $\frac{1}{1 - P(\text{paire MUT})}$. Nous avons alors estimé la probabilité $P(\text{paire MUT})$ à partir de la taille moyenne des UT d'*E. coli* ($\overline{L}_{UT}$) : $P(\text{paire MUT}) = \frac{\overline{L}_{UT} - 1}{\overline{L}_{UT}} = 0.53$ (la courbe de la densité de la loi géométrique ainsi estimée est représentée sur la **Figure 20**).

Tableau 4. Modèles de prédicteurs d'opérons étudiés pour le choix de DisTer.

1. Le premier modèle utilise la distance intergénique (D) et la simple présence/absence d'un terminateur de transcription (T), avec l'hypothèse que ces deux propriétés sont indépendantes.

$$p(\text{MUT} \mid D=d, T=t) = \frac{p(\text{MUT}, D=d, T=t)}{p(D=d, T=t)}$$
$$= \frac{p(D=d, T=t \mid \text{MUT})\,p(\text{MUT})}{p(D=d, T=t \mid \text{MUT})\,p(\text{MUT}) + p(D=d, T=t \mid \text{UTD})\,p(\text{UTD})}$$
$$= \frac{p(D=d \mid \text{MUT})\,p(T=t \mid \text{MUT})\,p(\text{MUT})}{p(D=d \mid \text{MUT})\,p(T=t \mid \text{MUT})\,p(\text{MUT}) + p(D=d \mid \text{UTD})\,p(T=t \mid \text{UTD})\,p(\text{UTD})}$$

2. Le deuxième modèle utilise la distance intergénique (D) et la présence/absence du terminateur (T), mais sans faire l'hypothèse d'indépendance entre ces deux propriétés des paires de gènes.

$$p(\text{MUT} \mid D=d, T=t) = \frac{p(D=d, T=t \mid \text{MUT})\,p(\text{MUT})}{p(D=d, T=t \mid \text{MUT})\,p(\text{MUT}) + p(D=d, T=t \mid \text{UTD})\,p(\text{UTD})}$$

3. Le troisième modèle utilise la distance intergénique (D) et le score de prédiction (TransTermHP, T_{score}) du terminateur de transcription.

$$p(\text{MUT} \mid D=d, T_{score}=s) = \frac{p(D=d, T_{score}=s \mid \text{MUT})\,p(\text{MUT})}{p(D=d, T_{score}=s \mid \text{MUT})\,p(\text{MUT}) + p(D=d, T_{score}=s \mid \text{UTD})\,p(\text{UTD})}$$

Avec ce type de méthode de prédiction la classe normalement attribuée à un objet, ici la paire de gènes adjacente tandem, est celle ayant la plus grande probabilité. Dans notre cas, comme seules deux classes sont possibles, MUT ou UTD, une paire de gènes devrait être classée comme MUT si la probabilité $P(\text{MUT} \mid \text{propriétés de la paire}) > 0.5$, et UTD dans le cas contraire. Au lieu de procéder de cette façon, nous avons cherché le seuil de probabilité le plus discriminant, en utilisant comme indice la précision et la valeur prédictive du modèle. Des valeurs comprises entre 0.05 et 1 ont été testées pour le seuil de probabilité, sur l'ensemble d'entrainement, ce qui nous a permis d'évaluer le taux d'erreur de chaque modèle sur l'ensemble d'entraînement (le même jeu de données est utilisé pour l'entraînement et pour la prédiction), associé à chacune des valeurs seuil. Ce taux d'erreur est un bon indicateur de l'incertitude de la règle de classification (**Figure 13**, à gauche et **Figure 14**). La qualité des prédictions a été évaluée à travers la sensibilité (la proportion de vraies paires MUT correctement prédites) et la spécificité (la proportion de vraies paires UTD correctement prédites).

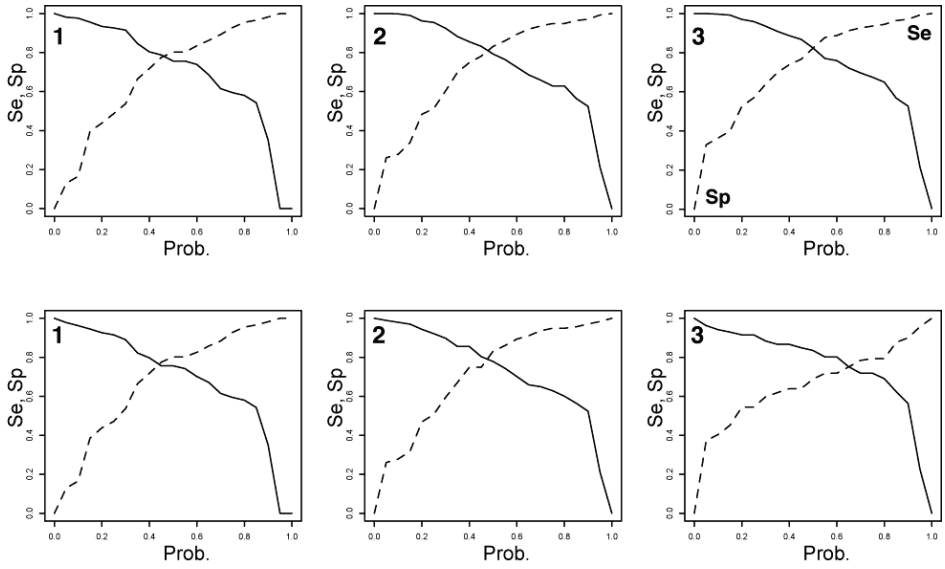


Figure 13. Qualité des prédictions faîtes avec les trois modèles de prédiction des unités de transcription, en fonction de la valeur seuil de la probabilité à partir de laquelle une paire est classée MUT. La sensibilité (Se) et la spécificité (Sp) représentées dans les cadres de la première ligne, ont été calculées dans le cas où le même ensemble de données est utilisé pour l'entraînement et pour la prédiction. La deuxième ligne représente les capacités prédictives des modèles, calculées avec la technique « leave-one-out » (le modèle est entraîné sur *n-1* paires, pour ensuite être utilisé pour déterminer la classe de la *n*ième paire).

La valeur prédictive de chaque modèle a été testée également pour chaque valeur seuil de probabilité en utilisant deux techniques différentes. Dans la première approche, l'ensemble d'unités de transcription connues d'*E. coli* est séparé en un ensemble d'entraînement (80% des paires) et un ensemble de vérification (20% des paires). La sensibilité et la spécificité sont donc calculées à partir des prédictions faites par le modèle sur l'ensemble de vérification (**Figure 14**, au milieu). La deuxième technique, appelée « leave-one-out » est une approche itérative dans laquelle chaque paire de gènes de l'ensemble est exclue de l'ensemble durant une itération. Le modèle est donc entraîné sans l'utiliser. Ensuite, le modèle ainsi entraîné est utilisé pour déterminer la classe de la paire exclue de l'entraînement (**Figure 14**, à droite).

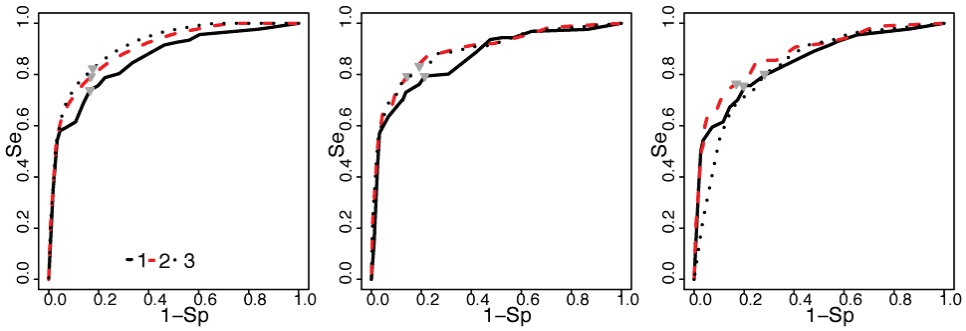


Figure 14. Courbes ROC mesurant le taux d'erreur d'entraînement (à gauche), la performance des modèles lorsqu'ils sont entraînés sur 80% de l'ensemble de données et que leur prédiction est faîte sur les 20% restants (milieu) et la performance des modèles estimée grâce à la technique « leave-one-out » (droite). Les lignes rouges représentent les performances du modèle 2, qui a été finalement choisi pour prédire les UT de *Buchnera* (DisTer). Les triangles gris représentent pour chaque courbe son point le plus proche de coin gauche supérieur.

Sur la **Figure 14**, pour chaque courbe, le point le plus près du coin supérieur gauche (indiqué par un triangle gris) est la configuration minimisant la somme des carrés des proportions des prédictions erronées. Nous avons considéré que le seuil de probabilité correspondant à ce point permet la meilleure performance. Les meilleurs seuils de probabilité, donnés par les différentes techniques d'évaluation des performances du modèle ne sont pas les mêmes, excepté pour le deuxième modèle (**Figure 14**). On peut remarquer que le deuxième modèle est le plus performant, sauf lorsque l'intégralité de l'ensemble de données d'*E. coli* est utilisé pour l'entraînement. Ce résultat était attendu, étant donné que le troisième modèle divise l'ensemble de données en plus de classes que le second. Ainsi, le second modèle apparaît comme ayant la meilleure capacité prédictive (avec Se = 79% et Sp = 83% comme performance d'apprentissage et Se = 78%, Sp = 84% comme performance). C'est donc bien le second modèle avec une valeur seuil de probabilité de 0.5 qui a été utilisé pour prédire les unités de transcription de *Buchnera APS*. Tous les calculs ont été faits avec le programme R 2.6.1[28].

[28] http://cran.r-project.org/

4.2 IGOIM : une méthode d'inférence des réseaux de gènes utilisant l'Information Mutuelle conditionnelle

4.2.1 Définition et propriétés de l'Information Mutuelle (IM)

L'IM[29] a été proposée par Shannon dans le cadre de la théorie de l'information. Les propriétés de cette mesure ont été bien étudiées (Steuer *et al.,* 2002; Brillinger 2005). La propriété la plus intéressante est sans doute que de par sa définition, l'IM est nulle, si et seulement si les deux variables sont indépendantes. Cette propriété est générale pour l'IM alors qu'elle ne s'applique au coefficient de corrélation que si les deux variables sont gaussiennes.

L'IM est aussi invariante par des transformations injectives[30], elle ne dépend pas des valeurs mesurées, mais des probabilités de ces valeurs. C'est une propriété importante dans la mesure où on ne travaille jamais sur la mesure réelle des concentrations des transcrits, mais sur une mesure relative plus ou moins distordue par la normalisation liée à la technique d'acquisition de données. L'IM est ainsi moins sensible que la corrélation aux points pivots par exemple et plus robuste au bruit.

Si du point de vue théorique, il paraît plus intéressant de choisir l'IM comme mesure de la dépendance plutôt que la corrélation, du point de vue pratique, l'estimation de l'IM à partir des données est un problème difficile. Par exemple, il n'existe pas un estimateur universellement accepté et non-biaisé pour les cas de sous-échantillonnage.

4.2.2 Les estimateurs de l'IM

29 *Définition de l'IM*. Considérons deux variables aléatoires *X* et *Y*, avec une distribution de probabilités jointes $p(x,y)$ et des distributions marginales $p(x)$ et $p(y)$. L'IM de *X* et *Y* se définit de la façon suivante : $I(X,Y) = \iint_{xy} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \mathrm{d}x\mathrm{d}y$.

30 Considérons le profil d'expression du gène 1 - g_1 et celui de g_2, et puis deux fonctions injectives *u* et *v*, alors $I(g_1,g_2) = I(u(g_1),v(g_2))$.

Les estimateurs de l'IM peuvent être divisés en deux grandes classes, les estimateurs paramétriques et non-paramétriques. Les plus intéressants pour nous sont les estimateurs non-paramétriques, car ils ne font aucune hypothèse quant à la nature de la loi des profils d'expression. La classe des estimateurs non-paramétriques se divise à son tour en deux classes. Tout d'abord, les estimateurs qui utilisent les observations telles qu'elles sont (des valeurs continues), pour estimer par intégration les fonctions de densité avec des fonctions à noyaux. Cette procédure est lourde du point de vue de la complexité des calculs, et de plus, nous n'avons pas trouvé d'études sur les propriétés de ces estimateurs dans les cas de sous-échantillonnage. La deuxième sous-classe d'estimateurs non-paramétriques est la classe des estimateurs utilisant une discrétisation des variables[31].

Un des avantages de l'approche discrète, non-paramétrique de l'estimation de l'IM est le coût des calculs qui est très inférieur par rapport à ceux des autres types d'estimateurs. De plus, on sait que l'$IM_{discrète}$ est inférieure ou égale à la vraie valeur de l'IM, et qu'on converge vers cette vraie valeur en affinant de plus en plus la partition de l'intervalle des valeurs (Paninski 2003). Enfin, le plus souvent, il n'est pas nécessaire de connaître la valeur précise de l'IM, mais seulement un ordre de grandeur ou un rang, pour différents ensembles de gènes, car le but n'est pas d'estimer entièrement la distribution des probabilités jointes, mais seulement de capturer un graphe le plus proche possible du modèle d'indépendance (Nemenman 2004).

On parle toujours des estimateurs non-paramétriques de l'IM discrète, alors qu'en fait il s'agit d'estimateurs non-paramétriques de l'entropie discrète (H) ; ceci s'explique par le lien[32] existant entre les deux quantités.

Les propriétés de certains de ces estimateurs (consistance, écart quadratique moyen, comportement asymptotique, voir Beirlant *et al*. (1997)

---

31 Considérons que la concentration du transcrit du gène *i* est une variable aléatoire continue, notée *X*. Le profil d'expression du gène *i* est un échantillon de *n* valeurs identiquement distribuées par la loi de *X*. La discrétisation se fait en divisant l'intervalle des valeurs possibles de *X* en *m* plus petits intervalles (pas forcement égaux). Notons *Y* la variable aléatoire discrète obtenue par cette technique à partir de *X*. *Y* suit une loi multinomiale avec comme paramètres (n, p_1,..., p_i,..., p_m), avec p_i la probabilité d'appartenir au *i*ème intervalle.

32 $I(X,Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) = H(X) + H(Y) - H(X,Y)$, cette relation se généralise pour le cas où on a plus de deux variables aléatoires, ou encore pour le cas de l'IM conditionnelle.

pour les définitions) dans le cas de sous-échantillonnage, ont été étudiées récemment, soit par une approche analytique, soit par des simulations.

On montre qu'il existe un estimateur "précis" de l'entropie quand n/m est petit (n : nombre d'observations ; m : nombre d'intervalles utilisé pour la discrétisation) (Paninski 2004).

De nombreux estimateurs de l'entropie discrète ont été développés dans la littérature. Leurs propriétés n'ont pas toujours été testées en régime de sous-échantillonnage. Nous avons retenu huit d'entre eux décrits si-dessous, dans notre travail (H^{MV}, H^{MV3}, $H^{Dirichlet}$, H^{NSB}, H^{shrink}, H^{ZIP}, H^{ZINB}, H^{BUB}).

Les estimateurs de maximum de vraisemblance

Le plus classique estimateur de la classe des estimateurs non-paramétriques de l'entropie discrète est l'estimateur de maximum de vraisemblance, H^{MV}. Nous allons voir ses propriétés dans le régime de sous-échantillonnage et les modifications proposées dans une approche statistique bayésienne adaptée aux conditions de sous-échantillonnage, ainsi que les propriétés de cet estimateur associé à une autre technique de discrétisation.

L'estimateur de maximum de vraisemblance (H^{MV}). H^{MV}, connu aussi sous le nom de "technique de l'histogramme", utilise les fréquences comme estimation pour les probabilités de chaque intervalle de la variable discrétisée[33]. Cet estimateur est optimal pour le cas où $n >> m$. H^{MV} sous-estime l'entropie, et donc surestime l'IM (cf. la relation entre IM est H). Dans les cas de sous-échantillonnage, ($n \leq m$), l'estimateur de maximum de vraisemblance est sous-optimal, et surtout, fortement biaisé, car les petites erreurs dans l'estimation des p_i se traduisent par de grandes erreurs dans l'estimation de l'entropie. Aussi, le meilleur estimateur des p_i ne correspond pas au meilleur estimateur de l'entropie.

Un des critères permettant de comparer les estimateurs est l'erreur quadratique moyenne, EQM[34]. Certaines corrections du H^{MV} permettant de

---

33 $\hat{p}_i = \frac{y_i}{n}$, avec $y_i = \sum_{k=1}^{n} \Theta_i(x_k)$, le nombre d'observations tombées dans l'intervalle i ($\Theta_i(x_k)$, fonction de Heaviside).

34 $EQM(\hat{H}) = E((\hat{H} - H)^2)$, on peut montrer que $EQM(\hat{H}) = Var(\hat{H}) + Biais^2(\hat{H})$.

diminuer l'EQM ont été proposées (Efron 1981), mais ces corrections doivent être appliquées seulement quand le nombre d'intervalles, *m*, reste bien inférieur à celui de la dimension de l'échantillon, *n*.

H^{MV} avec une technique de comptage basée sur les fonctions "B-spline" (H^{MV3}). Daub et al. (2004) proposent un estimateur beaucoup plus simple, très proche de la "technique de l'histogramme" mais qui calcule le nombre de valeurs dans un intervalle avec des fonctions splines. En fonction de l'ordre des fonctions splines, une valeur peut appartenir à un (*ordre=1*) ou plusieurs intervalles (*ordre > 1*). En plus de la faible complexité de cet estimateur, il possède d'autres propriétés qui le rendent très intéressant dans le cadre de l'étude des réseaux de gènes. Il est beaucoup moins sensible au nombre d'intervalles choisis pour la discrétisation, ses performances sont au moins aussi bonnes que celles de H^{BUB} (défini plus loin) et mieux encore, ses performances se rapprochent des performances des estimateurs à noyaux.

Les auteurs ont analysé les performances de cet estimateur en fonction de l'ordre des fonctions splines choisi, le meilleur choix serait l'ordre *3*, car au-delà les calculs sont complexifiés pour obtenir une amélioration insignifiante, de plus ce ne serait pas toujours réaliste de dire qu'un échantillon qui normalement appartient au cinquième intervalle est tombé dans le premier à cause du bruit (mais ceci dépend bien entendu de la largeur des intervalles).

Les estimateurs Bayésiens de l'entropie

Dans une approche statistique bayésienne, des estimateurs biaisés ont été proposés pour estimer les probabilités d'appartenance à un intervalle donné, de façon à diminuer l'EQM de l'entropie. Ces estimateurs ont été conçus et étudiés pour le cas de sous-échantillonnage.

Les estimateurs basés sur l'a priori de Dirichlet (Hausser 2006). L'approche classique consiste à prendre un a priori de Dirichlet. Utiliser cet a priori revient à introduire un certain nombre constant de comptages dans chaque intervalle, (a priori non-informatif). Les estimateurs basés sur l'a priori de Dirichlet améliorent l'EQM, sauf que le choix de l'a priori domine l'estimation de l'entropie même après que les données aient été observées (Hausser 2006). Le choix du nombre de comptages ajoutés aux comptages réels par l'a priori domine l'estimation face aux observations (Nemenman *et al.,* 2002).

L'estimateur basé sur l'a priori de Nemenman, Shafee et Bialek. Nemenman et al. (2002), proposent un nouvel estimateur, qui remplace l'a priori de Dirichlet avec un autre, construit de façon à ce que la distribution

de l'entropie sur l'intervalle de valeurs possibles $[0, \log(p)]$, soit uniforme. Cet estimateur est sous-optimal lorsque la distribution de l'entropie a une forme lisse.

L'estimateur réduit, H^{shrink}. H^{shrink} consiste à estimer la moyenne pondérée entre la fréquence des comptages ($\hat{p}_i^{MV}$), et $\frac{1}{m}$ qui est la probabilité d'appartenir à un intervalle permettant de maximiser l'entropie. Cet estimateur dépend d'un paramètre λ, qui est l'intensité de "réduction". Pour des valeurs précises du λ, on peut obtenir la valeur maximale de l'entropie, l'estimateur de maximum de vraisemblance ou encore l'estimateur bayésien basé sur l'a priori de Dirichlet. Le valeur optimale de λ est calculable analytiquement (Schäfer and Strimmer 2005), ce qui réduit énormément la complexité des calculs.

Les estimateurs basés sur des lois à excès de zéro. Les nouveaux estimateurs proposés par J. Hausser, (Hausser 2006), H^{ZIP} et H^{ZINB}, utilisent l'estimateur H^{shrink}, mais en estimant les p_i à partir des comptages, par des lois à excès de zéro, adaptées aux comptages (la loi de Poisson et la Binomiale négative)[35]. La méthode consiste à estimer les deux paramètres de la loi (trois pour ZINB), de façon à maximiser la vraisemblance des données, ensuite on élimine les intervalles dont la p_i est estimée à *0* et on applique le H^{shrink} sur le reste des intervalles.

L'estimateur de Paninski, (2003), H^{BUB} ("Best Upper Bound"). C'est un autre estimateur bayésien, nous ne l'avons pas beaucoup étudié, mais il est très souvent cité dans les travaux de Hausser (Hausser 2006) et de Steuer (2002). Grâce au code fourni par J. Hausser nous avons pu le tester.

Hausser (2006) compare les estimateurs qu'il propose (H^{ZIP} et H^{ZINB}), avec H^{shrink}, H^{NSB} et le H^{BUB} proposé par Paninski, grâce à des simulations. Dans ses conclusions, Hausser affirme que H^{NSB} est sous-optimal et qu'en dépit d'un codage optimisé, il demande énormément de moyens de calculs. Les résultats des estimations avec H^{BUB} ne sont pas fournis, car ils semblent très mauvais.

---

35 La loi de X qui est une variable aléatoire discrète suivant une loi de Poisson à excès de zéro, ZIP, est : $\begin{cases} X = 0 & \text{avec la probabilité } q \\ X \sim Poisson(\lambda) & \text{avec la probabilité } 1-q \end{cases}$. Une variable aléatoire suivant une loi ZINB se définit de manière homologue.

4.2.3 Les étapes du calcul dans IGOIM

La méthode de construction des réseaux de gènes que nous avons développée permet d'inférer des graphes de premier ordre, à l'aide de l'IM conditionnelle. Nous avons appelé cette méthode IGOIM pour Inférence de Graphes de premier Ordre avec l'IM.

IGOIM prend en entrée une matrice d'expression de gènes (ME), un nombre d'intervalles (m), avec lequel on divisera l'intervalle des valeurs de chaque gène (discrétisation), un nombre de permutations (N_{perm}), pour tester les différentes quantités, une méthode de comptage, et un estimateur de l'IM. La sortie n'est autre que la matrice d'adjacence[36] du graphe inféré. Pour l'instant, nous avons choisi d'utiliser le même intervalle de valeurs pour tous les gènes, c'est l'intervalle compris entre la valeur minimale et la valeur maximale de la ME.

Les calculs de IGOIM sont divisés en six étapes (**Figure 15**).

- *Etape 1 : Discrétisation et comptages*. L'intervalle des valeurs est divisé en m intervalles égaux, dans lesquels on répartit les différentes observations du profil d'expression. Les comptages peuvent être faits par la technique de l'histogramme (on compte tout simplement les échantillons qui tombent dans chaque intervalle), ou avec une méthode de lissage, par exemple les B-splines[37].

36 La matrice d'adjacence se définit ainsi : $a_{i,j}=1$ s'il existe une arête entre le gène i et le gène j, $a_{i,j}=0$ sinon.

37 Notons k l'ordre de la fonction B-spline (k indique alors à combien d'intervalles va appartenir une même observation). Une fonction B-spline associe à chaque échantillon une probabilité d'appartenir à chacun des k intervalles.

Une fonction B-spline se définit récursivement de la manière suivante : $B_{i,k}(x) = B_{i,k-1}(x)\frac{x - t_i}{t_{i+k-1} - t_i} + B_{i+1,k}(x)\frac{t_{i+k} - x}{t_{i+k} - t_{i+1}}$, avec $B_{i,1}(x) = \begin{cases} 1, \text{ si } t_i \le x \le t_{i+1} \\ 0, \text{ sinon} \end{cases}$ et $t_i = \begin{cases} 0, \text{ si } i < k \\ i - k + 1, \text{ si } k \le i \le m - 1 \\ m - 1 - k + 2, \text{ si } i > m - 1 \end{cases}$. Le comptage pour l'intervalle i est alors $\sum_{u=1}^{n} B_{i,k}(x_u)$ (Steuer *et al.* 2002).

- Si nous choisissons l'ordre des fonctions B-spline égal à 1, la technique de comptage B-spline devient équivalente à la technique de l'histogramme. Choisir la méthode de comptage revient à choisir l'ordre de la fonction B-spline. Les comptages sont faits pour chaque gène, mais aussi pour tous les couples de gènes. A cette étape, la matrice d'adjacence correspond à un graphe complètement connecté.
- *Etape 2* : *Estimation des* p_i. L'estimateur choisi (MV ou ZIP) est appliqué aux comptages.
- *Etape 3* : *Calcul de* H_{12}. H_{12} est une matrice carrée (nb de gènes x nb de gènes) contenant l'entropie de chaque couple de gènes. L'entropie est calculée à partir des p_i estimées à l'étape précédente. La diagonale contient donc l'entropie de chaque gène.
- *Etape 4 : Calcul de* I_0. I_0 c'est la matrice contenant l'IM de chaque couple de gènes, sur la diagonale, elle contient, comme H_{12}, l'entropie de chaque gène.
- *Etape 5 : Test* I_0. Nous permutons un certain nombre de fois (N_{perm}) la ME (une permutation de la matrice consiste à permuter individuellement chaque ligne de la matrice). Pour chaque permutation de la ME nous estimons la matrice I_{0perm}. Ensuite, pour chaque terme de la matrice I_0 nous vérifions si sa significativité est inférieure à 0.2, dans le cas contraire le coefficient est mis à zéro. Cette étape supprime les arêtes correspondant à des valeurs de IM « non significatives ».
- *Etape 6 : Test* I_1. Pour chaque coefficient non nul de I_0, nous calculons $I_{i,j|l}$ et nous testons par la même technique (N_{perm} permutations) cette quantité par rapport à zéro, si nous trouvons au moins un gène l tel que $I_{i,j|l}=0$, le coefficient (i,j) de la matrice d'adjacence est mis à zéro.

Comme I_0 et H_{12} sont des matrices symétriques, nous travaillons tout le temps sur la matrice triangulaire supérieure, ce qui divise par deux le temps de calcul.

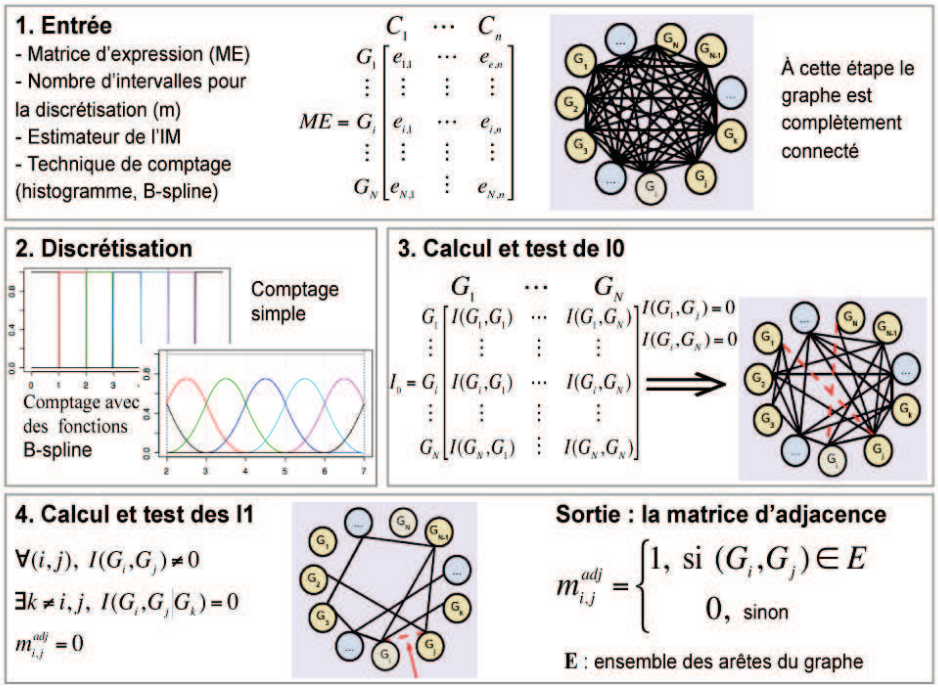


Figure 15. Etapes du calcul dans IGOIM, méthode d'inférence des graphes de premier ordre à l'aide de l'information mutuelle conditionnelle.

4.2.4 Les jeux de données d'expression utilisés pour l'inférence de réseaux

Étant donné que l'IM et le formalisme des modèles probabilistes graphiques offrent un certain espoir pour l'inférence des réseaux de gènes, nous avons décidé d'analyser une méthode d'inférence de graphes 0-1, utilisant l'IM pour le test d'indépendance et pour le test d'indépendance conditionnelle d'ordre 1. L'inférence des réseaux de gènes avec des graphes 0-1 a déjà été réalisée en utilisant la corrélation comme mesure de dépendance (Magwene and Kim 2004), mais jamais avec l'IM. De plus, parmi les estimateurs non-paramétriques de l'entropie discrète que j'ai décrit plus haut, à part celui du maximum de vraisemblance, aucun n'a été utilisé pour le moment pour l'inférence de réseaux de gènes.

Nous allons aussi comparer notre approche (associée à différents estimateurs de l'IM) avec celle des modèles gaussiens graphiques (Schäfer and Strimmer 2005; Schäfer *et al.,*) (même si ces modèles font l'hypothèse de la normalité des profils d'expression, ils ont l'avantage d'inférer des graphes complets) et avec une autre approche issue de la théorie de l'information, ARACNE (utilisant l'IM simple, avec un estimateur à noyaux ;

l'originalité de cette méthode est dans le fait qu'elle éclate les cycles de trois gènes[38]) (Margolin *et al.,* 2006).

Dans un premier temps pour le choix des estimateurs de l'IM et ensuite pour la comparaison de IGOIM avec deux autres approches d'inférence (modèles gaussiens graphiques (Schäfer and Strimmer 2005; Schäfer *et al.,*), ARACNE (Margolin *et al.,* 2006)), nous avons utilisé cinq ensembles de jeux de données :

- *dataset1* : 20 échantillons identiquement distribués d'une loi normale bivariée, générés avec la fonction *mvrnorm* de la librairie MASS de R. Chaque loi marginale est une normale de moyenne 0 et de variance 1, la corrélation entre les deux marginales étant de 0.8. Ce jeu a servi à la comparaison de la complexité du calcul des différents estimateurs. La normale bivariée a été choisie pour la simple raison que c'est la seule loi pour laquelle la relation entre les variables est entièrement connue, et pour laquelle on dispose d'une fonction de simulation.
- *dataset2* : contient des échantillons de trois tailles différentes (10, 20 et 50) provenant de différents types de lois normales bivariées. Plus précisément, nous avons pris 5x20 lois normales bivariées différentes. Chaque loi est caractérisée par le vecteur moyenne (*m*), le vecteur variance (*va*) et la corrélation entre les marginales (*co*). Ces 100 lois sont issues de différentes combinaisons des valeurs des paramètres, *m*=0, 1, 5, *va*=1, 3 et *co*=0.1, ..., 0.95.
- *dataset3* : jeu de données contenant des échantillons de différentes tailles de deux variables couplées par une relation de type "dose-réponse". Une des variables est échantillonnée uniformément et l'autre est calculée en fonction de la première avec un bruit gaussien. La différence entre ce jeu de données et celui de *dataset2* est la forme de la fonction liant les deux variables.
- *dataset4* : jeux de données de Bansal et al. (2007). Ce sont des données d'expression provenant des réseaux de 10 ou 100 gènes, de perturbations locales[39] ou globales[40]. Le modèle qui a été utili-

38 Si les gènes G_1 et G_3 n'interagissent qu'à travers le gène G_2 et si aucune autre interaction directe n'existe entre G_1 et G_3, alors IM(G_1,G_3)≤ min(IM(G_1,G_2),IM(G_2,G_3)).

39 Dans le cas des *perturbations locales*, seul un petit nombre de gènes sont perturbés, un transgène surexprimé, ou à l'inverse en inhibant spécifiquement un gène donné (technique de "knock out"). Dans ce cas la perturbation est parfaitement connue. Dans le jeu de Bansal *et al*. (, 2007), dans une perturbation locale un seul gène et perturbé.

sé pour la simulation est un ensemble d'équations différentielles linéaires couplées. A chaque profil d'expression un bruit normal a été ajouté. Le programme de simulation a été implémenté dans Matlab.

– *dataset5* : jeu de données de Mendes et al. (2003). Les données ont été générées avec le programme de simulation GEPASI. Ce jeu comprend trois ensembles de données d'expression après perturbation locale. Chaque ensemble de données correspond à un réseau de topologie différente : aléatoire, scale-free ou petit-monde. Ces réseaux à la base de la simulation sont très creux. Il y a en tout 200 interactions dans chacun de ces réseaux. Les données ne contiennent pas de bruit, mais nous l'avons introduit sous la forme d'une loi normale de moyenne nulle et d'écart type 5, 10 ou 15 % de la valeur maximale du profil d'expression. Ce jeu de données est décrit comme l'un des plus proches du modèle réel, d'abord à cause de la non-linéarité de la relation entre les gènes, mais aussi parce que la dégradation des ARN est incluse dans le modèle, et enfin parce qu'une variabilité intrinsèque est introduite pour simuler des processus stochastiques.

Les trois premiers jeux de données seront utilisés pour les choix d'estimateurs à inclure dans la méthode d'inférence des réseaux de gènes à large échelle. Les deux derniers serviront surtout pour comparer notre méthode avec les modèles gaussiens graphiques et avec ARACNE.

Les fonctions des différents estimateurs que nous avons analysés, ainsi que la méthode d'inférence que nous proposons, ont été implémentées dans R. Les calculs ont été réalisés avec R version 2.4.0.

5 Tests statistiques

Toutes les analyses et les tests statistiques utilisés dans ces travaux ont été réalisés grâce au logiciel R[41], (versions 2.6.1 à 2.10.1).

---

40 Les *perturbations globales* concernent des traitements biotiques ou abiotiques qui sont appliqués sur les cellules, dans ce cas les gènes qui seront perturbés (généralement nombreux) ne sont pas connus à l'avance. Pour le jeu de données *dataset4* dans une perturbation globale, chaque gène est perturbé avec un certain coefficient.

41 http://cran.r-project.org/

Partie III

Résultats

1 Le génome de *Buchnera aphidicola* vu comme un sous-ensemble du génome d'*E. coli*

1.1 Analyse fonctionnelle - Conservation des classes de gènes

1.1.1 L'annotation Gene Ontology

Buchnera a évolué dans un environnement particulier, ce qui a eu pour conséquence une conservation originale des gènes, liée aux nécessités du couple symbiotique et à la vie intracellulaire. Il a déjà été noté (cf. Introduction, §3.4.1), la proportion relative plus forte des gènes métaboliques et plus faible des gènes régulateurs chez *Buchnera APS* par rapport à *E. coli*.

Nous avons utilisé l'annotation GO pour détecter si d'autres biais significatifs de conservation par fonction, sont attestés dans le génome de *Buchnera*. Sachant que l'annotation GO a une structure hiérarchique et que les différents niveaux de son schéma ne sont pas aussi bien décrits, les termes GO associés à chaque gène de *Buchnera* ont été systématiquement convertis en termes GO de niveau 3, et ce sont ces annotations que nous avons utilisées pour nos analyses, afin d'éviter des biais d'annotation. La **Figure 16** montre que le génome de *Buchnera* contient des proportions significativement supérieures pour de nombreuses classes par rapport à *E. coli*. Les proportions de gènes de la classe des transporteurs, de la régulation traductionnelle, de la réponse à un stimulus, de la localisation et de la régulation biologique sont significativement inférieures chez *Buchnera APS* par rapport à *E. coli*. Parmi les classes dans lesquelles des gènes d'*E. coli* ont été annotés mais ne comportant aucun gène de *Buchnera APS* (classes absentes de la **Figure 16**) il est important de mentionner : la signalisation (131 gènes chez *E. coli*), l'activité de transduction moléculaire (101 gènes chez *E. coli*), les processus multi-organismes (56 gènes chez *E. coli*), l'activité de régulation traductionnelle (sept gènes chez *E. coli*), l'adhésion

biologique (34 gènes chez *E. coli*), la mort cellulaire (20 gènes chez *E. coli*) et la reproduction (25 gènes chez *E. coli*).

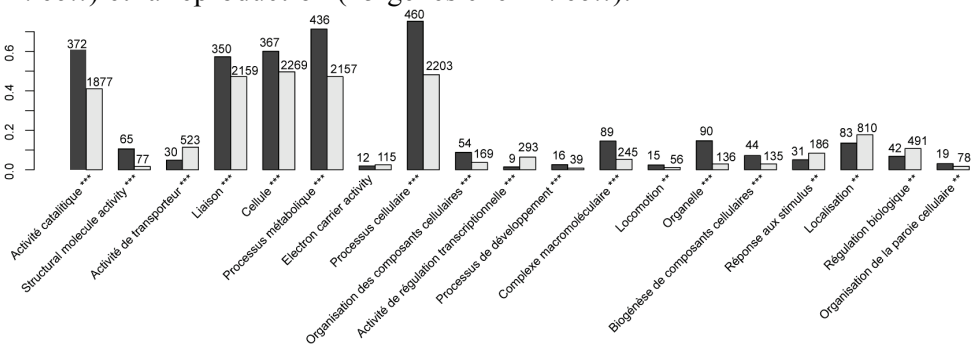


Figure 16. Comparaison des proportions relatives des génomes (nombre de gènes dans le génome/nombre total de gènes) annotés à des classes d'ontologie de niveau 3, chez *Buchnera APS* (noir) *vs. E. coli* (blanc). Les classes pour lesquelles le test de différence des proportions est significatif sont marquées * si p-valeur <0.01, ** si p-valeur<0.001 et *** si la p-valeur<0.0001 (R *prop.test*, librairie stats).

Etant donné la faible taille de son génome, nous avons supposé que pour assurer les différentes fonctions *Buchnera* a dû conserver plutôt les gènes des protéines multifonctionnelles. Pour vérifier cette hypothèse, nous avons alors construit les distributions des gènes de *Buchnera* et d'*E. coli* selon le nombre de termes GO qui leur a été associé. On peut en effet constater que dans le génome de *Buchnera* les gènes multifonctionnels ont été plus conservés (**Figure 17**, A) que chez *E. coli*. Par ailleurs, il faut noter que chez *Buchnera* il y a moins de gènes sans terme GO annoté par rapport à *E. coli*. Pour comprendre cette multifonctionnalité, nous avons analysé les termes GO de niveau trois qui co-apparaissent le plus souvent chez *Buchnera*. Nous avons ainsi réalisé que ce biais vers la multifonctionnalité chez *Buchnera* vient des gènes impliqués dans le métabolisme, ou plus précisément des gènes associés au GO:0008152 (correspondant au terme « processus métabolique »). Ce terme est souvent associé avec « processus cellulaire », « cellule » ou « activité catalytique ». Les différences entre *Buchnera* et *E. coli* disparaissent lorsque nous construisons les distributions en éliminant les gènes associés au terme « processus métabolique » (**Figure 17**, B). Cette analyse a été réalisée en utilisant tous les termes GO de niveau 3, toutes ontologies confondues (processus biologique - BP, composant cellulaire - CC et fonction moléculaire - MF). Nous avons aussi fait les analyses pour chacune des ontologies et la conclusion reste la même : *Buchnera APS* a plus conservé les gènes multifonctionnels, qui codent dans la plupart des cas pour des protéines impliquées dans le métabolisme.

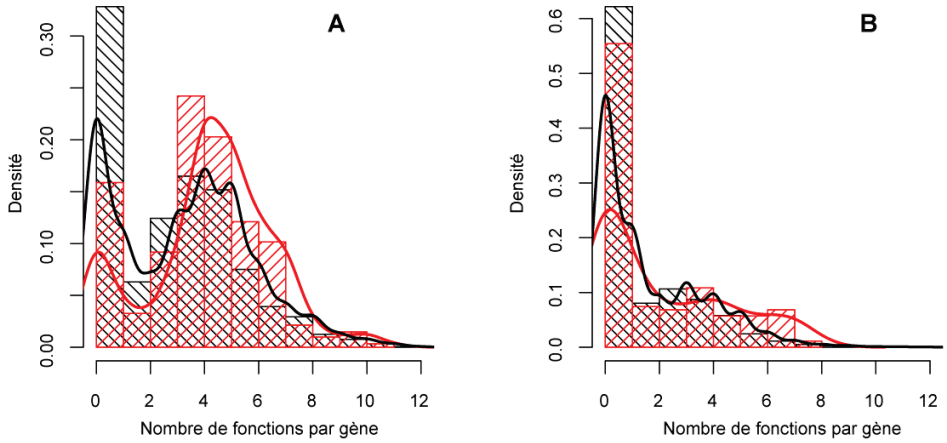


Figure 17. Distribution des gènes de *Buchnera* (rouge) et *E. coli* (noir) en fonction du nombre de termes GO. (A) Annotation au niveau 3, utilisant la totalité des génomes ; (B) Annotation au niveau 3 en éliminant les gènes associés au « processus métabolique ».

Nous avons aussi voulu étudier la localisation des gènes sur le chromosome, chez *E. coli* et chez *Buchnera APS* en relation avec leur fonction. Pour cela nous avons élaboré divers types de distances (normalisées par les tailles des génomes) et divers types de tests. Néanmoins, étant donné la différence d'échelle de la taille du génome, mais aussi la différence des nombres de fonctions distinctes pouvant être assurées pour les deux organismes, les distances entre les fonctions des gènes (mesurées en nombre de gènes), par fonction, nous ont semblé pas comparables entre les deux espèces. Pour cette raison nous ne présenterons pas ici les résultats de ces analyses, pour lesquelles nous n'avons d'ailleurs pas trouvé d'interprétations biologiques. Par contre, nous avons regardé si le test des groupements multiples et du nombre de suites sont significatifs chez *Buchnera APS* et chez *E. coli* pour chaque terme GO de niveau 3. Ces analyses n'ont pas permis de mettre en évidence des différences entre les deux bactéries.

1.1.2 Le métabolisme et la régulation

Le métabolisme a été sous l'emprise d'une sélection forte dans l'évolution du génome de *Buchnera*. Ainsi, nous avons analysé plus précisément les gènes métaboliques de cette dernière, et notamment leur régulation relativement à *E. coli*.

Dans les études se penchant sur le métabolisme d'*E. coli*, trois grandes classes de gènes sont généralement considérées : les gènes cataboliques, les gènes anaboliques et les gènes du métabolisme central et de l'énergie (Seshasayee *et al.,* 2009). Comme on peut le constater à partir du **Tableau 5**, les gènes dont les produits sont impliqués dans l'anabolisme sont les moins régulés chez *E. coli* : 40% seulement contre approximativement 80% des gènes régulés pour le catabolisme et le métabolisme central et énergétique. Chez *Buchnera APS*, ont surtout été conservés les gènes de l'anabolisme (38%) et du métabolisme central et énergétique (31%) alors que les gènes du catabolisme ont été largement perdus (9%). Étonnamment, l'existence d'une régulation des gènes métaboliques chez *E. coli* ne joue pas de rôle sur la probabilité de sa conservation chez *Buchnera APS*, la partition des orthologues régulés chez *E. coli* dans chacune des trois classes étant la même chez *E. coli* et *Buchnera APS*, *e.g.* dans le cas de l'anabolisme, chez *Buchnera APS* 128 gènes ont été conservés, parmi lesquels 52 (40%) sont des orthologues de gènes régulés chez *E. coli*. De même, parmi les 16 gènes du catabolisme chez *Buchnera APS*, 14 étaient régulés chez *E. coli* (87%). Dans ces deux classes 40 et 83% des gènes sont régulés chez *E. coli*.

Tableau 5. Regroupement des gènes métaboliques chez *E. coli* et de leurs orthologues chez *Buchnera* dans trois classes fonctionnelles « Anabolisme », « Catabolisme » et « Métabolisme central et énergétique ».

	Anabolisme	Catabolisme	Métabolisme central et énergétique
E. coli-Total	339	186	109
Gènes régulés chez *E. coli*, la proportion est calculée par rapport aux gènes d'*E. coli*	136 (40%)	155 (83%)	85 (78%)
Orthologues de BAp des gènes métaboliques d'*E. coli* (la proportion est calculée par rapport aux gènes d'*E. coli*)	128 (38%)	16 (9%)	34 (31%)
Orthologues de BAp des gènes métaboliques régulés chez *E. coli*, (la proportion est calculée par rapport aux gènes de BAp)	52 (40%)	14 (87%)	26 (76%)

Il semble alors que la probabilité de conservation d'un gène est influencée par deux facteurs. Premièrement par le type de métabolisme dans lequel le produit codé par le gène est impliqué, ce qui reflète les fonctions qui doivent être assurées pour permettre à *Buchnera APS* à la fois de sur-

vivre dans son environnement et de satisfaire aux contraintes de l'association symbiotique. Deuxièmement, par la régulation du gène, ce facteur est vraisemblablement un indicateur de la « centralité » ou l'importance du produit codé par le gène, au sein du métabolisme considéré.

1.2 L'agencement des gènes sur le chromosome

Buchnera APS a plus conservé les gènes impliqués dans le métabolisme et en a perdu d'autres qui n'étaient vraisemblablement plus nécessaires dans son nouvel environnement. Par ailleurs, les génomes de *Buchnera APS* et *d'E. coli* ont été fortement réarrangés dans leurs lignées respectives il y a quelques 200 M d'années (cf. Introduction, §3.3). Nous avons alors cherché à comprendre si l'implication des gènes dans les différents réseaux d'interactions biologiques (métabolique, protéique et de régulation) fait partie des forces de sélection qui ont structuré le génome de *Buchnera APS*. En d'autres termes, nous avons cherché à relier des aspects fonctionnels avec la préservation de certaines régions chromosomiques dans les deux lignées (les synthons).

Grâce au programme C3P de Boyer et al. (2005), nous avons calculé les synthons (ensembles de plusieurs gènes adjacents sur le chromosome d'*E. coli* et dont les orthologues chez *Buchnera APS* sont placés dans le même ordre, sans considérer l'orientation des gènes), les interactons[42], les métabolons[43] et les transcriptons[44] d'*E. coli* (cf. Matériels et Méthodes, §3.6). S'il existe une force de sélection indirecte sur le positionnement relatif des gènes sur le chromosome (possibilité de co-régulations et sensibilité moindre aux réarrangements) alors les interactons, les métabolons et les transcriptons *d'E. coli* devraient être conservés de façon intègre chez *Buchnera APS* (*i.e.* leurs gènes et leurs adjacences sont conservés). De plus, nous avons cherché si au cours de l'évolution réductrice de *Buchnera APS* de nouveaux interactons, métabolons ou transcriptons ont été créés.

42 *Interacton* - un ensemble de plusieurs gènes adjacents sur le chromosome qui sont connectés dans le sous-réseau d'interaction protéine-protéine, constitué par les protéines de l'ensemble.

43 *Métabolon* - un ensemble de plusieurs gènes adjacents sur le chromosome et qui sont connectés dans le sous-réseau métabolique, constitué par les produits de l'ensemble.

44 *Transcripton* – ensemble de plusieurs gènes adjacents sur le chromosome et connectés dans les sous-réseau de régulation de la transcription, induit par les gènes de l'ensemble.

Trente-huit interactons ont été déterminés chez *E. coli*, dont 10 conservés chez *Buchnera APS* et leurs gènes se trouvent dans un unique synthon pour neuf des 10 interactons. Par conséquent les réarrangements du génome de *Buchnera APS* n'ont pas entraîné la désintégration des interactons dont les gènes ont été conservés.

On trouve chez *E. coli* 239 transcriptons. Parmi ces 239 transcriptons seuls 38 ont plus d'un gène avec un orthologue chez *Buchnera APS*. Chez *Buchnera APS*, les orthologues de 37 (97%) de ces transcriptons sont adjacents et sont contenus dans un seul synthon. Le transcripton restant a été désintégré par les réarrangements génomiques. Un seul gène orthologue est retrouvé chez *Buchnera APS* pour 27 autres transcriptons d'*E. coli* ; l'orthologue de 14 (52%) de ces 27 transcriptons se situent dans un synthon. Pour savoir si des nouveaux transcriptons se sont formés suite aux réarrangements génomiques nous avons cherché les transcriptons de *Buchnera APS* en utilisant le réseau de régulation de *E. coli*. Ainsi nous trouvons 31 transcriptons chez *Buchnera APS*, dont 12 (39%) se sont formés grâce aux réarrangements. La façon dont la recherche des transcriptons a été faite a été orientée de façon à détecter non pas seulement les gènes strictement connectés dans le réseau de régulation d'*E. coli*, mais aussi les ensembles de cibles de régulation d'un même facteur et colocalisées sur le chromosome. Comme nous savons que chez *Buchnera APS* seuls quelques facteurs généraux de régulation ont été conservés, les transcriptons que nous avons retrouvés sont des ensembles de gènes corégulés. Nous pouvons conclure que les réarrangements génomiques ont conduit aux regroupements génomiques des cibles de ces mêmes régulateurs dans 39% de cas.

E. coli possède 107 métabolons. A 23 de ces métabolons correspondent plus d'un orthologue chez *Buchnera APS* et les orthologues de chacun des ces métabolons sont situés dans un unique synthon. On peut donc les considérer comme des métabolons de *Buchnera APS*. Nous avons aussi cherché à savoir si des nouveaux métabolons se sont formés chez *Buchnera APS*, suite aux réarrangements. Nous n'avons trouvé aucun nouveau métabolon qui serait apparu suite aux réarrangements génomiques.

Ce travail fournit donc une preuve que l'implication des gènes dans les complexes protéiques ou dans les composantes connexes du réseau métabolique sont des pressions de conservation du positionnement relatif des gènes. Ce sont donc des paramètres importants pour la transcription et l'expression de ces gènes. D'autre part, la régulation semble au moins toute aussi importante, car comme pour les interactons et les métabolons, les transcriptons pour lesquels plus d'un gène a été conservé chez *Buchnera APS* n'ont pas été désintégrés, mieux encore, les réarrangements ont conduit à la formation de nouveaux groupements de cibles de régulation.

2 La machinerie transcriptionnelle de *Buchnera aphidicola*

La régulation de la transcription des gènes chez les bactéries se fait de façon directe ou indirecte, à travers les divers acteurs moléculaires interagissant avec l'ADN et les ARN. Il existe deux type d'acteurs : les protéines et les ARN.

Les acteurs protéiques peuvent être à leur tour regroupés dans plusieurs classes : les facteurs de transcription, les facteurs de structuration et/ou de compaction du chromosome et les facteurs de réplication. Nous

nous sommes intéressés aux deux premières classes de facteurs chez *Buchnera APS*. Bien sûr, il existe des protéines qui jouent les rôles des différents types de facteurs (*e.g.* FIS est à la fois un facteur de transcription et un facteur de structuration du chromosome). L'analyse des facteurs protéiques a consisté à dresser un inventaire de ces éléments de la régulation et à analyser leur fonctionnalité.

Les acteurs nucléiques dont l'implication importante dans la régulation de la transcription et l'expression des gènes est connue depuis peu, sont les petits ARN. Ils interviennent notamment dans la stabilité des ARNm. L'analyse des petits ARN de *Buchnera* n'a pas été abordée dans le cadre de cette thèse, mais elle est actuellement étudiée dans le cadre de notre équipe.

Il existe une troisième classe de « facteurs » de régulation qui n'est constituée ni d'éléments protéiques, ni nucléiques. Il s'agit de la structure physique et thermodynamique du chromosome. De sa capacité à se compacter, s'enrouler ou s'ouvrir dépendra l'effet des divers autres acteurs de la régulation de la transcription. Cette structure physique est liée d'une part à la composition en bases du chromosome et d'autre part à la présence de certains facteurs de stabilisation comme les toporégulateurs (cf. Introduction, §4.3).

Enfin, l'organisation des gènes sur le chromosome ainsi que leur structuration constituent une autre partie importante de cette thèse.

2.1 Les acteurs protéiques de la transcription et de la régulation transcriptionnelle chez *Buchnera aphidicola*

L'inventaire de l'appareil de régulation de *Buchnera APS* (**Tableau 6**) a été réalisé principalement par orthologie avec *E. coli* (en utilisant l'annotation d'*E. coli*), mais aussi par la recherche systématique de domaines hélice-tour-hélice, le domaine protéique le plus fréquent dans les domaines de liaison à l'ADN des facteurs de transcription bactériens (Aravind *et al.,* 2005) et par la recherche systématique des protéines de BAp étant associées aux familles « régulation » et « régulation transcriptionnelle bactérienne » de la base de données Pfam.

En cherchant les orthologues des facteurs de transcription d'*E. coli* inventoriés dans RegulonDB, nous avons trouvé les gènes suivants : *dnaA*, *hupA*, *himA*, *hns*, *himD*, *pepA*, *fis*, *alaS*, *bolA*, *ychA*, *yrbA*, *metR*, *ybaB* et deux gènes codant pour deux facteurs σ, *rpoD* (σ^{70}) et *rpoH* (σ^{32}).

La recherche systématique du motif structural hélice-tour-hélice nous a permis de détecter des tels domaines dans 10 protéines de *Buchne-*

ra APS. Néanmoins, seuls les motifs de RpoD et de FIS ont été détectés avec un score de confiance de 100 %, les scores des autres protéines étant compris entre 25 et 50 %.

Chez *Buchnera APS*, les orthologues des protéines *d'E. coli* attribuées à la classe « r » comme régulation par M. Riley (Riley 1998), sont codées par les gènes *hupA*, *himA*, *dksA*, *hns*, *himD*, *fis*, *csrA*, *bolA*, *cspE*.

La recherche des protéines de *Buchnera APS* appartenant à des familles de régulation transcriptionnelle bactérienne (Pfam) nous a permis de rajouter à cette liste les gènes *cspC*, *cspE* et *csrA*.

Ces différents gènes et protéines peuvent être classés en fonction de leur spécificité de régulation. Le résumé de l'inventaire et de leur classification se trouve dans le **Tableau 6**.

Tableau 6. Inventaire des acteurs de la régulation de la transcription chez *Buchnera APS*. Les alignements ont été faits avec Geneious (Drummond *et al.,* 2010), alignement global, Blosum62. La similarité indique la proportion des acides aminés de l'alignement identique ou ayant des propriétés physico-chimiques similaires.

Gène	Type de régulateur	Identité *E. coli*/BAp (%)	Similarité *E. coli*/BAp (%)	BAp	BSg	BBp	BCc
rpoD	Facteurs σ	82.11	89.92	+	+	+	+
rpoH		72.18	89.08	+	+	+	+
alaS	Spécifiques	50.11	71.18	+	+	+	+
bolA		27.62	57.14	+	+	+	+
pepA		52.09	72.37	+	+	+	-
metR		-	-	pseudo	+	-	-
dksA	Bifonctionnels	70.89	88.08	+	+	+	-
cspC		91.30	97.10	+	-	-	-
cspE		94.20	97.10	+	+	+	+
csrA		86.89	91.80	+	+	+	+
ychA	Hypothétiques	57.62	74.35	+	+	+	+
yrbA		41.67	71.43	+	+	+	+
dnaA	Toporégulateurs	69.16	83.50	+	+	+	+
fis		65.31	84.69	+	+	-	+
hns		60.58	78.83	+	-	-	-
hupA		66.30	83.70	+	+	+	-
himA		59.80	75.49	+	+	-	+
himD		68.08	79.79	+	+	-	+
ybaB		81.65	91.74	+	+	+	+
topA		55.43	73.33	+	+	-	-
gyrA		63.58	79.57	+	+	+	+
gyrB		62.11	79.13	+	+	+	+

2.1.1 Les facteurs σ

Les deux génomes reconstruits (Moran and Mira 2001; Silva *et al.,* 2001) du dernier ancêtre commun à *Buchnera* et à *E. coli*, contiennent six facteurs σ sur les sept présents actuellement chez *E. coli* (cf. Introduction, §4.1.1), seul manque le facteur σ^{19} impliqué dans le transport de fer. Ainsi, quatre facteurs σ semblent avoir été perdus dans la lignée de *Buchnera* lors du passage au mode de vie intracellulaire. Ces facteurs ont été vraisemblablement perdus très précocement dans la lignée de *Buchnera*, puisque les mêmes deux facteurs ont été retrouvés chez les quatre espèces de *Buchnera* séquencées et particulièrement dans l'espèce que nous avons étudiée : σ^{70}, le facteur σ constitutif des gènes bactériens et σ^{32}, le facteur σ du choc thermique. La conservation chez *Buchnera* des gènes composant les régulons de ces deux facteurs σ ainsi que celle des autres facteurs chez *E. coli* est indiquée dans le **Tableau 7**.

Tableau 7. Conservation des régulons σ d'*E. coli* chez *Buchnera APS*. La dernière ligne du tableau indique le nombre de gènes dans le régulon orthologue de *Buchnera APS*, pour lesquels un promoteur σ^{70} a été trouvé.

	E. coli σ^{24}	*E. coli* σ^{28}	*E. coli* σ^{32}	*E. coli* σ^{38}	*E. coli* σ^{54}	*E. coli* σ^{19}
	119	54	156	144	110	0
# de gènes conservés chez BAp (% du régulons conservé)	20 (17 %)	15 (28 %)	43 (28 %)	13 (9 %)	1 (1 %)	0 (0 %)
# de gènes de BAp avec une promoteur σ^{70} prédit	20	15	40	13	1	-

2.1.2 Les facteurs de transcription

2.1.2.1 Les régulateurs spécifiques

Ils sont au nombre de quatre listés ci-dessous :

AlaRS réprime la transcription de son propre gène *alaS*, en s'associant à une région de l'ADN chevauchant le site d'initiation de la transcription du gène. La régulation d'*alaS* se fait exclusivement de cette manière chez *E. coli*.

BolA ne montre que deux interactions dans le réseau d'*E. coli*. Néanmoins, BolA est décrit comme un important facteur dans la réponse au stress (thermique, osmotique, oxydatif et nutritionnel) et surtout durant la phase stationnaire (Santos *et al.,* 1999). BolA contient un domaine hélice-tour-hélice putatif (Aldea *et al.,* 1989) qui expliquerait sa capacité à se lier à l'ADN et à réguler la transcription des gènes responsables de la bonne

morphologie de la cellule sous conditions de stress (Santos *et al.,* 2002; Freire *et al.,* 2009). C'est pour cette raison que *bolA* est parfois appelé morphogène. Ce motif n'a pas été déterminé avec un score significatif par l'algorithme HTH (Dodd and Egan 1990) ni chez *E. coli*, ni chez *Buchnera APS*.

PepA est l'aminopeptidase A, une protéine multifonctionnelle, agissant en tant qu'enzyme mais aussi en tant que facteur de transcription de l'opéron *carAB*. Son domaine de liaison à l'ADN n'ayant pas de structure qui rassemble aux motifs structurels connus comme caractérisant les domaines de liaison (Charlier *et al.,* 2000).

MetR n'es plus codant chez *Buchnera APS* (pseudogène). Cette pseudogénisation est apparue récemment dans l'évolution puisque le gène *metR* code une protéine fonctionnelle chez *Buchnera Sg* par exemple. La comparaison de la conservation du gène *metR* dans les différentes lignées de *Buchnera* sera abordée dans la discussion générale.

2.1.2.2 Les régulateurs bifonctionnels

CspC, CspE, CsrA et DksA sont des facteurs régulant la transcription, mais pas de façon « classique », comme les facteurs de transcription. CspC est connue pour son activité d'antiterminaison de la transcription Rho indépendante chez *E. coli* (Bae *et al.,* 2000). Les protéines CspC et CspE sont connues toutes les deux pour leur capacité à se lier à l'ARN et à l'ADN simple brin. De par leur capacité à s'associer à l'ADN simple brin, elles interviennent dans la compaction du chromosome et affectent indirectement la transcription des gènes (Phadtare and Inouye 1999). CsrA est connue principalement pour son rôle de répresseur post-transcriptionnel de la biosynthèse du glycogène (Sabnis *et al.,* 1995; Yang *et al.,* 1996) et sa capacité à s'associer à l'ARN (Romeo 1998).

DksA est connue pour sa capacité à réguler l'élongation de la transcription en s'associant à l'ARN polymérase. La protéine DksA amplifie l'effet de l'alarmone (ppGpp) sur la transcription, mais chez *Buchnera* ppGpp ne peut pas être synthétisée car les enzymes nécessaires à sa synthèse (RelA et SpoT), ne sont pas présentes chez *Buchnera*, ni chez le puceron (sous l'hypothèse de transfert de gène de *Buchnera* vers la puceron). DksA intervient dans la déstabilisation des complexes de transcription, lors de la réplication. Elle peut aussi avoir un rôle de régulateur traductionnel (RpoS).

Vu la forte conservation de ces quatre protéines et leur présence chez toutes les *Buchnera* (sauf CspC), il semble qu'elles fassent partie de l'appareil essentiel de la transcription chez la bactérie symbiotique.

2.1.2.3 Les régulateurs hypothétiques

YchA et YrbA sont des facteurs de transcription prédits chez *E. coli* sans qu'on ait à notre disposition d'études expérimentales permettant de faire des hypothèses sur leur rôle précis. YrbA est un paralogue de BolA. Leur conservation forte chez les quatre espèces de *Buchnera* les rend intéressantes pour notre étude.

2.1.2.4 Les toporégulateurs

Les Nucleoid Associated Proteins (NAP)

Comme pour les facteurs σ et les régulateurs bifonctionnels, les séquences primaires des NAP sont fortement conservées. Les alignements de ces protéines (**Figure 18**) révèlent une conservation homogène le long des séquences de toutes les NAP sauf pour FIS. Les substitutions dans la séquence de cette dernière sont concentrées dans la moitié N-terminale de la protéine (test des rangs, p-value~10^{-5}). Cette partie de la protéine est essentielle à son implication dans les processus de recombinaison, mais pas à sa liaison à l'ADN, ni à son activité de facteur de transcription. FIS est un exemple d'évolution par domaine.

Nous avons inspecté si les sites essentiels à leur fonction de régulateur et leur capacité à se lier à l'ADN, déterminés par mutations ponctuelles chez *E. coli*, ont été conservés chez *Buchnera* (**Figure 18**). Tous ces sites sont conservés à l'identique, sauf quelques exceptions, correspondant à des substitutions par des acides aminés ayant des propriétés physico-chimiques similaires.

Tableau 8. Activité de facteurs de transcription des NAP chez *E. coli* et la conservation des cibles de régulation des NAP chez *Buchnera APS*. Les cibles sont celles décrites dans la base RegulonDB et correspondant à des ensembles très restreints. Dans la réalité, même chez *E. coli*, il y a beaucoup plus de cibles de régulation de NAP.

Facteur de transcription	*dnaA*	*fis*	*hns*	*hupA*	*ihfAB*
Nombre de cibles chez *E.coli*	10	168	116	9	210
% des cibles réprimées chez *E. coli*, conservées chez BAp	43%	16%	6%	0%	27%
% des cibles activées chez *E. coli*, conservées chez BAp	100%	21%	0%	0%	5%
% des cibles duales chez *E. coli*, conservées chez BAp	0%	75%	0%	0%	17%
% total des cibles conservées chez BAp	60%	21%	4%	0%	13%

Nous avons également mis dans cette classe des NAP la protéine YbaB. Cette petite protéine est largement conservée dans les génomes bactériens et a commencé à être décrite comme NAP depuis récemment (Cooley *et al.,* 2009). Chez *E. coli*, YbaB comme son orthologue chez *Borrelia burgdorferi*, est capable de s'associer spécifiquement à l'ADN. Néanmoins, les séquences ADN s'associant avec YbaB diffèrent entre *E. coli* et *B. burgdorferi*. Il se peut que la spécificité d'association soit donnée par la topologie de l'ADN. Sa structure ne ressemble à aucune des familles des protéines liant l'ADN (Cooley *et al.,* 2009).

Les topoisomérases

Sur les quatre topoisomérases connues chez *E. coli* seuls les gènes codant pour deux d'entre ces topoisomérases sont retrouvés (*topA* – la topoisomérase I et *gyrAB* – la gyrase), et ceci pas de façon systématique, puisque *topA* n'est pas présente chez BBp et BCc, les deux espèces *Buchnera* ayant les génomes les plus riches en bases A et T parmi les 4.

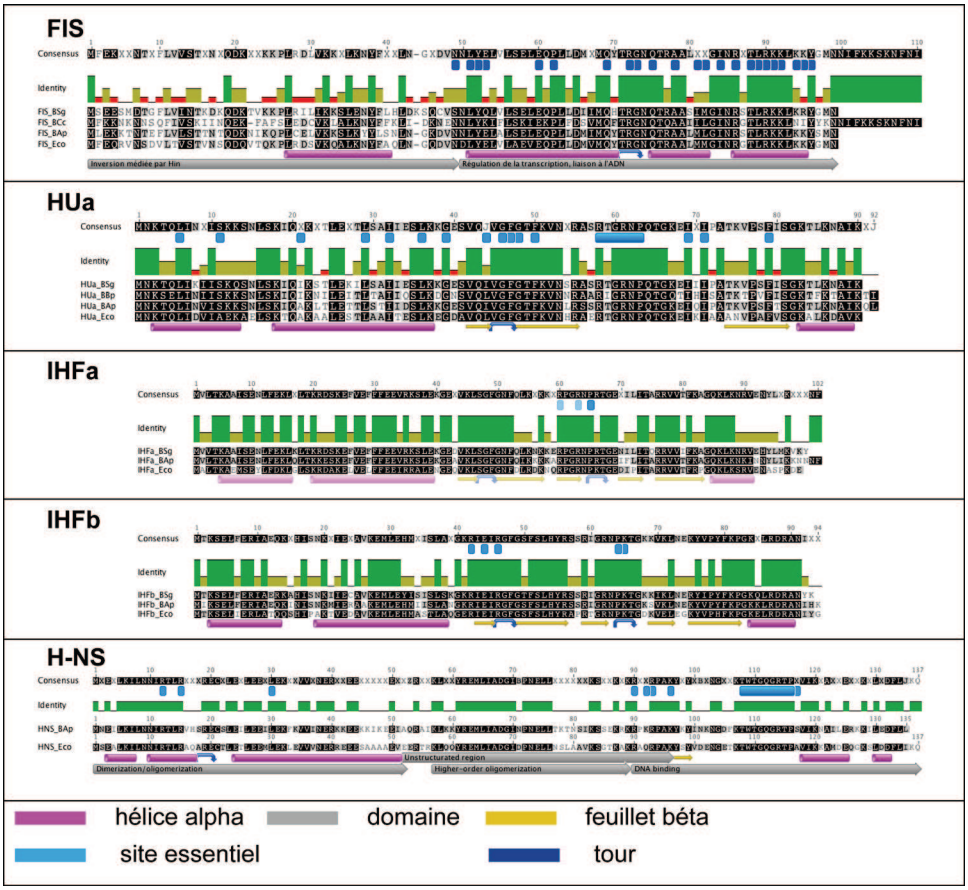


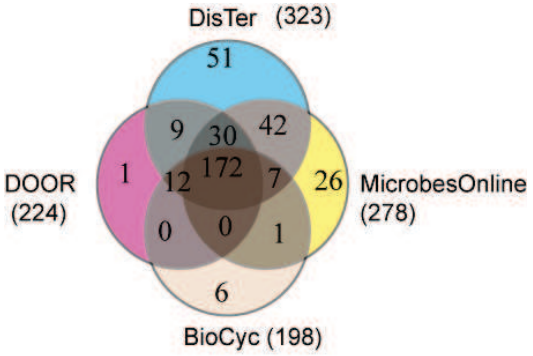
Figure 18. Alignements des séquences protéiques des NAP des 4 *Buchnera* et de *E. coli*.

2.2 Architecture génomique de la transcription chez *Buchnera aphidicola*

2.2.1 La carte opéronique, une vision d'ensemble des unités de transcription de *Buchnera*

Le génome de *Buchnera* est organisé en 443 paires de gènes adjacents, ayant la même direction de transcription et 167 paires de gènes adjacents ayant des directions de transcription opposées. Nous avons utilisé notre prédicteur d'unités de transcription (UT), DisTer (cf. Matériels et Méthodes, §4.1), sur les 443 paires de gènes unidirectionnels et cette prédic-

tion a été comparée aux autres prédictions des UT de *Buchnera APS*, présentes dans la littérature (**Figure 19**). La prédiction DisTer diffère significativement des autres prédictions. MicrobesOnline[45], la prédiction la plus proche de DisTer, ne s'accorde pas sur le statut de 99 paires parmi les 443 (**Figure 19**).



86

Figure 19. Comparaison des prédictions des statuts des paires de gènes adjacents (MUT ou UTD), chez *Buchnera APS*, par DisTer et trois autres méthodes trouvées dans la littérature : La méthode de MicrobesOnline (Price *et al.,* 2005), de BioCyc[46] (Romero and Karp 2004) et de DOOR[47] (Dam *et al.,* 2007).

Nous avons proposé une nouvelle carte opéronique de *Buchnera APS*, contenant 288 UT (**Tableau A1**, dans l'annexe), dont 155 sont des UT monocistroniques. Le nombre moyen de gènes contenu dans les UT de *Buchnera* est 2.12 (1.63 chez *E. coli*), dans les UT polycistroniques cette moyenne vaut 3.43 gènes (3.17 chez *E. coli*). Les UT de *Buchnera* contiennent en moyenne plus de gènes que les UT d'*E. coli* et la distribution des tailles des UT de *Buchnera* est significativement déplacée à droite, comparée à celle d'*E. coli* (**Figure 20**).

45 http://www.microbesonline.org/operons/

46 http://biocyc.org/

47 http://csbl1.bmb.uga.edu/OperonDB/displayNC.php?id=87

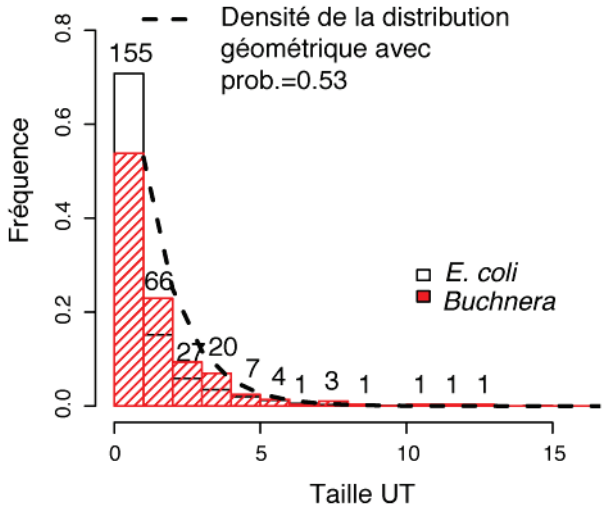


Figure 20. Comparaison de la distribution des tailles des unités de transcription (en nombre de gènes) chez *Buchnera APS* et chez *E. coli*. Les deux distributions sont significativement différentes (Wilcoxon test, p-value ~ 10^{-9}). Les nombres au-dessus des barres de l'histogramme représentent le nombre total d'UT de cette taille chez *Buchnera.* La courbe en pointillé représente l'estimation de la distribution des tailles des unités de transcription par une loi géométrique.

2.2.1.1 Validation expérimentale indirecte de la carte opéronique de Buchnera APS *avec des données transcriptomiques*

Afin de comparer les quatre cartes opéroniques prédites de *Buchnera APS* (DisTer, MicrobesOnline, DOOR et BioCyc), nous avons utilisé les données d'expression obtenues par Reymond et al. (2006) en faisant l'hypothèse suivante : la variabilité des niveaux d'expression des gènes à l'intérieur d'une UT doit être plus faible que la variabilité des niveaux d'expression entre des gènes appartenant à des UT distinctes. Cette comparaison a été réalisée une fois les UT monocistroniques exclues. Un modèle d'ANOVA à un facteur a été ajusté sur les données d'expression des gènes log-transformées en utilisant les UT comme facteur qualitatif (**Tableau 9**).

Etant donné que les quatre cartes opéroniques prédites contiennent des nombres différents d'UT polycistroniques, nous avons dû employer une version ajustée du R^2 (pénalisé par le nombre de catégories de la variable qualitative) pour leur comparaison. La prédiction DisTer montre une plus grande valeur de corrélation avec les données d'expression, néanmoins les quatre prédictions ont des valeurs ajustées de R^2 très similaires. Pour comparer ces valeurs de R^2, nous avons également analysé les p-valeurs associées. Ces p-valeurs ont été calculées par une approche non-paramétrique :

à partir de chaque carte opéronique, 10 000 autres cartes ont été simulées en mélangeant les noms des UT mais en gardant la même distribution des tailles des UT que la carte de départ. Pour chacune de ces simulations, nous avons fait une ANOVA à un facteur, comme pour la carte de départ, et conservé la valeur F du test. La p-valeur non-paramétrique est la fraction des F obtenus à partir des cartes opéroniques simulées, supérieures à la valeur de F obtenue avec la carte opéronique de départ. Bien que toutes les valeurs de F soient significatives, les valeurs obtenues avec la carte prédite par DisTer et par MicrobesOnline sont plus faibles.

Tableau 9. Analyse de variance à un facteur des données d'expression de *Buchnera*, dans le but de comparer les différentes cartes opéroniques prédites pour le chromosome de *Buchnera APS*.

	DisTer	BioCyc	MicrobesOnline	DOOR
# des paires prédites MUT	323	198	278	224
# total d'UT prédites	288	413	333	387
R^2 ajusté	0.44	0.38	0.40	0.41
p-value non-paramétrique	<1e-04	0.0143	1e-04	0.0026

2.2.1.2 Validation expérimentale directe des unités de transcription de Buchnera APS *par RT-PCR*

Nous avons utilisé l'amplification par RT-PCR afin de vérifier expérimentalement si les paires prédites comme MUT sont effectivement co-transcrites dans un même ARNm. Pour chacune des paires de gènes testée par cette méthode, des paires d'amorces ont été construites afin d'amplifier la région non-codante située entre les deux gènes et les 300 pb flanquantes de cette région. La RT-PCR utilisant ces amorces ne devrait amplifier un produit que si un ARNm polycistronique contient les deux gènes de la paire testée. Afin de vérifier cette technique sur *Buchnera*, nous avons utilisé deux témoins positifs (**Figure 21**, paires CP) et quatre témoins négatifs (**Figure 21**, paires CN). Les paires des témoins positifs font partie de l'opéron *trpABCD*, qui avait été validé préliminairement par NothernBlot par Bauman et al. (1999). Les témoins négatifs ont été choisis parmi les paires de gènes divergents (trois gènes) et convergents (un gène) de *Buchnera APS*. Les produits correspondant aux contrôles positifs ont été amplifiés par RT-PCR, alors qu'aucun des témoins négatifs ne l'a été. Ainsi cette

méthode permet de déterminer expérimentalement le statut des paires de gènes (MUT ou UTD).

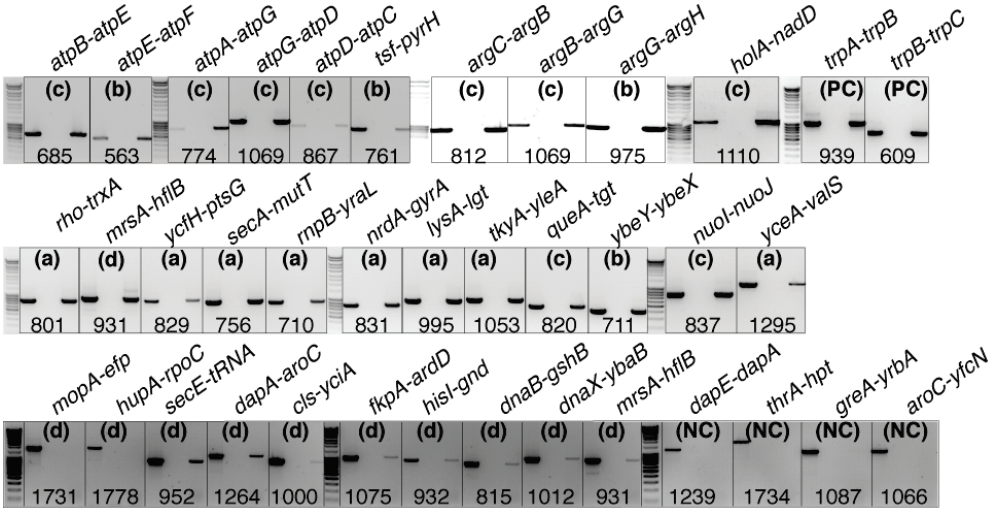


Figure 21. Vérification expérimentale du type des paires de gènes adjacents (MUT ou UTD) par amplification RT-PCR. Pour chaque paire de gènes, un contrôle positif de la RT-PCR a été fait en utilisant de l'ADNg (première colonne de chaque rectangle), un contrôle négatif de la RT-PCR, dans lequel nous n'avons pas ajouté d'enzyme (deuxième colonne de chaque rectangle). La troisième colonne de chaque rectangle correspond à l'amplification du produit à partir des ADNc. La taille de chaque amplicon est indiquée en bas de chaque rectangle et, en haut, est désigné le type de paire de gènes testée : (CP) contrôle positif ; (CN) contrôle négatif ; (a) paire de gènes prédite MUT seulement par DisTer ; (b) paire de gènes prédite MUT par DisTer et une ou deux autres méthodes ; (c) paire de gènes prédite MUT par toutes les méthodes ; (d) paire de gènes prédite UTD par DisTer.

Nous avons testé expérimentalement les prédictions que nous avons faites avec DisTer pour 31 paires de gènes, parmi lesquelles : huit paires ont été prédites MUT seulement par DisTer (**Figure 21**, les paires *a*), quatre paires ont été prédites comme MUT par DisTer et une ou deux autres méthodes de prédiction (**Figure 21**, les paires *b*), neuf paires qui ont été prédites MUT par les quatre méthodes de prédiction (**Figure 21**, les paires *c*) et enfin, 10 paires de gènes prédites UTD par DisTer (**Figure 21**, les paires *d*). Toutes les prédictions de DisTer pour les paires a, b et c ont été validées expérimentalement. Par contre, huit des 10 paires prédites comme UTD par DisTer (mais aussi par les trois autres méthodes exceptée la paire the *dnaX-ybaB*), ont été amplifiées. Il n'est pas exclu que des ADNc chimériques aient été créés et amplifiés lors de cette expérience de PCR.

L'apparition de telles structures est possible dans le cas de deux ARNm monocistroniques disjoints dont les gènes sont voisins ; l'amorçage chimérique de l'un sur l'autre s'effectuant à cause de la similarité des régions 3' et 5'UTR très AT-riches chez *Buchnera APS*. Ces ADNc chimériques devraient montrer des tailles inférieures aux tailles prédites par l'analyse génomique. Bien que nos amplicons montrent des tailles correspondant aux prédictions, la précision de l'estimation de cette taille ne nous permet pas de rejeter complètement cette hypothèse. Ce résultat suggère que même si notre méthode de prédiction classe le plus grand nombre de gènes dans des opérons, la vraie carte opéronique de *Buchnera APS* pourrait être encore plus polycistronique.

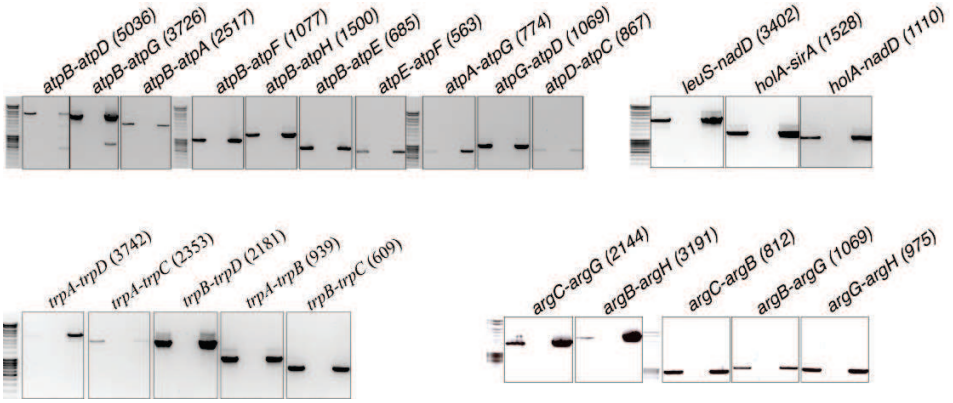


Figure 22. Validation expérimentale des opérons *atpBEFHAGDC*, *argCBGH*, *trpABC* et *leuSholAnadDsirA*, par RT-PCR. Pour chaque paire d'amorces, et donc pour chaque produit amplifié, nous avons utilisé un contrôle négatif (deuxième colonne), lors de la RT nous n'avons pas introduit la reverse transcriptase, et un contrôle positif (troisième colonne) en utilisant de l'ADNg à la place de l'ADNc. La taille des amplicons (en pb) est indiquée entre parenthèses, à coté de son nom.

Bien que nous n'ayons pas toujours réussi à amplifier les ARNm complets, l'amplification de fragments chevauchants de plus de deux gènes nous a permis de valider indirectement la présence des ARNm correspondant aux UT longues suivantes : *atpBEFHAGDC, argCBGH*, *trpABCD* et *leuSholAnadDsirA* (**Figure 22**).

2.2.1.3 Les unités de transcription et leur évolution chez Buchnera APS

Trois différentes études ont établi le répertoire de gènes du dernier ancêtre commun d'*E. coli* et de *Buchnera* (Shigenobu *et al.,* 2000; Moran and Mira 2001; Silva *et al.,* 2001). Néanmoins, cette information partielle sur l'ancêtre (*i.e.* présence/absence des gènes) ne permet pas de reconstruire sa

carte opéronique. Afin d'étudier l'évolution de la carte opéronique de *Buchnera APS*, nous l'avons comparée à la carte opéronique d'*E. coli*, car pour 95% de gènes de *Buchnera APS* un orthologue a pu être identifié dans le génome d'*E. coli*, exception faite de trois gènes orphelins et de 15 ARNt dont l'orthologie ne peut pas être définie précisément. Ainsi nous avons constaté que seuls 67% des UT de *Buchnera APS* se trouvent à l'intérieur des synthons de *Buchnera APS*/*E. coli*, les autres UT se trouvent dans des segments non synthoniques, ou à cheval sur plusieurs synthons. Les UT contenues dans des segments non-synthoniques ou à cheval, sont clairement des UT ayant suivi des évolutions distinctes dans les deux lignées de bactérie.

Tableau 10. Paires de gènes testées expérimentalement dont le produit a été amplifié par RT-PCR. Le type de la paire de gènes est en fonction des prédictions des quatre méthodes de prédiction d'opéron : ++++ pour les paires prédites opéroniques par les quatre méthodes ; +++ pour les paires prédites opéroniques par DisTer et une ou deux autres méthodes ; + pour les paires prédites opéroniques uniquement par DisTer et enfin – pour les paires prédites non-opéroniques par DisTer.

Gène1	Gène2	Type de la paire de gènes	Distance intergénique	Terminateur	BioCyc	DOOR	DisTer	MicrobesOnline
atpB	*atpE*	++++	37	+	MUT	MUT	MUT	MUT
atpE	*atpF*	+++	119	+	UTD	UTD	MUT	MUT
atpA	*atpG*	++++	33	-	MUT	MUT	MUT	MUT
atpG	*atpD*	++++	24	-	MUT	MUT	MUT	MUT
atpD	*atpC*	++++	27	-	MUT	MUT	MUT	MUT
tsf	*pyrH*	+++	51	+	UTD	MUT	MUT	MUT
argC	*argB*	++++	21	+	MUT	MUT	MUT	MUT
argB	*argG*	++++	30	+	MUT	MUT	MUT	MUT
argG	*argH*	+++	72	-	MUT	UTD	MUT	MUT
holA	*nadD*	++++	23	-	MUT	MUT	MUT	MUT
trpA	*trpB*	CP	19	-	MUT	MUT	MUT	MUT
trpB	*trpC*	CP	38	-	MUT	MUT	MUT	MUT
rho	*trxA*	+	130	+	UTD	UTD	MUT	UTD
ycfH	*ptsG*	+	97	+	UTD	UTD	MUT	UTD
secA	*mutT*	+	76	+	UTD	UTD	MUT	UTD
rnpB	*yraL*	+	80	+	UTD	UTD	MUT	UTD
nrdA	*gyrA*	+	73	+	UTD	UTD	MUT	UTD
lysA	*lgt*	+	66	-	UTD	UTD	MUT	UTD
tkyA	*yleA*	+	67	-	UTD	UTD	MUT	UTD
queA	*tgt*	++++	41	-	MUT	MUT	MUT	MUT
ybeY	*ybeX*	+++	81	+	UTD	UTD	MUT	MUT
nuoI	*nuoJ*	++++	10	+	MUT	MUT	MUT	MUT
yceA	*valS*	+	57	-	UTD	UTD	MUT	UTD

secE	*tRNA-Thr*	-	314	-	UTD	UTD	UTD	UTD
dapA	*aroC*	-	576	+	UTD	UTD	UTD	UTD
cls	*yciA*	-	381	+	UTD	UTD	UTD	UTD
fkpA	*argD*	-	463	+	UTD	UTD	UTD	UTD
hisI	*gnd*	-	359	+	UTD	UTD	UTD	UTD
dnaB	*gshB*	-	247	+	UTD	UTD	UTD	UTD
mrsA	*hflB*	-	220	+	UTD	UTD	UTD	UTD
dnaX	*ybaB*	-	320	+	UTD	UTD	UTD	MUT

Afin d'étudier les différents chemin évolutifs des UT de *Buchnera APS*, nous avons défini cinq types d'UT : identiques, similaires, fragmentées, fusionnées et réorganisées (schématisés dans la **Figure 23**). Les définitions de chacune de ces classes sont données ci-dessous. La distribution et les caractéristiques des UT de *Buchnera APS* dans ces différentes classes, sont données dans le **Tableau 11**.

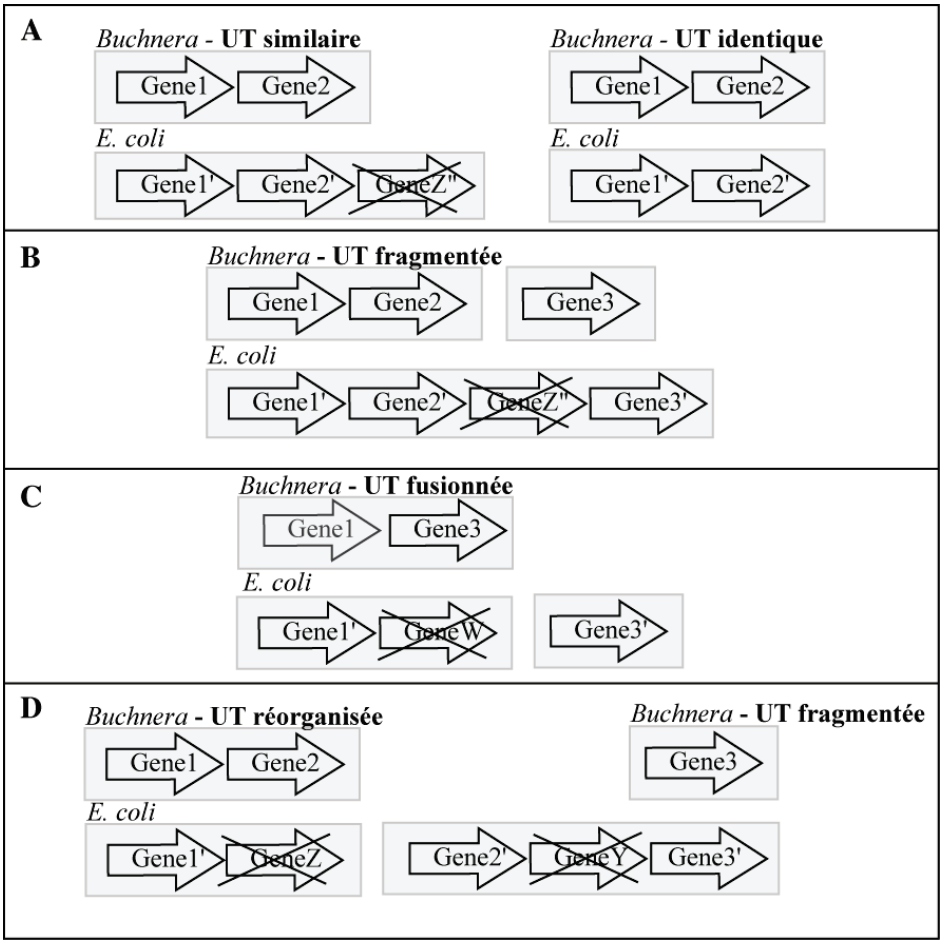


Figure 23. Schéma des classes des unités de transcription de *Buchnera APS* en fonction de leur évolution par rapport à *E. coli*.

Les UT identiques entre Buchnera APS *et* E. coli ***(Figure 23**, A)*

Nous appelons UT identique, toute UT de *Buchnera APS* composée de gènes dont les orthologues chez *E. coli* forment une UT unique, identique à celle de *Buchnera APS*. On peut dans ce cas parler d'UT orthologue. Ces UT n'ont pas été modifiées ni par les réarrangements ni par l'évolution de la séquence d'ADN dans les deux lignées. Parmi les 121 UT identiques, 99 sont des UT monocistroniques. Ces 99 UT monocistroniques sont distribuées aléatoirement sur le chromosome de *Buchnera APS* et 78% d'entre elles n'appartiennent pas à un fragment synthénique ; les 22% d'UT restantes se trouvent dans un fragment synthénique et 10 de ces UT forment 5 paires d'UT adjacentes. Les UT identiques polycistroniques contiennent principalement des gènes codant pour des enzymes ou des sous-unités d'enzymes, des protéines ribosomales ou des protéines liant l'ATP/GTP.

Les UT similaires entre Buchnera APS *et* E. coli ***(Figure 23**, A)*

Nous appelons UT similaire, toute UT de *Buchnera APS* composée de gènes dont les orthologues chez *E. coli* forment aussi une unique UT, mais qui de plus contient un ou plusieurs gènes n'ayant pas d'orthologue chez *Buchnera APS*.

Les UT d'*E. coli*, correspondant aux UT similaires de *Buchnera APS*, contiennent en plus des gènes orthologues, 80 autres gènes, n'ayant pas d'orthologue chez *Buchnera APS*, 63 de ces 80 gènes ont été perdus dans la lignée de *Buchnera APS*. En effet, il s'agit de gènes présents dans au moins un des génomes suivants : le dernier ancêtre commun de *Buchnera APS* et d'*E. coli* Silva et al. (2001), *V. cholerae*, *H. influenzae* ou *P. aeruginosa*. Les 17 autres gènes sont des gènes acquis spécifiquement dans la lignée d'*E. coli*.

Trente huit des UT similaires sont monocistroniques et 80 % (30) d'entre elles n'appartiennent pas à un fragment synthénique. Ainsi, la présence de ces UT similaires monocistroniques montre qu'en plus des multiples délétions qui ont eu lieu dans le génome de *Buchnera APS*, il y a eu aussi des réarrangements de gènes conservés provoquant des changements importants de contexte génomique.

Parmi les 16 UT similaires polycistroniques, six ont perdu leur premier gène (et donc probablement le promoteur et la régulation de l'ancêtre commun), trois ont perdu leur(s) dernier(s) gène(s), et cinq ont perdu des gènes internes. Les gènes perdus dans les deux UT similaires restantes ne sont pas adjacents dans les UT d'*E. coli*.

Les UT fragmentées chez Buchnera APS *par rapport à* E. coli ***(Figure 23,** B)*

Nous appelons UT fragmentée (chez *Buchnera APS* par rapport à *E. coli*) toute UT de *Buchnera APS* composée de gènes dont les orthologues chez *E. coli* appartiennent à une unique UT, et l'UT correspondante est formée par les orthologues de plus d'une UT chez *Buchnera APS*. Remarquablement, toutes les UT de *Buchnera APS* correspondant à la même UT d'*E. coli* sont adjacentes sur le chromosome de *Buchnera APS*. Ainsi, les UT fragmentées sont des exemples d'UT ayant évolué dans les deux lignées uniquement par évolution locale de leurs séquences, sans intervention de réarrangements. Ici encore, les gènes d'*E. coli*, sans orthologue chez *Buchnera APS*, se trouvant dans les UT d'*E. coli*, correspondant aux UT fragmentées de *Buchnera APS* ont été majoritairement perdus dans la lignée de *Buchnera APS* (17 des 23 gènes).

Les UT fusionnées chez Buchnera APS *par rapport à* E. coli ***(Figure 23,** C)*

Une UT fusionnée est une UT de *Buchnera APS* dont les orthologues chez *E. coli* appartiennent à plusieurs UT. De plus, tous les gènes d'une UT d'*E. coli* correspondant à une UT fusionnée, s'ils ont un orthologue chez *Buchnera*, doivent être situés dans une même UT. Par définition, cette classe d'UT ne peut pas contenir d'UT monocistroniques. Parmi les 64 UT fusionnées de *Buchnera APS*, 18 seulement correspondent à des UT adjacentes chez *E. coli* et représentent donc des exemples de réarrangements locaux de frontières. Sur les 99 gènes appartenant à des UT d'*E. coli* correspondant aux UT fusionnées de *Buchnera APS*, 71 ont été perdus dans la lignée de *Buchnera APS*. Ces UT ont donc été remaniées majoritairement par les réarrangements génomiques. Les UT d'*E. coli* correspondant à 33 des 64 UT fusionnées de *Buchnera APS* ne sont pas régulées de façon spécifique chez *E. coli* (*i.e.*, elles sont transcrites de façon constitutive).

Les UT réorganisées chez Buchnera APS *par rapport à* E. coli ***(Figure 23,** D)*

Nous appelons UT réorganisée, toute UT de *Buchnera APS* dont les orthologues chez *E. coli* appartiennent à plusieurs UT. La différence entre une UT fusionnée et une UT réorganisée et qu'aux UT d'*E. coli*, correspondant à une UT réorganisée, correspondent plusieurs UT de *Buchnera APS*. Vingt sept des 30 gènes présents dans les UT réorganisées chez

E. coli mais pas dans *Buchnera APS* ont été perdus dans la lignée de *Buchnera.* Six des UT réorganisées de *Buchnera APS* sont composées de paires de gènes ancestraux (les paires de gènes ancestrales ont été identifiées grâce au score OperonDB[48], les scores de ces paires étant supérieurs à 86 %). Les UT réorganisées sont donc des exemples d'UT ayant gardé une architecture ancestrale chez *Buchnera APS* et qui ont été modifiées par des réarrangements dans la lignée d'*E. coli*. Sept des 16 UT réorganisées d'*E. coli* ne sont pas régulées spécifiquement.

Tableau 11. Caractérisation des UT de *Buchnera* prédites avec DisTer.

Type UT	# UT	# gènes	# UT monocistroniques	# UT polycistroniques
Idéntiques	121	162	99	22
Similaires	54	88	38	16
Fragmentées	23	47	11	12
Fusionnées	64	231	0	64
Réorganisées	16	70	0	16

2.2.1.4 Evolution de la carte opéronique de Buchnera APS – *dynamique locale et globale*

Le but est ici d'étudier l'évolution de la carte génomique et opéronique de *Buchnera APS* comme dans la partie précédente, mais cette fois-ci en utilisant les paires de gènes adjacents comme unité d'étude et non plus les UT de *Buchnera APS*. Cette autre approche permet de comparer les deux bactéries, de façon symétrique. En regardant les paires de gènes adjacents, et plus précisément celles qui ont été conservées dans les deux lignées nous pourrions avoir une meilleure idée de la pression de conservation sur le positionnement des gènes.

Parmi les 611 paires de gènes/pseudogènes adjacents de *Buchnera APS*, 320 sont des paires ancestrales (formées par 411 gènes), ce sont des gènes présents dans *Buchnera APS* et *E. coli* (237), ou dans *Buchnera APS* et d'autres bactéries plus éloignées (83). Ces paires ancestrales sont distribuées le long du chromosome avec une tendance à former des grou-

[48] http://operondb.cbcb.umd.edu/cgi-bin/operondb/operons.cgi

pements. Nous avons compté ainsi 68 fragments ancestraux ayant en moyenne 6.5 gènes. Par conséquent le génome de *Buchnera APS* est composé d'une alternance de fragments ancestraux et de fragments réorganisés. Ces derniers sont plus courts, 2.5 gènes en moyenne. Les cartes génomique et opéronique de *Buchnera APS* sont représentées dans la **Figure 24**.

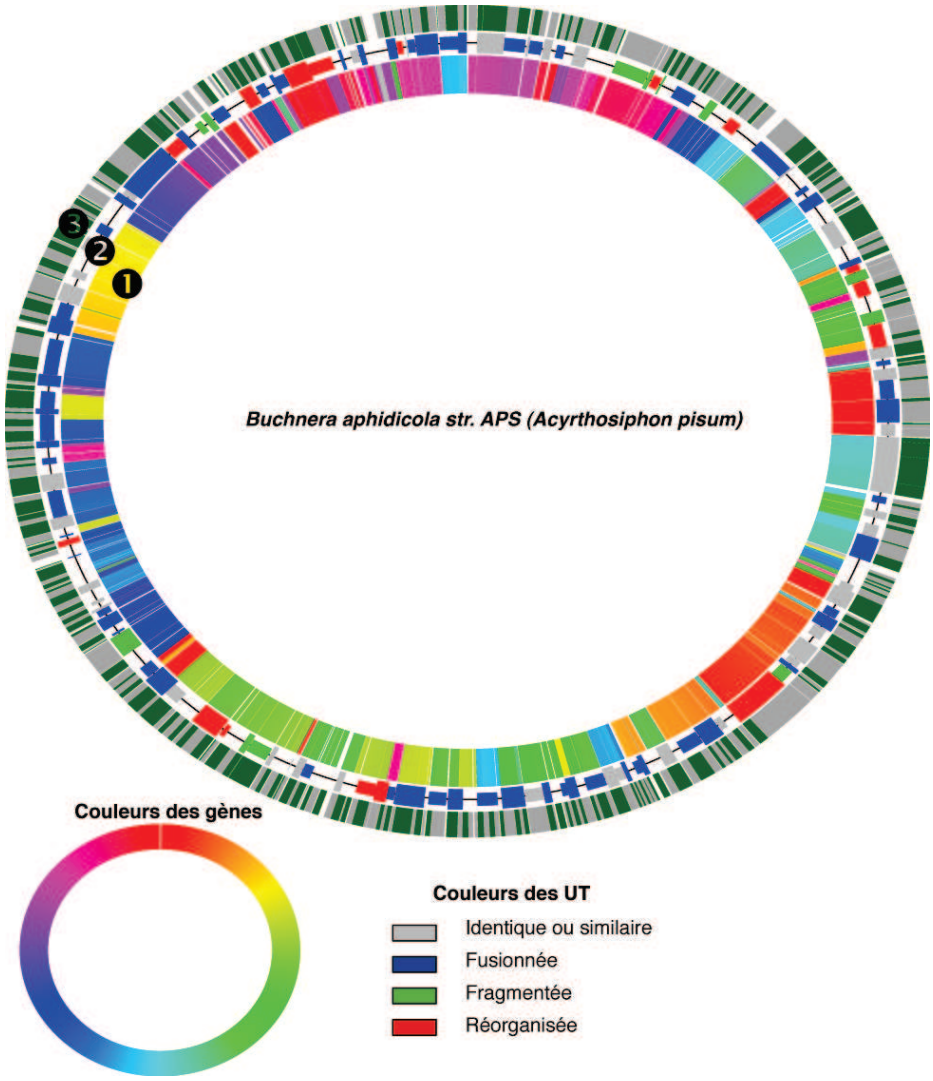


Figure 24. Carte génomique et opéronique de *Buchnera APS*. Le cercle en bas à gauche de la figure représente l'ordonnancement des orthologues des gènes de *Buchnera APS* chez *E. coli*, sur le génome d'*E. coli*, chaque gène a une couleur propre, les gènes se succédant ont des couleurs similaires et les gènes sans orthologue chez *Buchnera APS* ont été exclus. Le 1er cercle représente le génome de *Buchnera APS*, ses gènes ayant été colorés avec les mêmes couleurs que leur orthologue chez *E. coli*. Le 2ème cercle représente la carte opéronique de *Buchnera APS*, seules les UT polycistroniques sont visualisées. Le 3ème cercle représente

des blocs de gènes qui sont adjacents chez *Buchnera APS* et dont les orthologues sont adjacents sur le chromosome d'*E. coli*, de plus chez *E. coli* les gènes de chaque bloc provient d'une seule UT. Les blocks sont colorés alternativement en vert et gris de façon à pouvoir distinguer deux blocs voisins. La carte synthénique entre *Buchnera APS* (*vs. E. coli*) n'est pas représentée ici, néanmoins les couleurs des gènes visualisent partiellement les nombreux réarrangement génomiques qui ont eu lieu depuis la divergence des deux lignées.

Parmi les 237 paires de gènes présents chez *Buchnera APS* et chez *E. coli*, 188 (79.3%) sont des paires MUT ; 15 (6.3%) sont des paires UTD, et 34 (14.4%) sont des paires divergentes ou convergentes (**Tableau 12**). Par conséquent, la plupart des paires de gènes ancestraux conservés dans les deux lignées, sont des paires de gènes de type MUT. Les 188 paires sont distribuées de la façon suivante : 41 (21.8%) dans des UT identiques, 27 (14.4%) dans des UT similaires, 22 (11.7%) dans des UT fragmentées, 68 (36.2%) dans des UT fusionnées et 30 (15.9%) dans des UT réorganisées.

Tableau 12. Comparaisons des paires de gènes adjacents entre *E. coli* et *Buchnera APS*.

	Paires de gènes tandem		Paires de gènes à direction de transcription opposée	
	443		167	
	Paires MUT	Paires UTD	Paires convergentes	Paires divergentes
	323	120	84	83
Les gènes orthologues chez *E. coli* des gènes de la paire, sont adjacents et forment une paire MUT chez *E. coli*	188	13	-	-
Les gènes orthologues chez *E. coli* des gènes de la paire, sont adjacents et forment une paire UTD chez *E. coli*	21	15	3	14
Les gènes orthologues chez *E. coli* des gènes de la paire, ne sont pas adjacents mais appartiennent à la même UT	18	6	-	-

Les gènes orthologues chez *E. coli* des gènes de la paire, ne sont pas adjacents et n'appartiennent pas à la même UT	96	86	81	69

Les gènes formant les 83 paires ancestrales (paires dont les gènes ne sont pas adjacents chez *E. coli*) sont inclus chez *E. coli* dans des UT apparues suite aux réarrangements génomiques dans la lignée.

2.2.2 Les régions codantes - évolution de la taille des séquences codantes de *Buchnera APS*

La relation entre la structuration en UT et la dynamique de la taille des séquences codantes a été analysée en comparant *Buchnera APS* à *E. coli*. Afin d'orienter l'évolution, un groupe externe a été utilisé (voir Matériels et Méthodes). Nous trouvons ainsi 122 gènes de *Buchnera APS* ayant la même taille que leurs orthologues chez *E. coli*. Les 412 autres gènes ayant pu être analysés sont distribués de la façon suivante : 57 (66) ont augmenté (diminué) en taille dans la lignée d'*E. coli*, et 96 (193) ont augmenté (diminué) en taille dans la lignée de *Buchnera APS*. Par conséquent, les séquences codantes de *Buchnera* ont été majoritairement raccourcies durant l'évolution de cette lignée, par rapport aux séquences codantes d' *E. coli*, comme cela avait déjà été observé par Charles et al. (1999), sur un jeu de données beaucoup plus restreint.

Nous avons remarqué une différence marquée entre l'évolution des séquences qui ont été raccourcies chez *Buchnera APS* par rapport à *E. coli* et celles qui ont été rallongées (**Figure 25**). Les séquences codantes de *Buchnera APS* qui sont plus courtes que leur orthologue chez *E. coli* appartiennent à une des deux classes : soit la séquence a été raccourcie durant l'évolution de la lignée de *Buchnera*, soit elle a été rallongée dans la lignée d'*E. coli*. Aucune différence significative des valeurs de vitesse d'évolution mesurée par le Ka (Tamas *et al.,* 2002) n'a été retrouvée entre ces deux classes de gènes. Les mêmes observations ont été faites pour les séquences codantes de *Buchnera APS* plus longues que leurs orthologues chez *E. coli*. Enfin, aucune corrélation ne semble exister entre la structuration des gènes en UT et la dynamique de la taille de leur séquence codante.

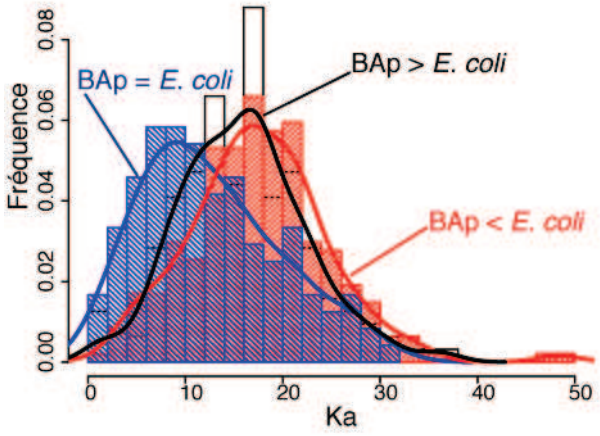


Figure 25. Distributions de la vitesse d'évolution (Ka) des gènes de *Buchnera APS* divisées en trois groupes : les gènes ayant la même taille de séquence codantes que leur orthologue chez *E. coli* (bleu) ; les gènes ayant une taille de la séquence codante plus petite que leur orthologue chez *E. coli* (rouge) ; et enfin les gènes ayant une séquence codantes plus longue que leur orthologue chez *E. coli* (noir). Les distributions rouge et noire sont significativement différentes de la distribution bleu (Kruskal-Wallis, p-valeur=$5*10^{-9}$).

2.2.3 Les régions non-codantes (intergéniques)

La distribution des distances intergéniques dans les génomes bactériens contenant des UT polycistroniques possède un pic caractéristique dans la région de -20 (superposition des séquences codantes) à +30 pb, suggérant que les opérons sont universellement compacts (font exception à cette règle quelques opérons complexes utilisant des sites d'initiation de la transcription alternative) (Moreno-Hagelsieb and Collado-Vides 2002). *Buchnera APS* ne fait pas exception à cette règle (**Figure 26**). La distribution totale des distances intergéniques de *Buchnera APS* est significativement différente de celle d'*E. coli* (test de Wilcoxon, p-value = 0.05). Comme les distributions (de *Buchnera APS* et *d'E. coli*) des distances intergéniques entre les gènes adjacents ayant des sens de transcription opposés, ne sont pas significativement différentes (test de Wilcoxon, p-value = 0.78), nous en concluons que la différence entre les distributions globales devrait venir de la différence significative existant entre les distributions des distances intergéniques entre les gènes adjacents, ayant la même direction de transcription. Notre hypothèse s'est avérée être juste (test de Wilcoxon, p-value = 8.10^{-4}). La distribution des distances intergéniques entre les gènes des

paires tandem de *Buchnera APS*, est légèrement déplacée vers la droite (**Figure 26**). Ceci est dû au fait que *Buchnera APS* contient moins de paires de séquences codantes se superposant sur le chromosome (distance intergénique négative). La différence de cette distribution entre les deux espèces est aussi due au fait que *Buchnera APS* possède plus de régions intergéniques comprises entre 20 et 100 bp, et moins de régions comprises entre 220 et 300 pb que *E. coli* (test de Chi2 avec la correction de Holm des p-values).

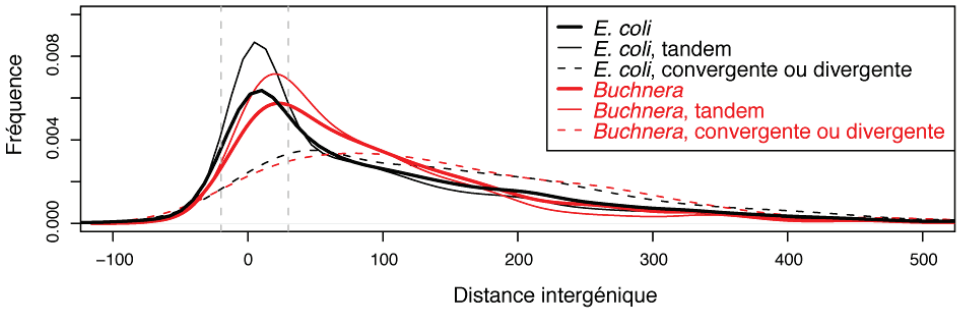


Figure 26. Distributions des distances intergéniques entre les gènes ayant la même direction de transcription (tandem), ou des directions de transcription opposées (convergents ou divergents) chez *Buchnera APS* et chez *E. coli*.

Parmi les 521 gènes formant des paires dont les séquences codantes se superposent chez *E. coli*, seulement 28 (5%) ont un orthologue chez *Buchnera APS*, et 20 de ces gènes se superposent aussi chez *Buchnera APS*. Par ailleurs, *Buchnera APS* a une proportion significativement plus petite de paires de gènes superposés : 6% contre 13% chez *E. coli*. Il semblerait donc que l'apparition de séquences superposées s'est rarement produite durant l'évolution de *Buchnera APS*. Pour regarder la dynamique d'évolution des distances intergéniques, nous avons comparé les distances intergéniques chez *E. coli* avec les tailles des régions intergéniques « orthologues » chez les *Buchnera* des pucerons *Schizaphis graminum* (BSg), de *Baizongia pistaciae* (BBp) et d'*Acyrthosiphon pisum* (BAp). Une région intergénique « orthologue » est une région intergénique entre deux gènes adjacents donc les orthologues sont aussi adjacents dans l'autre espèce (**Figure 27**). La variabilité plus grande des distances intergéniques inter-UT que celle des distances intergéniques intra-UT chez les 3 espèces de *Buchnera* et chez *E. coli* indique que certaines contraintes doivent s'exercer pour contrôler les distances entre les gènes à l'intérieur des UT (**Figure 27**, C et D).

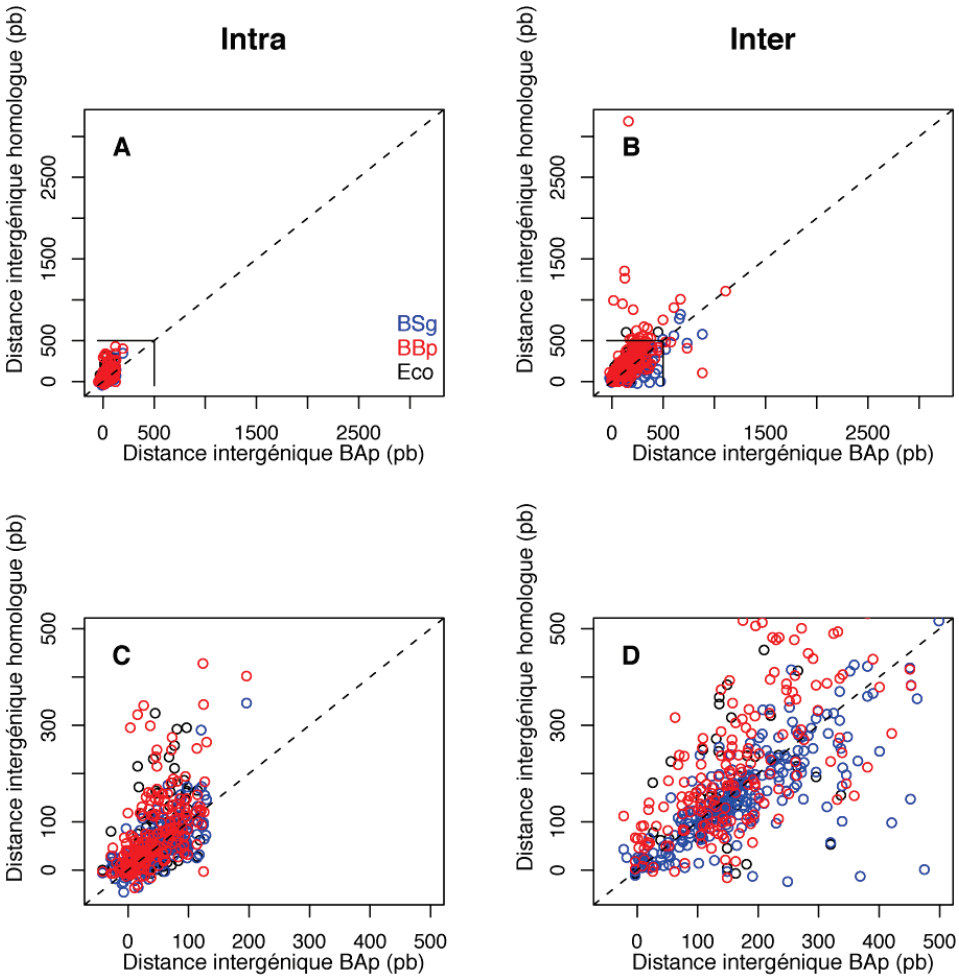


Figure 27. Distances intergéniques en paires de bases d'*E. coli*, de BSg et de BBp, en fonction de leur distance intergénique « orthologue » chez BAp. A et C : les distances intergéniques de BAP intra-UT ; B et D : les distances intergéniques de BAp inter-UT. C et D représentent l'agrandissement de la zone entourée dans A, respectivement B.

Comme nous l'avons déjà mentionné plus haut, les distances intergéniques entre les gènes ayant des directions de transcription opposées, sont similaires entre *Buchnera APS* et *E. coli*. Les distributions restent similaires entre les deux espèces, même lorsque nous séparons ce type de distance intergénique en distance intergénique de paire de gènes convergents et distances intergéniques des paires de gènes divergents (**Tableau 12**). Un autre aspect remarquable est que les distances intergéniques des paires de gènes convergents sont significativement plus courtes que les distances intergéniques des paires de gènes divergents, ceci aussi bien chez *Buchnera APS* que chez *E. coli* (**Figure 28** et **Tableau 12**).

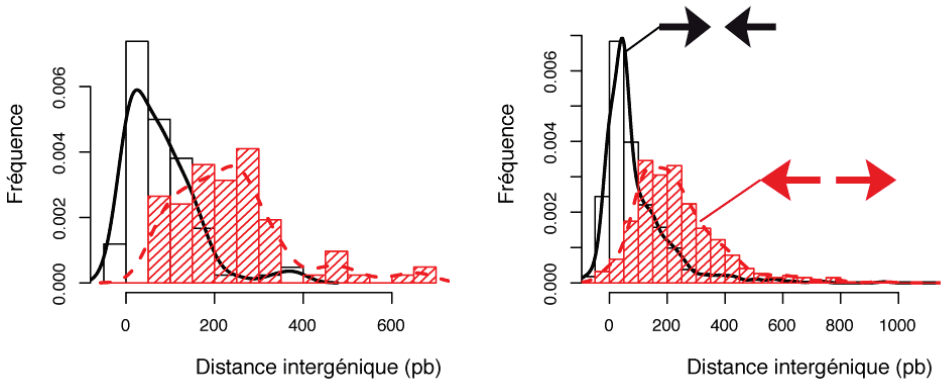


Figure 28. Comparaison entre la distribution des distances intergéniques convergentes et celle des distances divergentes, chez *Buchnera APS* (à gauche) et *E. coli* (à droite).

2.2.4 Les séquences de fixation des facteurs de transcription

Un aspect essentiel de la régulation de la transcription est constitué par les motifs d'ADN qui vont servir comme support de l'interaction entre les divers acteurs protéiques et l'ADN. En fonction de leur présence ou absence, de leur nombre et de leur séquence, ces motifs vont permettre à leur tour une régulation du niveau de transcription. Dans cette partie, nous avons fait une recherche de sites de fixation des facteurs de transcription de *Buchnera APS* et des promoteurs de ses facteurs σ. Cette analyse comporte 3 objectifs : savoir si *Buchnera APS* utilise des promoteurs constitutifs semblables aux autres bactéries ou s'il s'agit de promoteurs dégénérés, étudier le devenir chez *Buchnera APS* des gènes dont l'orthologue chez *E. coli* appartient à d'autres régulons que σ^{70} et trouver les cibles potentielles de régulation spécifique des facteurs de transcription de *Buchnera APS*.

2.2.5 Promoteurs des facteurs σ^{70} et σ^{32} de l'ARN polymérase

Chez *Buchnera APS*, on trouve 699 promoteurs σ^{70} dans les 500 pb en amont de 574 séquences codantes (*i.e.* 94% des gènes) ou en amont de 96% de ses UT. En comparaison, chez *E. coli* on trouve 4793 promoteurs σ^{70} dans les 500 pb en amont des 4062 (*i.e.* 89% de son génome) séquences codantes ou en amont de 92% de ses UT. Nous avons noté une différence de

proportion de gènes intra-UT pour lesquelles un promoteur σ^{70} a été prédit entre *Buchnera APS* (92%) et *E. coli* (84%). Les promoteurs prédits pour les gènes intra-UT ont des scores de prédictions plus faibles que ceux des gènes intra-UT, dans les deux bactéries.

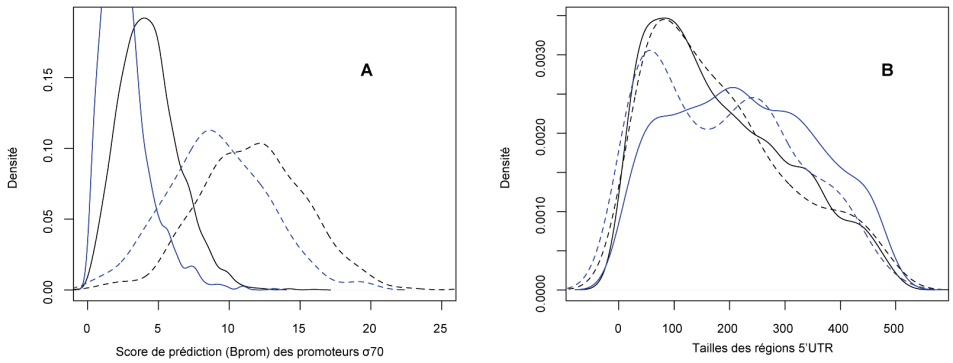


Figure 29. Comparaison des distributions des valeurs (A) des scores de prédiction (Bprom) des sites de fixation du facteur σ^{70}, en début d'UT (noir) et à l'intérieur des UT (bleu), chez *E. coli* (continue) et chez *Buchnera APS* (pointillée) et (B) des tailles des régions 5'UTR.

Les sites de fixation de σ^{70}, prédits avec Bprom, diffèrent entre *E. coli* et *Buchnera APS*, la fonction linéaire discriminante prend des valeurs significativement plus grandes pour les sites de fixation de σ^{70} de *Buchnera APS* relativement à ceux d'*E. coli* (**Figure 29**). Les sites de fixation de σ^{70} trouvés chez *Buchnera APS* sont plus semblables au modèle des promoteurs σ^{70}, utilisé pour la recherche. Un résultat remarquable reste la distinction claire des sites de fixation du facteur σ^{70} prédits en amont des gènes en début d'UT, de ceux prédits en amont des gènes à l'intérieur des UT, ceci aussi bien chez *E. coli* que chez *Buchnera APS*, alors que cette propriété n'a pas été utilisée pour la prédiction des UT.

Des sites de fixation du facteur σ^{32} ont été trouvés en amont de 29 (46%) des gènes dont l'orthologue appartient au régulon σ^{32} d'*E. coli* (**Tableau 13**). Certains des gènes dont l'orthologue appartient au régulon σ^{32} chez *E. coli*, mais pour lesquels le site de fixation de ce facteur n'a pas été trouvé, ont une expression différentielle en condition de choc thermique chez BSg (Wilcox *et al.,* 2003) (**Tableau 13**). On trouve aussi des sites de

fixation de σ^{32} devant 45%[49] des gènes dont l'orthologue appartient au régulon σ^{24} chez *E. coli*, impliqué aussi dans la réponse au stress thermique. Enfin, nous avons détecté des promoteurs σ^{32} en amont de gènes dont l'orthologue chez *E. coli* n'appartient pas au régulons σ^{32} (*ychF*, *fpr*, *rpsF*, *tmk*, *trpD*).

Tableau 13. Recherche des sites de fixation σ^{32} en amont des gènes orthologues du régulons σ^{32} *d'E. coli*. En rouge sont marqués les gènes significatifs dans l'expérience de Wilcox et al. (2003).

	RegulonDB	RegulonDB & Richmond (Richmond *et al.*, 1999)	Richmond (Richmond *et al.*, 1999)
Un promoteur σ^{32} a été trouvé dans les 500 pb en amont du codon start	*lspA, yfhC, gapA, ybeX, thiL, nusB, ppiD, rrs, lipB, pyrF, rrf, rrl*	*mopA, dnaK, grpE, topA, htpX, htpG, hslU*	*carB, carA, ycfC, glyA, flgE, pyrD, pyrI, endA, eno, fba*
Seul un promoteur σ^{70} a été trouvé dans les 500 pb en amont du codon start	*lytB, ileS, vals, hflB, ribH, yhgI, yggW, hflC, hflK, mutL, rpmE, yhiQ, glnS, ybeD*	*rpoD, dnaJ, grpE1, ftsJ, ybeY, clpP, clpX, miaA, hslV, ibpA, groES, lon*	*nuoCD, cvpA, pta, lpdA, purB, pyrC, pyrB, yheL*

Néanmoins, nous avons trouvé un promoteur σ^{32} en amont de 240 gènes de *Buchnera* APS, soit environ 40% des gènes situés sur son chromosome.

Nous avons donc constaté que *Buchnera APS* possède des promoteurs σ^{70} pour la majorité de ses gènes. Le fort biais mutationnel n'a pas provoqué un changement d'architecture des promoteurs de *Buchnera APS*, ils ont la même composition en bases (ce qu'on peut constater à travers les forts scores de prédiction Bprom) et les tailles des régions 5'UTR ont des distributions similaires entre *Buchnera* et *E. coli*. Certains gènes semblent avoir deux types de promoteurs (σ^{70} et σ^{32}). Dans ces cas, chez *E. coli* les promoteurs seront reconnus par l'un ou l'autre des facteurs σ, en fonction des conditions environnementales. Chez *Buchnera APS*, il se peut que le même mécanisme soit utilisé, néanmoins les distances entre les deux promoteurs, quand ils co-apparaissent sont généralement importantes (environ

[49] *surA*, *apaH*, *smpA*, *htrA*, *yaeT*, *dnaE*, *yfiO*, *ybaB*, *fkpA*

300 pb), et nous ne pouvons pas nous prononcer quant à la fonctionnalité d'un tel site distant.

2.2.5.1 Les sites de fixations des protéines associées au nucléoïde (NAP)

Parmi les cinq protéines ayant été décrites comme NAP, on dispose de séquence consensus de leur site de fixation pour FIS, IHF, H-NS et DnaA (Prodoric[50]). Néanmoins ces séquences sont toutes dégénérées et riches en AT. Nous avons fait une recherche de ces séquences sur le génome de *Buchnera APS*. Contrairement à *E. coli* chez laquelle le nombre d'occurrences est significativement plus grand qu'attendu par hasard (estimation faite avec une chaîne de Markov et la variance limite) chez *Buchnera APS* les sites sont significativement sous-représentés.

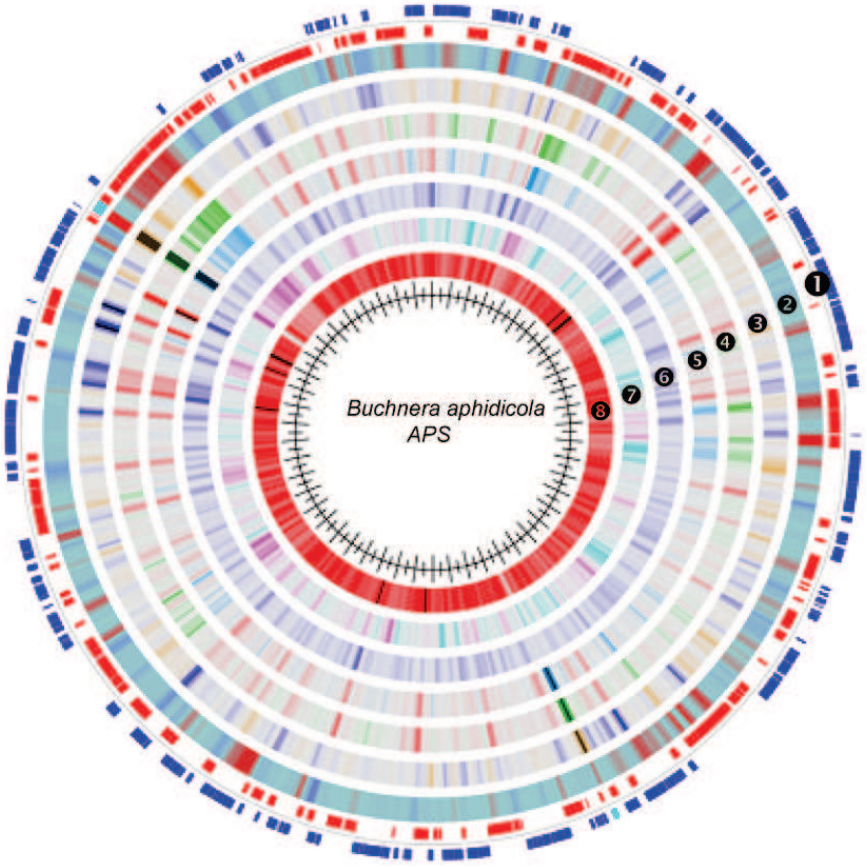
2.3 Propriétés physico-chimiques et structurelles séquence-dépendantes du génome de *Buchnera APS*

Le fait que l'organisation du chromosome joue un rôle important dans la régulation de l'expression des gènes chez les bactéries a été vérifié à plusieurs reprises (Peter *et al.,* 2004; Blot *et al.,* 2006). L'ADN n'est pas seulement un support inerte de l'information, mais aussi un élément actif, qui par ses propriétés chimiques, physiques et structurelles participe aux processus de transcription. Certaines de ces propriétés structurelles dépendent dans une grande mesure de la séquence nucléotidique (Brukner *et al.,* 1990; Bolshoy *et al.,* 1991; Hassan and Calladine 1996; Olson *et al.,* 1998; Sinden *et al.,* 1998; Pedersen *et al.,* 2000). Nous avons utilisé l'outil web GeneWiz[51] (Hallin *et al.,* 2009) pour estimer le long du chromosome de *Buchnera APS* les propriétés structurelles séquence dépendantes suivantes : la courbure intrinsèque (que nous appellerons simplement la courbure), l'énergie d'empilement, l'angle de torsion du brin d'ADN et le SIDD (leur caractérisation est donnée dans l'Introduction). Comme nous l'avons mentionné plus haut (cf. Introduction, §4.3.6), la courbure ainsi que le SIDD permettraient de pointer des régions promotrices plus favorables à la trans-

[50] http://prodoric.tu-bs.de/

[51] http://www.cbs.dtu.dk/services/gwBrowser/

cription, l'énergie de l'empilement et l'angle de torsion donnant une idée de la stabilité du brin de l'ADN et de sa flexibilité isotopique. Nous avons utilisé ces quatre propriétés structurelles séquence-dépendantes pour dresser l'atlas du génome de *Buchnera APS* (**Figure 30**), comparer ces propriétés avec celles du chromosome d'*E. coli* et enfin, pour analyser les régions promotrices de *Buchnera APS*.



1 Annotations
CDS +
CDS -
ARNr
ARNt

2 Niveau de transcription
-2,104 2,624

3 Courbure intrinsèque
0,212 0,287

4 Energie d'empilement des bases
-7,316 -6,333

5 Angle de torsion
-15,07 -13,679

6 SIDD
5 10

7 GC-skew
-0,086 0,088

8 Composition en bases A et T
0,2 0,8

Figure 30. Atlas du chromosome de *Buchnera APS* résumant les propriétés structurelles séquences dépendantes (la courbure, l'énergie d'empilement des bases, l'angle de torsion et le SIDD), les données d'expression normalisée par l'ADNg, le biais GC et la composition en bases A et T.

2.3.1 Analyse comparative globale

Nous avons comparé les distributions des quatre paramètres entre *E. coli* et *Buchnera APS*, en utilisant deux modèles de référence pour le chromosome de *Buchnera APS*, le premier est une distribution uniforme des permutations des bases du génome de *Buchnera APS*, pour le deuxième modèle la permutation des bases a été faite à l'intérieur de chacune des parties codantes et non-codantes du génome. Nous n'avons pas utilisé un modèle nul pour le chromosome d'*E. coli*, car ces données sont publiées par Pedersen et al. (Pedersen *et al.,* 2000).

Le positionnement des distributions de *Buchnera APS* et d'*E. coli* reflètent fortement leurs compositions relatives en bases G et C. Le chromosome de *Buchnera APS* est globalement plus courbé, plus flexible est moins stable que celui d'*E. coli* (**Figure 31**). La position du génome de *Buchnera APS* par rapport à la distribution du premier modèle est la même que pour *E. coli*. Les distributions du deuxième modèle nul se rapprochent beaucoup des vraies distributions. Dans la discussion, nous reprendrons les résultats de cette comparaison de distribution étendus aux quatre espèces de *Buchnera APS*.

2.3.2 Analyse des régions géniques de *Buchnera APS*

La comparaison des distributions globales des propriétés structurelles, par rapport aux deux modèles nuls suggère que l'alternance de régions géniques, ayant des compositions distinctes en bases, intervient dans l'allure de ces distributions (**Figure 31**). Les types de régions géniques (les régions intergéniques tandem, convergentes, divergentes et les régions codantes) sont des régions se distinguant par leurs éléments structurels, et aussi par leur composition : les régions divergentes sont supposées contenir plus de promoteurs et être plus prônes à l'initiation (ouverture) de la transcription alors que les régions convergentes ne devraient pas contenir de promoteurs. Nous avons alors comparé leurs propriétés structurelles dans le but de mettre en évidence des propriétés notamment des régions enrichies en promoteurs (**Figure 32**).

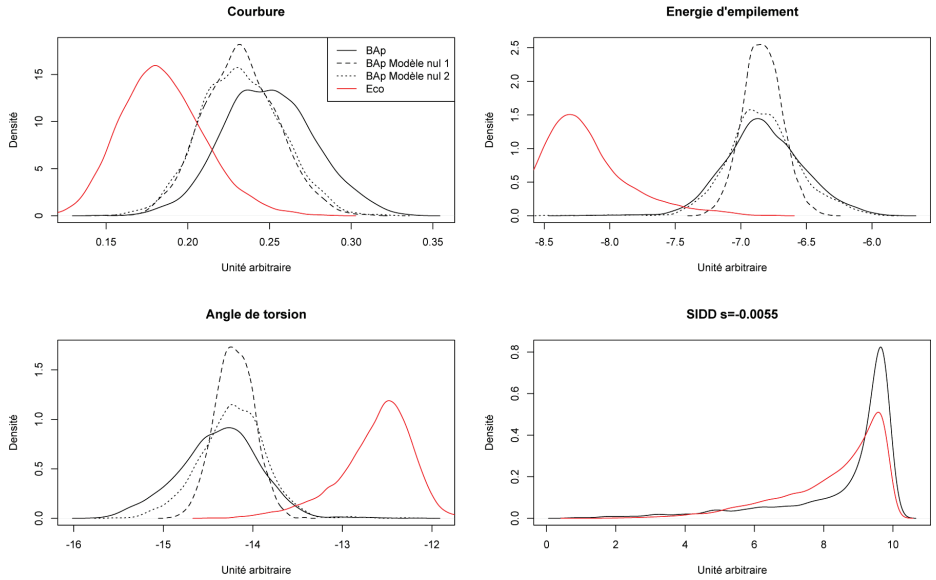


Figure 31. Comparaison des distributions globales de la courbure, de l'énergie d'empilement des bases, de l'angle de torsion et du SIDD de *Buchnera APS* par rapport à *E. coli*.

La courbure ne nous permet pas de distinguer ces différents types de régions entre elles. L'énergie d'empilement des bases et l'angle de torsion opposent les régions intergéniques aux régions codantes, mais ne permettent pas de distinguer les différents types de régions intergéniques. Les distributions du SIDD de ces différents types de régions sont par contre bien différentes. Les séquences codantes sont les plus stables. Les séquences convergentes et les régions tandem ont tendance à être stables aussi. Les régions divergentes, au contraire sont plutôt instables.

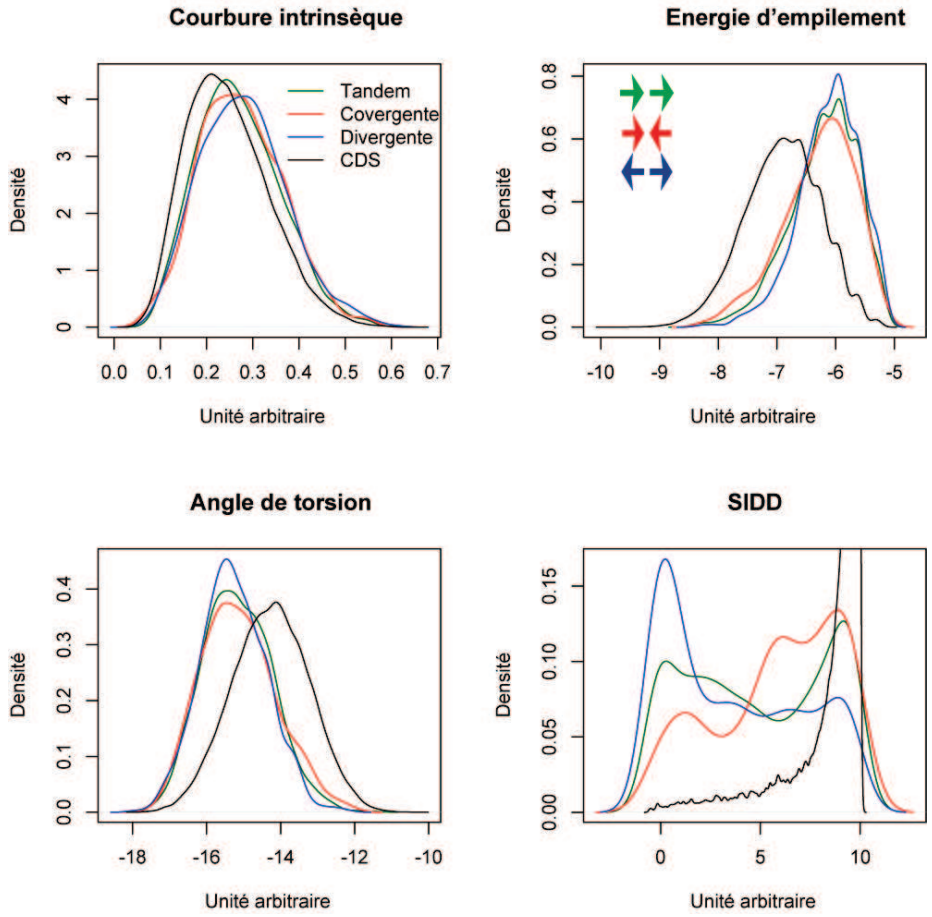


Figure 32. Distributions des paramètres physiques dans les régions intergéniques (tandem, convergentes et divergentes) et dans les régions codantes de *Buchnera APS*.

Nous avons pensé que l'aspect « bi-modal » de la distribution des régions tandem pourrait être dû au mélange des régions intergéniques tandem inter-UT et des régions tandem intra-UT. En effet, en séparant les distributions de ces régions (inter : courbes noires et intra : courbe rouge, **Figure 33**), nous constatons qu'une proportion plus importante des régions intra-UT situées autour du pic des valeurs stables. Les régions inter-UT semblent être distribuées de façon équivalente entre les valeurs « stables » et « instables » de SIDD. Nous pouvons interpréter cette observation de deux manières : soit parmi les UT, certaines sont stables et nécessitent un plus grand apport d'énergie pour l'initiation de la transcription, soit les régions inter-UT stables sont des faux positifs et correspondent à des UT longues comme semble l'indiquer notre validation expérimentale des UT par RT-PCR.

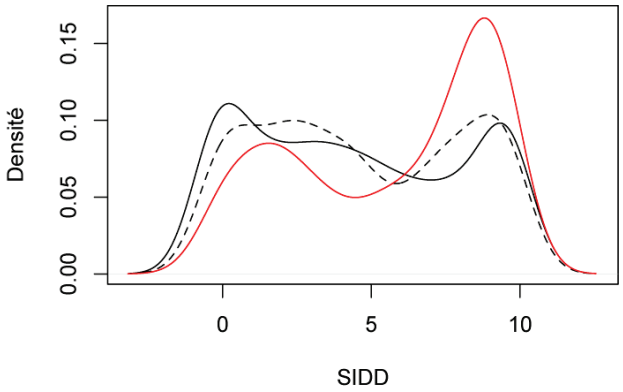


Figure 33. SIDD des régions intergéniques tandem. Ces régions intergéniques tandem ont été séparées en trois classes : début d'opéron (noir, continu), début de singleton (noir, discontinu) et intérieur d'opéron (rouge). On peut remarquer que les régions tandem à l'intérieur des opérons ont une tendance plus marquée à être stables.

2.3.3 Analyse des régions promotrices

Le but final de cette analyse étant de séparer, si possible, des groupes de régions promotrices, et donc de gènes, en espérant que ces groupes soient cohérents (corrélés) avec certaines fonctions biologiques ou certains types de métabolisme par exemple, nous avons décidé de faire une analyse des régions promotrices.

Nous avons commencé par construire les profils des régions centrées autour du codon start des séquences codantes, avec 250 pb de chaque coté (**Figure 34**).

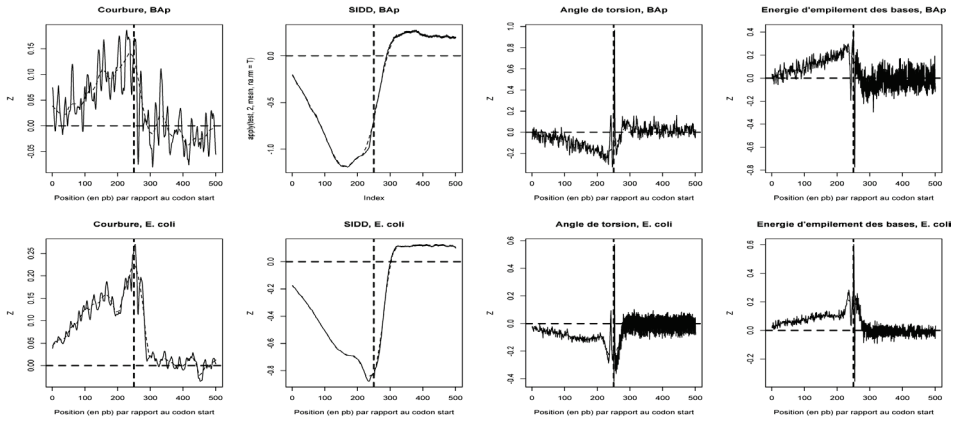


Figure 34. Profils de la courbure, du SIDD, de l'angle de torsion et de l'énergie d'empilement des bases des séquences de 500 pb, centrées autour du codon start de chaque gène. La courbe en pointillé représente le profil lissé.

Le paramètre pour lequel la variation le long du profil est la plus forte et le SIDD. La forme du profil de la courbure de *Buchnera APS* ressemble à celle d'*E. coli* (augmentation à partir de 200 pb en amont du codon start pour un maximum autour du codon start).

Le SIDD semble avoir le même aspect pour les deux bactéries, par contre le minimum est situé à 100 à 150 pb en amont du codon start chez *Buchnera APS* alors qu'il est autour du codon start chez *E. coli*. Cet aspect ainsi que la position des promoteurs σ^{70} prédits nous ont déterminés à utiliser les régions promotrices RP_{150} pour la suite de notre analyse.

Enfin, nous avons aussi séparé la courbure et le SIDD de *Buchnera APS*, chacun en deux profils : le profil des gènes en début d'UT et le profil des gènes intra-UT (**Figure 35**). La zone de variation du profil et plus étroite est plus rapprochée du codon start dans le profil des gènes intra-UT. Dans un premier temps nous avons pensé que ce résultat s'explique par le fait que les régions intergéniques en amont de ces gènes sont plus courtes, et probablement plus riches en bases G et C. La comparaison des compositions des régions intergéniques, ainsi que des régions -80 à +40 autour du codon start ont infirmé ces hypothèses. La seule explication possible est que dans le calcul des valeurs de SIDD, l'algorithme tient compte des régions voisines assez éloignées, contrairement aux autres méthodes de calcul de propriétés structurelles.

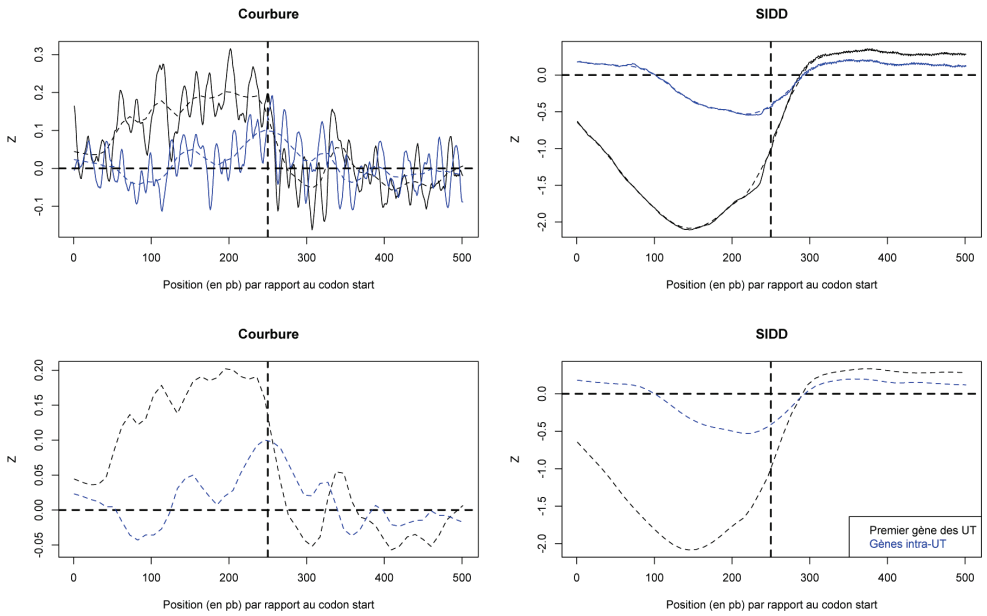


Figure 35. Profils de la courbure et du SIDD de *Buchnera APS* pour les gènes en début d'unité de transcription et les gènes intra-unités de transcription.

Relation entre la stabilité des régions promotrices des gènes et leur classe fonctionnelle

Nous avons analysé les valeurs de SIDD des RP_{150} en fonction de la classe fonctionnelle annotée par Riley et al. (1998) des gènes associés.

Seules les classes « r » (régulation) et « t » (transporteurs) sont significativement associées avec les RP_{150} à faible valeur de SIDD (test de Wilcox, p-value = 0.0002 et 0.03). Par conséquent, les gènes impliqués dans la régulation et les gènes des transporteurs ont des régions promotrices qui ont une plus grande probabilité de s'ouvrir et sont donc plus favorables à l'initiation de la transcription.

Relation entre la structure des régions promotrices et la régulation par les NAP de leur orthologue chez *E. coli*

Il semble que les NAP se lient à l'ADN majoritairement de façon non-spécifique, en reconnaissant plus une structure qu'une séquence bien définie (*e.g.* une séquence d'ADN courbée (Johnson *et al.,* 2005)). En tenant compte de cet aspect, nous avons regardé si les cibles des différents NAP chez *E. coli* sont caractérisées par des RP_{150} plus courbées ou avec de valeurs de SIDD plus faibles ou plus fortes. Aucune association significative n'a été trouvée. Le résultat reste non significatif lorsque nous faisons la même analyse mais avec les gènes orthologues de *Buchnera APS* des cibles des NAP chez *E. coli*.

Relation entre le type des régions promotrices des gènes de *Buchnera APS* et la sensibilité de l'expression de leur orthologue chez *E. coli* au niveau de surenroulement

Peter et al. (2004) proposent une liste de 306 gènes d'*E. coli* qui sont sensibles aux variations de surenroulement (expression différentielle significative). Parmi ces 306 gènes, 61 ont un orthologue chez *Buchnera APS*. Nous avons regardé si les 306 gènes chez *E. coli* et les 61 chez *Buchnera APS* sont significativement associés avec les régions promotrices instables ou stables. Les tests sont non significatifs, et ils restent non significatifs même quand on sépare les gènes en deux groupes : ceux qui sont activés par la relaxation du chromosome et ceux qui sont inhibés.

Relation entre le type des régions promotrices de *Buchnera APS* et la classe métabolique de ses gènes

La classification de la fonction métabolique des gènes d'*E. coli* et par orthologie celle des gènes de *Buchnera APS*, est celle proposée par Seshasayee et al., 2009. Chez *E. coli*, les gènes de l'anabolisme et du catabolisme ont des régions promotrices (RP_{150}) significativement moins courbées que le reste du génome. Néanmoins, chez *Buchnera APS* nous ne constatons pas d'association significative. Par contre, les gènes du catabolisme de *Buchnera APS* sont significativement associés avec des RP_{150} plus stables que le reste du génome (**Tableau 14**).

Tableau 14. p-values des tests de l'association des types de régions promotrices (RP_{150}, stables, instables) avec le type de métabolisme (anabolisme, catabolisme, central et énergétique) dans lequel le gène est impliqué (tests de Wilcox, p-values) et effectifs de gènes conservés chez *Buchnera APS*.

	Anabolisme (339 chez *E. coli*)	Catabolisme (186 chez *E. coli*)	Métabolisme central et énergétique (109 chez *E. coli*)
E. coli - SIDD	0.80	0.65	0.48
BAp - SIDD	0.53	0.02*	0.40
E. coli - courbure	0.01*	0.002*	0.2907
BAp - courbure	0.31	0.56	0.36
Conservé chez BAp	128 (37.76 %)	16 (8.6 %)	34 (31.19 %)

Relation entre le type des régions promotrices de *Buchnera APS* et leur expression différentielle dans les différentes population de *Buchnera APS* (bactériocytaires *vs.* les embryonnaires)

Les valeurs du SIDD des RP_{150} de gènes différentiellement exprimés dans les conditions testées par Bermingham *et al.* (Bermingham *et al.,* 2009) ne sont pas significativement corrélées avec le niveau basal de transcription de ces gènes, mésuré par Viñuelas *et al.* (2007) – (données non présentées).

Relation de l'expression des gènes chez *Buchnera* et les propriétés de leurs régions promotrices

Nous avons recherché une corrélation entre les propriétés structurelles séquence-dépendantes des régions promotrices et le niveau basal de transcription de ces gènes, mésuré par Viñuelas *et al.* (2007) chez *Buchnera APS*. Aucune corrélation significative n'a été trouvée.

3 Le réseau de régulation de la transcription chez *Buchnera APS*

3.1 Reconstruction du réseau

En utilisant exclusivement les relations d'orthologies, établies par BBH (bidirectional best BLAST hits), nous avons construit un premier réseau, composé de 84 gènes connectés par 110 interactions à partir du réseau de régulation transcriptionnelle d'*E. coli* (RegulonDB, version 6.1).

Ce réseau orthologue a été étendu en utilisant les UT prédites avec notre prédicteur, DisTer. Si un des gènes d'une UT intervient dans une interaction du réseau d'orthologie, l'interaction est étendue à l'UT entière. Le réseau ainsi développé, contient 140 gènes et 194 interactions régulatrices (**Figure 36**).

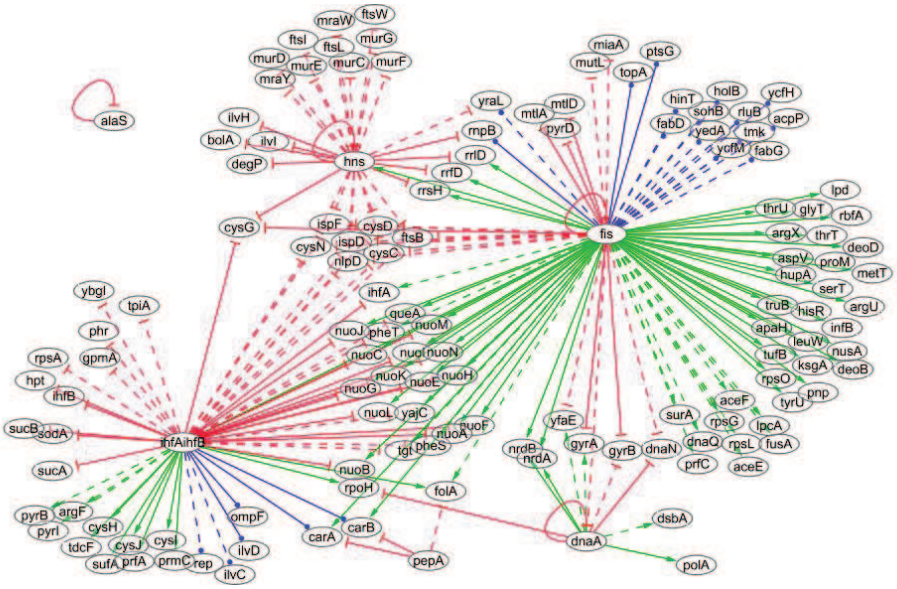


Figure 36. Réseau orthologue étendu de *Buchnera APS*. Les nœuds portent le nom du gène du régulateur ou du gène régulé. Les arêtes du réseau sont colorées de la façon suivante : rouge pour indiquer une inhibition, vert pour une activation et bleu pour une régulation duale. Les interactions déduites à partir des unités de transcription prédites avec DisTer sont tracées en pointillé, en continu étant tracées les interactions déduites par relation d'orthologie avec *E. coli*.

Nous avons voulu étendre encore ce réseau en rajoutant le régulon du facteur σ^{32} et en insérant dans le réseau des cibles des NAP grâce à la recherche de leurs sites de fixation, et en cherchant d'autres indicateurs comme par exemple les paramètres physiques qui caractérisent les régions promotrices nécessitant une régulation chez *E. coli*. Or, comme nous l'avons vu plus haut, les sites de fixation des NAP bien que présents en grand nombre n'apparaissent pas plus souvent qu'attendu sous l'hypothèse du hasard. Par conséquent, nous n'avons pas pu utiliser ces sites de fixation pour l'enrichissement du réseau.

3.1.1 Les régions promotrices à SIDD faible

La liaison des facteurs de transcription à l'ADN est connue pour délocaliser le niveau d'instabilité de l'hélice d'ADN, de façon à faciliter l'ouverture de l'hélice au niveau du promoteur (Sheridan *et al.,* 1998; Sheridan *et al.,* 1999; Dorman 2008). En nous inspirant de cette idée, nous avons cherché à déterminer si le réseau de régulation d'*E. coli* est enrichi en gènes ayant des régions promotrices plutôt instables ou plutôt stables, le but étant, si cette

analyse est significative, d'introduire les gènes de *Buchnera APS* ayant de régions semblables dans son réseau de régulation. Nous avons ainsi trouvé chez *E. coli* que les régions promotrices des gènes codant pour les régulateurs sont significativement instables (SIDD plus faibles), ce qui ne semble pas être le cas des régions promotrices des gènes régulés. L'ensemble des régions promotrices des gènes codant pour les régulateurs que nous avons énumérés au §2.1 a été montré comme étant significativement associé avec des valeurs faibles de SIDD. Néanmoins, les observations faite sur *E. coli* ne nous permettent pas d'étendre le réseau de *Buchnera APS*.

Finalement, ni les recherches des sites de fixation, ni l'analyse des paramètres physiques ne nous ont permis d'étendre plus loin le réseau de *Buchnera APS*.

3.1.2 Confrontation du réseau de *Buchnera APS* avec les données de transcription

Pour essayer de tester le réseau que nous avons construit, nous avons regardé si les gènes trouvés exprimés différentiellement pour Birmingham et al., (2009) entre les différents stades de développement des embryons de puceron, sont présents dans ce réseau. La demande métabolique du puceron varie beaucoup selon son stade de développement, aussi, nous pouvons penser que les *Buchnera APS* qui sont très peu nombreuses dans les très jeunes stades embryonnaires sont en phase de croissance exponentielle (ou du moins rapide), alors que dans les embryons plus âgés, elles passent en phase stationnaire. Or, chez *E. coli* par exemple, le passage d'une stade exponentiel à un stade stationnaire de croissance implique un changement global du profil de transcription et la participation de nombreux facteurs de transcription. Par conséquent, les gènes s'exprimant de façon différentielle dans l'expérience de Birmingham et al. devraient apparaître dans le réseau de régulation. Néanmoins, le résultat est non-significatif (**Tableau 15**). Ce résultat peut s'expliquer de deux manières. La première est que le réseau de régulation déduit par orthologie des facteurs de transcription d'*E. coli* nécessite implicitement l'hypothèse très forte de la conservation des mêmes sites de fixation et de leur fonctionnalité chez *Buchnera APS*. Or, nous savons que les sites de fixation d'un facteur de transcription peuvent changer entre deux lignées même très proches. De plus l'association des facteurs de transcription globaux, comme ceux qui ont été conservés chez *Buchnera APS*, avec l'ADN dépend beaucoup du contexte structurel, *e.g.* du niveau local de surenroulement de la région contenant le site de fixation. Chez *Buchnera APS* ces paramètres ne sont pas forcement les mêmes étant donné sa polyploïdie et son appareil de structuration du chromosome minimaliste.

La deuxième explication, est que les populations embryonnaires de *Buchnera APS* analysées dans l'expérience de Birmingham et al. (2009) correspondent à des embryons déjà très développés et que la différence de phases de croissance de *Buchnera APS* ne soit que très peu marquée. Des analyses sur les très jeunes embryons sont prévues dans l'UMR BF2I (F. Calevro).

Tableau 15. p-values des tests Chi2 d'indépendance des listes de gènes différentiellement exprimés dans les expériences de Peter et al. (2004) chez *E. coli* et Bermingham et al. (2009) chez *Buchnera APS*, avec les gènes du réseau de régulation de *Buchnera APS* reconstruit par orthologie avec *E. coli*.

	Liste Peter *et al*. 2004	Liste Birmingham *et al*. 2009
Réseau BAp orthologue	0.7568	0.9265
Réseau BAp étendu (avec les UT)	0.9033	0.698

3.2 Vers une vision d'un système de régulation généraliste de *Buchnera APS* – une régulation par la topologie

Le réseau de régulation de *Buchnera APS* que nous avons reconstruit est gouverné essentiellement par les toporégulateurs, ce qui est représenté sur la **Figure 37** dans laquelle seules les interactions entre les régulateurs du réseau ont été représentées. Nous avons constaté qu'il s'agit exclusivement d'interactions entre des toporégulateurs, bien que dans la description de la machinerie de transcription de *Buchnera APS* (cf. Résultats, §2) nous avons décrits d'autres acteurs protéiques. On sait que chez *E. coli* l'action conjointe de ces régulateurs régit des grands changements de régime d'expression. Nous pensons que chez *Buchnera APS*, qui doit simplement être capable d'alterner entre l'expression des gènes impliqués dans les différents types de métabolisme, il suffirait d'un système de régulation variant la conformation du chromosome de façon à rendre accessible à la transcription les gènes nécessaires, et simultanément conciliant le régime de transcription avec les divers autres processus impliquant des transactions d'ADN comme la réparation, la réplication ou la compaction de l'ADN.

Le profil de transcription basal périodique de *Buchnera APS* (Viñuelas *et al.,* 2007), qui est probablement dû, comme chez les bactéries à forme de vie libre, à la conformation du chromosome (Jeong *et al.,* 2004; Carpentier *et al.,* 2005), semble confirmer cette hypothèse. Nous avons alors cherché à déceler les causes génomiques de cette périodicité.

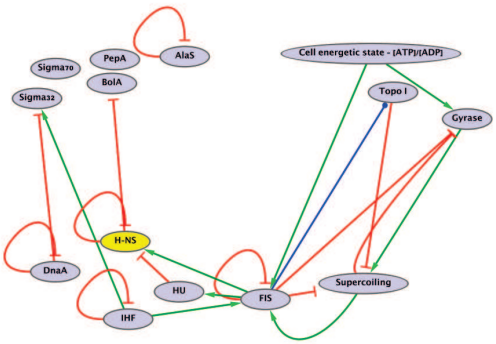


Figure 37. Réseau de régulation de *Buchnera APS* simplifié – le réseau de régulation topologique.

3.2.1 Analyse de la périodicité de la transcription des gènes chez *Buchnera APS*

Le but de cette étude est d'inspecter s'il existe une quelconque relation entre le profil périodique de la transcription (Viñuelas *et al.,* 2007) et le profil des propriétés structurelles du chromosome.

3.2.2 Analyse spectrale à pas constant du niveau d'expression des gènes, de la courbure, du SIDD et du taux de GC le long du chromosome de *Buchnera APS*

Dans l'étude de Viñuelas et al. (2007) le pas utilisé dans l'analyse spectrale a été le gène, il s'agissait donc d'un pas non constant (car les gènes n'ont pas tous la même taille). Moyenner sur des fenêtres de 100 pb (cf. Matériels et Méthodes) nous permet d'un coté de refaire l'analyse spectrale de l'expression à pas constant le long du chromosome, mais aussi de comparer les composantes (périodes) de l'expression des gènes avec celles des paramètres physiques (courbure, SIDD et taux GC).

Avant de faire l'analyse spectrale, nous avons vérifié que ces données contiennent une composante périodique significative avec le test Kappa de Fisher. Tous les tests ont été significatifs.

Sur la **Figure 38** sont représentés les périodogrammes de l'expression des gènes, du SIDD, de la courbure et du taux de GC. En rouge sont marquées les trois périodes les plus importantes (amplitude et

variance), indiquées en pb. L'expression, la courbure et le taux de GC montrent des longues périodes (de l'ordre de 100 kb). La période la plus importante de l'expression des gènes de 91.51 kb est cohérente avec la principale période trouvée par Viñuelas et al. (2007) (89.89 gènes), en considérant qu'un gène bactérien mesure en moyenne 1 kb. Les périodogrammes de la courbure et du taux de GC contiennent tous les deux la période de 106.67 kb avec une forte amplitude (première et deuxième période les plus importantes), ceci est dû à la corrélation qui existe entre ces deux paramètres. En même temps, parmi tous les paramètres physiques mentionnés, la courbure est la moins corrélée avec le taux GC (données non présentées). Donc, la corrélation forte entre un paramètre physique et le taux GC n'explique pas complètement la présence des mêmes périodes, puisque le SIDD et le taux GC ne sont pas caractérisés par les mêmes périodes, alors que le SIDD est plus corrélé avec le taux de GC que la courbure.

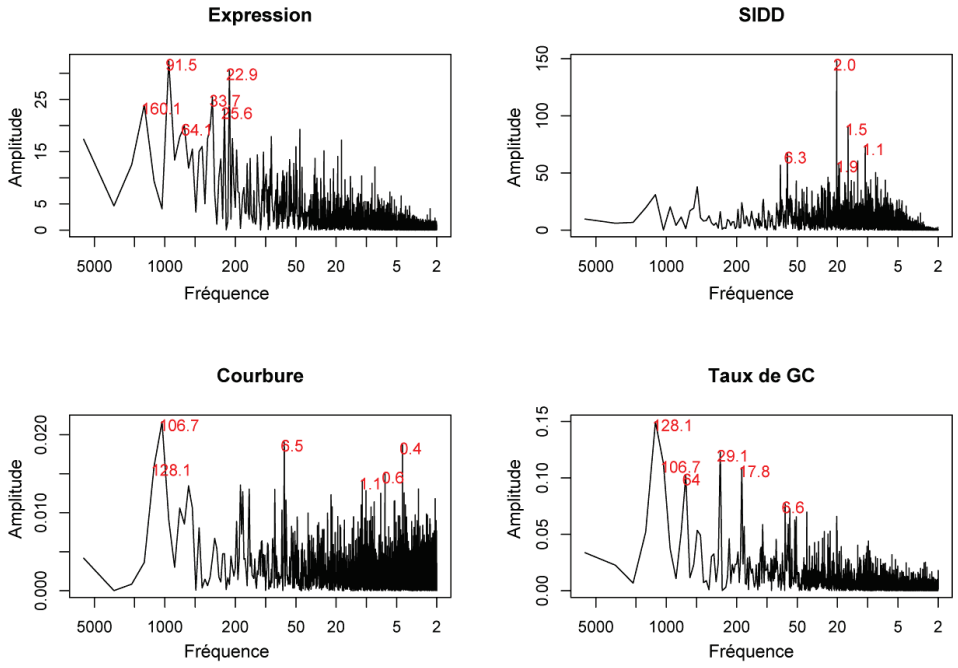


Figure 38. Périodogrammes de l'expression, de la courbure et du SIDD. Le chromosome a été divisé en fragments de 100 pb non chevauchants. En rouge, sont marquées (en kb) les cinq périodes les plus fortes de chaque série. On peut constater que la série SIDD a un périodogramme se distinguant des trois autres. Unité d'analyse : 100 pb.

Pour voir si la taille de périodes détectées n'est pas influencée par la taille de la fenêtre utilisée pour construire les profils, nous avons fait la

même analyse avec une fenêtre deux fois plus petite (50 pb). Une fois encore nous retrouvons les mêmes périodes (données non montrées).

En conclusion, il y a une régularité dans ces différents profils à une échelle approximative de 100 kb. Lorsque nous superposons le profil de la courbure du SIDD et de l'expression et en traçant des barrières tous les 100 kb, on peut voir un pic de chaque profil dans chacune des périodes (**Figure 39**). Aussi nous avons comparé les distributions de la courbure, de l'expression, du SIDD et du taux de GC dans les fragments de 100 kb non chevauchants. Les distributions sont similaires ce qui indique une régularité à échelle de 100 kb.

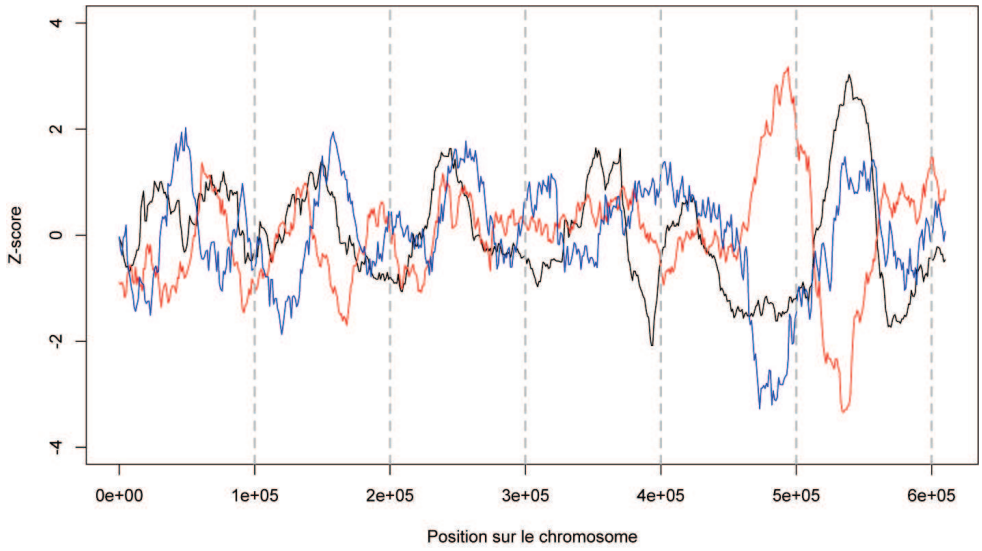


Figure 39. Profil de la courbure (rouge), du SIDD (bleu) et de l'expression (noir), du chromosome de *Buchnera APS*.

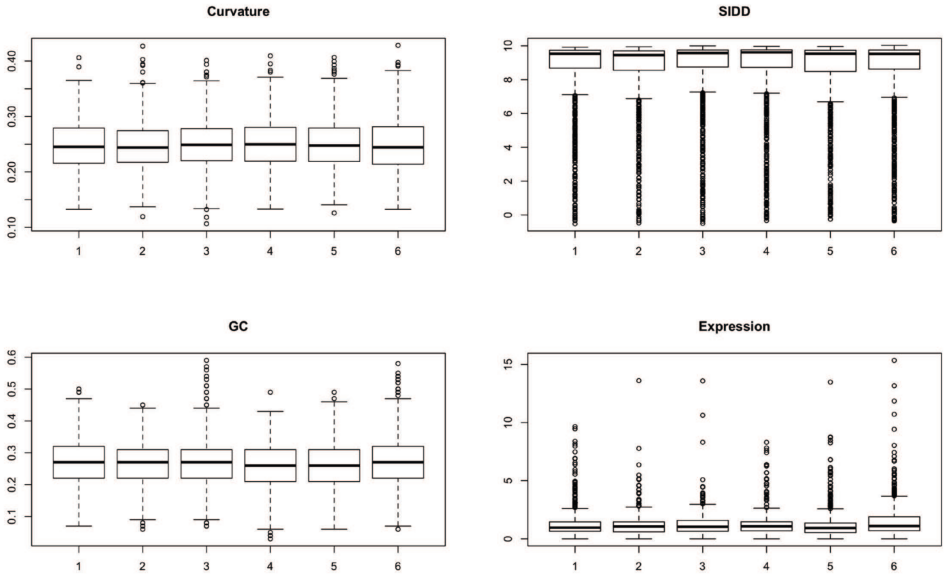


Figure 40. Boxplots des valeurs dans les domaines de 100 000 pb, définis selon les périodes trouvées dans la figure des périodogrammes (Figure 38).

3.2.2.1 *Analyse spectrale du niveau d'expression des gènes, de la courbure et du taux de GC des régions promotrices et des séquences codantes, le long du chromosome de* Buchnera APS

Afin de voir si la périodicité de l'expression des gènes concorde avec la périodicité de la courbure (le taux de GC) des régions promotrices ou plutôt avec la courbure (taux de GC) des séquences codantes nous avons fait l'analyse spectrale des valeurs moyennes des propriétés structurelles séquence-dépendantes, des régions promotrices et des séquences codantes le long du chromosome (**Figure 41**). La période 86.71 de l'expression est retrouvée dans le profil du taux GC des séquences codantes. Par contre, la périodicité à échelle de 100 kb semble être due aux régions promotrices.

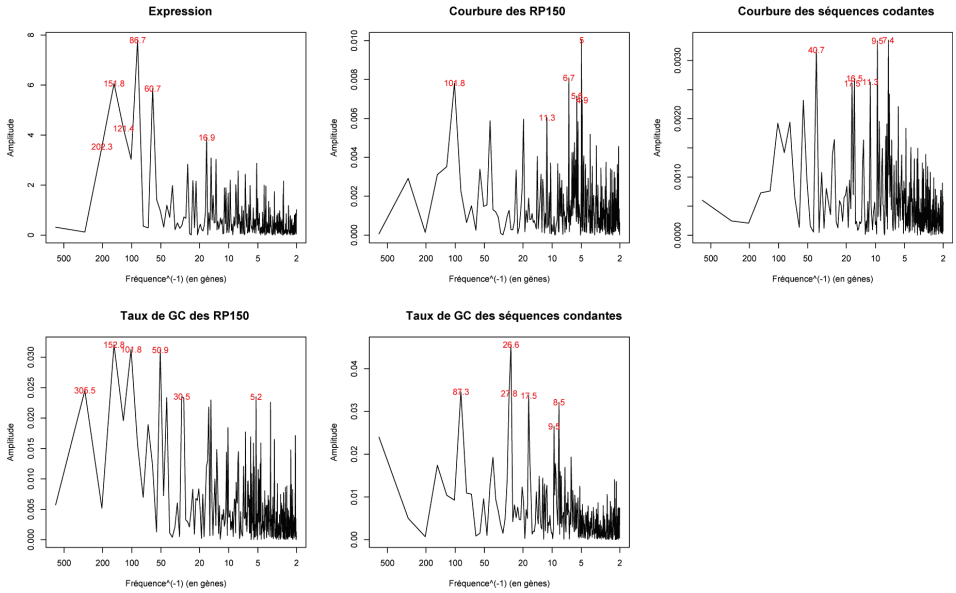


Figure 41. Périodogrammes de l'expression, de la courbure des RP_{150}, de la courbure des séquences codantes, du taux de GC des RP_{150} et du taux de GC des séquences codantes. En rouge, sont marquées les cinq périodes les plus fortes de chaque série.

4 Reconstruction descendante de réseaux de régulation à partir de données d'expression

4.1 IGOIM - inférence des graphes d'ordre 0-1 avec l'information mutuelle conditionnelle

Après avoir analysé la bibliographie des différentes approches pour l'inférence des réseaux génétiques (cf. Introduction, §5.1), nous avons décidé de nous intéresser plus particulièrement aux méthodes probabilistes. En raison du caractère stochastique de l'expression des gènes, le formalisme des modèles probabilistes graphiques semblait bien approprié pour notre étude, car se rapprochant du modèle biologique sous-jacent. Ainsi, en explorant des propriétés finalement simples des profils d'expression des gènes, les modèles probabilistes graphiques semblaient nous permettre de construire des modèles synthétiques et qualitatifs des réseaux de gènes.

La clé de voûte de toutes les méthodes d'inférence que nous avons étudiées est la dépendance conditionnelle. Comme il n'y a pas de façon univoque de tester cette indépendance conditionnelle, nous nous sommes dirigés vers l'IM (cf. Matériels et Méthodes, §4.2.1). En choisissant cette quantité, nous ne faisons aucune hypothèse sur le réseau de gènes à étudier.

Les étapes du calcul d'IGOIM sont résumées dans la **Figure 15**, leur description détaillée étant présentée dans la section 4.2.3 des Matériels et Méthodes.

La première étape de ce travail a été de choisir parmi les estimateurs existants, celui qui convenait le mieux pour l'inférence des réseaux de gènes. Le type de données (peu d'observations) ainsi que le temps de calcul nous ont obligés à réduire notre sphère de recherche aux estimateurs non-paramétriques discrets (cf. Matériels et Méthodes).

4.1.1 Comparaison du temps de calcul des estimateurs (*dataset1*)

La méthode d'inférence que nous avons implémentée nécessite un grand nombre d'estimations d'IM. Par conséquent, la complexité de calcul des estimateurs est un paramètre de choix important. Ainsi, parmi les estimateurs mentionnés au paragraphe 4.2.2, nous avons éliminé IM^{NSB} à cause de sa complexité de calcul (Hausser 2006).

En revanche, nous avons implémenté et comparé en plus de ces estimateurs, les estimateurs bayésiens ($IM^{shrink3}$, IM^{ZIP3}, IM^{ZINB3}). Les temps de calcul ont été estimés en testant chaque estimateur sur le jeu de données *dataset1* (cf. Matériels et Méthodes). La **Figure 42** donne un bon aperçu de la complexité des estimateurs, son allure ne change pas si on utilise une autre loi jointe de deux variables, ou encore si on utilise l'IM conditionnelle et non pas l'IM simple (données non présentées).

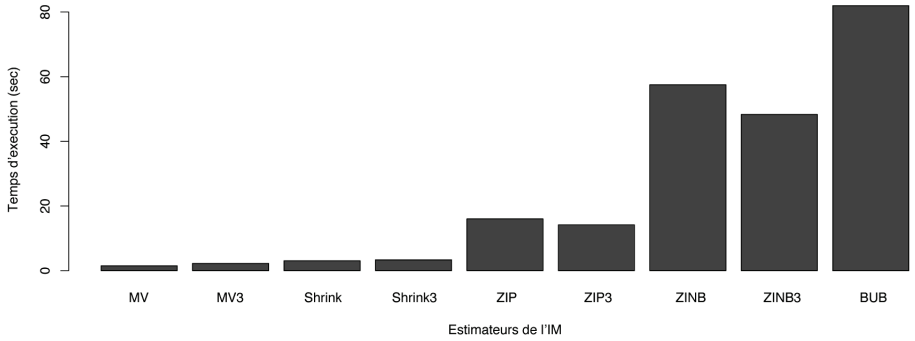


Figure 42. Comparaison des temps de calcul de différents estimateurs sur le jeu de données *dataset1* correspondant à 20 échantillons identiquement distribués d'une loi normale bivariée avec la matrice de covariance $\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$ et de moyenne $m = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

Cette étape de comparaison préliminaire nous a poussé à éliminer de la liste l'estimateur IM^{BUB}, très coûteux avec des performances ne dépassant pas celles de IM^{MV3} (Steuer *et al.,* 2002). Nous avons aussi éliminé les estimateurs IM^{ZINB} et IM^{ZINB3}. Ces deux estimateurs sont plus lents que ceux du type ZIP, alors que leurs performances, telles qu'elles ont été décrites dans (Hausser 2006) ne sont pas meilleures pour autant. De plus, Hausser (2006) avait conclu que l'estimateur ZIP convient mieux aux con-

ditions de sous-échantillonnage, ses propriétés en termes d'EQM sont meilleures pour l'estimation d'une loi multinomiale.

4.1.2 Comparaison des estimateurs de l'IM discrète (*dataset2*, *dataset3*)

Les estimateurs que nous avons comparés, en utilisant le jeu de données *dataset2* (correspondant à des échantillons de taille 10, 20, ou 50, générés avec des loi normales bivariées, cf. Matériels et Méthodes) sont IM^{MV}, IM^{MV3}, IM^{shrink}, $IM^{shrink3}$, IM^{ZIP} et IM^{ZIP3}. La propriété que nous recherchions dans ces estimateurs était leur capacité à déceler une relation de dépendance. Afin d'évaluer cette capacité, nous avons utilisé la significativité (S)[52]. Nous avons donc appliqué chacun des estimateurs sur chaque jeu de données de *dataset2*. Dans les lois bi-normales que nous avons utilisées pour les simulations ce qui est important ne sont pas les valeurs précises des paramètres de ces lois, mais plutôt leurs propriétés. Pour les trois différentes tailles d'échantillons (10, 20 et 50), nous avons testé les estimateurs sur les cinq types de lois bi-normales :

- les marginales sont identiques et peu dispersées ;
- les marginales sont identiques mais beaucoup plus dispersées par rapport au premier cas ;
- les marginales sont peu dispersées et leurs moyennes sont très différentes ;
- les moyennes sont identiques, mais les dispersions sont différentes ;
- les moyennes sont différentes et les dispersions aussi.

52 Pour chaque échantillon d'un couple de gènes, nous déterminons l'IM. Ensuite, nous permutons un certain nombre de fois chaque profil d'expression et nous estimons l'IM sur les profils permutés. Cette quantité doit être nulle car la permutation détruit la corrélation. On estime d'abord la moyenne et l'écart-type des valeurs de l'IM supposées être nulles, et on calcule $S = 1 - prob\left(\frac{IM - IM_0}{\sigma_{IM_0}}\right)$, en considérant que sous l'hypothèse nulle $\frac{IM - IM_0}{\sigma_{IM_0}} \sim N(0,1)$.

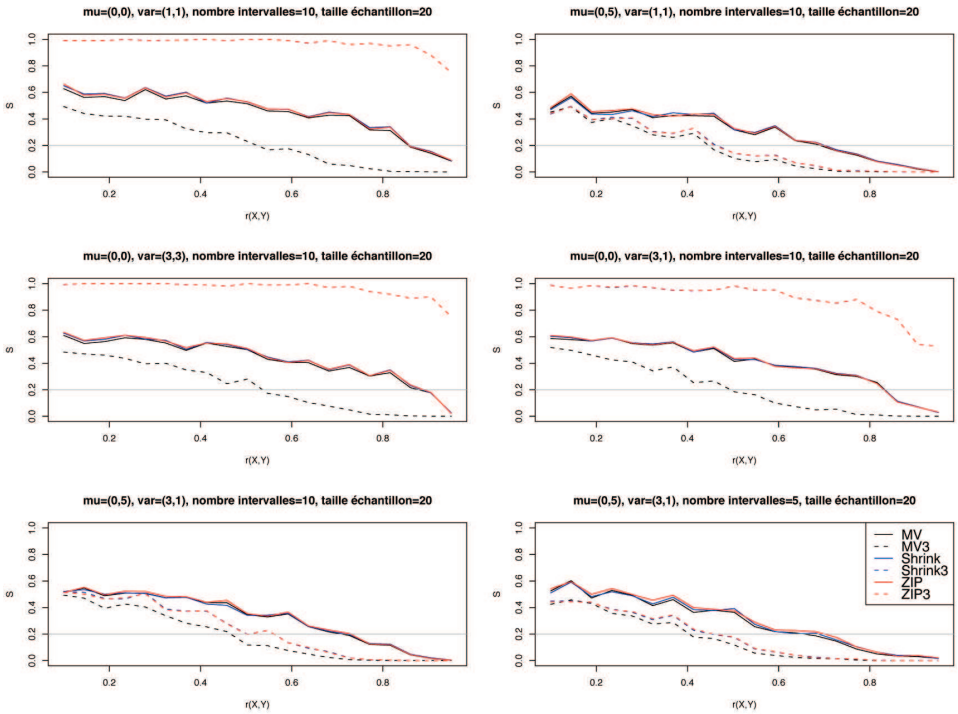


Figure 43. Comparaison des performances des estimateurs non-paramétriques de l'IM discrète sur le jeu de données *dataset2* (correspondant à des échantillons de taille 10, 20, ou 50, générés avec des loi normales bivariées, cf. Matériels et Méthodes).

Enfin, pour le dernier cas nous avons comparé les performances de chaque estimateur pour deux nombres différents d'intervalles (10 et $\sqrt{n}$, comme suggéré par Tukey (Mosteller and Wilder Tukey 1977) et proche de celui de Bendat et Piersol (Bendat and Piersol 1968)). Etant donné que les résultats sont quasiment les mêmes pour les trois tailles d'échantillons, nous ne présentons que ceux obtenus pour n = 20. Néanmoins, la mesure de la significativité S d'une seule simulation est relativement variable, surtout lorsqu'il s'agit d'échantillons de petite taille. Nous avons donc déterminé S sur plusieurs échantillons simulés selon la même loi, et la moyenne de S a été représentée sur la **Figure 43**. Le graphe des variances n'est pas présenté car mis à part pour IM^{MV3}, pour tous les autres estimateurs, la variance de S est du même ordre de grandeur, et elle reste constante en fonction de r. Pour IM^{MV3} cette variance est nettement inférieure aux variances des autres estimateurs, de plus elle diminue si le coefficient de corrélation augmente. Enfin, IM^{MV3} est toujours plus sensible que les autres pour le même niveau de significativité.

On peut remarquer que les courbes caractérisant les estimateurs IM^{MV}, IM^{shrink}, et IM^{ZIP} se superposent. En effet, comme nous l'avons décrit dans le deuxième chapitre, chacun de ces estimateurs est une amélioration du précédent, s'ils se superposent, c'est que les données sont telles que les estimateurs IM^{ZIP} et IM^{shrink} n'ont pas détecté d'intervalles de probabilité nulle, et que leurs intensités de "shrinkage" ont été nulles.

Grâce à ces résultats, nous avons choisi d'utiliser pour notre méthode d'inférence, un seuil de significativité S=0.2. Avec ce seuil, on peut espérer pouvoir détecter des corrélations supérieures à 0.4, tout en conservant une certaine spécificité.

Afin de nous assurer que les performances de l'ensemble des estimateurs n'étaient pas particulières à la loi bi-normale et que l'estimateur IM^{MV3} restait toujours le meilleur, nous avons simulé un autre jeu de données *dataset3*, cf. Matériels et Méthodes. Les variables qui ont servi à générer le *dataset3* sont liées par une fonction du type "dose-réponse", plus proche de la relation régulateur/effecteur. Le désavantage d'utiliser une fonction, même avec un certain bruit, est que les données ne sont plus simulées à partir d'une loi de probabilité et que l'échantillonnage n'est pas homogène pour les deux variables (on échantillonne uniformément pour la première variable, mais pas pour la deuxième qui est fonction de la première). Pour ce jeu de données IM^{MV3} est toujours l'estimateur le plus sensible, et cette fois ce sont les estimateurs IM^{shrink} et IM^{ZIP} qui se rapprochent de lui (**Figure 44**).

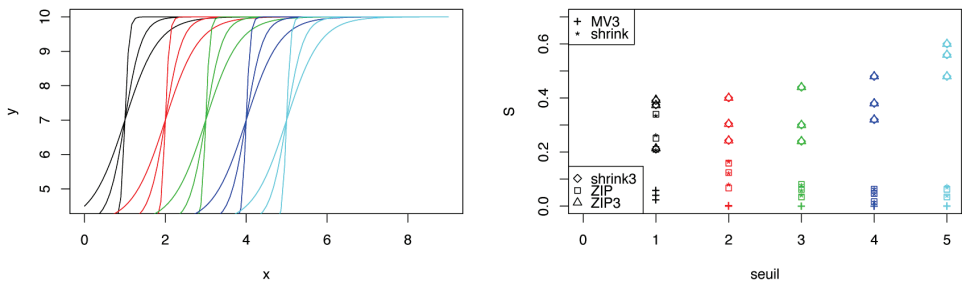


Figure 44. Comparaison des performances des estimateurs non-paramétriques de l'IM discrète sur le jeu de données *dataset3* (les variables sont dans ce cas liées par une fonction de type « dose-réponse »). Les couleurs permettent d'établir la correspondance entre les S des différents estimateurs et la fonction qui a généré l'échantillon sur lequel ces estimateurs ont été appliqués.

Finalement, pour implémenter notre méthode d'inférence, nous avons décidé de garder la possibilité de choisir parmi trois estimateurs de

l'IM : IM^{MV3} (le plus sensible), IM^{ZIP} (le plus proche de IM^{MV3} pour le *dataset3*) et IM^{ZIP3} (le plus proche de IM^{MV3} pour le *dataset2*).

4.1.3 Analyse et comparaison d'IGOIM grâce à des données d'expression simulées

L'étape suivante a été d'implémenter notre méthode IGOIM, et de la confronter aux méthodes préalablement publiées : les modèles gaussiens graphiques (Schäfer and Strimmer 2005) et ARACNE (Margolin *et al.,* 2006), en utilisant divers jeux de données simulées et en testant l'influence de différents paramètres.

Il n'est pas possible de générer des données d'expression statiques d'un réseau de gènes, il est nécessaire de passer par la modélisation d'un système dynamique. Nous avons alors choisi d'utiliser les deux ensembles de jeux de données de Bansal et di Bernardo (2007), et de Mendes *et al*. (2003), basés sur des systèmes d'équations différentielles couplées. La principale différence entre ces deux ensembles de jeux de données est que le modèle dynamique du premier est un système d'équations différentielles linéaires alors que celles du deuxième sont non linéaires. Le jeu de données de Mendes est décrit comme l'un des plus proches du modèle biologique, cependant, l'analyse visuelle du jeu de données a été très décevante. Les profils d'expression étaient constitués par des valeurs très peu variables, beaucoup de valeurs (correspondant aux différentes perturbations) étaient identiques. Ce jeu de données a finalement été conservé même s'il ne nous laissait que peu d'espoir d'obtenir une validation de nos méthodes.

Dans le cadre d'une analyse rigoureuse, on se doit d'étudier non pas un seul réseau artificiel, mais plusieurs ayant les mêmes propriétés générales (topologie, taille, bruit), afin que les conclusions concernent une classe de réseaux, et non pas un réseau particulier. À cause du temps de calcul, cette analyse systématique n'a été réalisée que pour les réseaux de 10 gènes du jeu de données linéaires. Ceci nous a permis de constater que pour les réseaux de 10 gènes, les résultats restaient les mêmes d'un réseau à l'autre.

4.1.3.1 Analyse des réseaux de gènes linéaires (dataset4)

Le *dataset4* contient des données simulées avec des réseaux linéaires de 10 gènes et 100 gènes. Pour chaque type de réseau nous avons appliqué l'ensemble des méthodes sur (1) le jeu complet de données à l'équilibre après perturbations locales ; (2) le jeu incomplet de données à l'équilibre après

perturbations locales (quand cela a été possible) ; (3) le jeu de données à l'équilibre après perturbations globales.

Les performances de chaque méthode sont mesurées par la spécificité (Sp[53]) et la sensibilité (Se[54]). Sachant qu'aucune des méthodes étudiées ne détecte les autorégulations et que dans les réseaux de *dataset4*, chaque gène s'auto-inhibe, nous avons décidé de ne pas comptabiliser ce type d'interaction dans les calculs de spécificité et de sensibilité.

Pour les MGG la probabilité associée au coefficient de corrélation partiel est utilisée pour inférer la présence d'une arête. Le seuil de cette probabilité a été choisi à 0.2 (le même que pour notre méthode).

Pour ARACNE, on peut intervenir au niveau de deux paramètres : la p-value de l'IM et la tolérance sur les cycles. Pour la tolérance les auteurs conseillent une valeur optimale que nous avons choisie, et pour la p-value de l'IM nous avons utilisé un seuil de 0.1.

Analyse de petits réseaux linéaires (10 gènes)

Ces données (un sous-ensemble de *dataset4*) ont été générées avec des petits réseaux de 10 gènes, chaque gène est lié en moyenne, à deux autres gènes (Bansal and di Bernardo 2007).

Les résultats des différentes méthodes sur les jeux de données complets après perturbations locales sont présentés dans la **Figure 45**. Ils sont les mêmes pour presque tous les réseaux. Dans la majorité des cas la meilleure estimation est obtenue avec notre méthode (estimateur IM^{MV3}). On peut également constater l'invariabilité de la méthode IM^{MV3} aux différents nombres d'intervalles.

Pour les données des réseaux de 10 gènes nous n'avons pas pu utiliser un sous-ensemble des données de perturbations locales, car avec des échantillons de taille inférieure à 10 aucune confiance ne peut être obtenue avec les résultats.

53 $Sp = \frac{VP}{VP + FP}$, *VP* est le nombre de vrais positifs, et *FP* le nombre de faux positifs ; *Sp* est le taux de vrais positifs parmi les interactions inférées.

54 $Se = \frac{VP}{VP + FN}$, est la proportion de vraies interactions détectées parmi toutes les interactions.

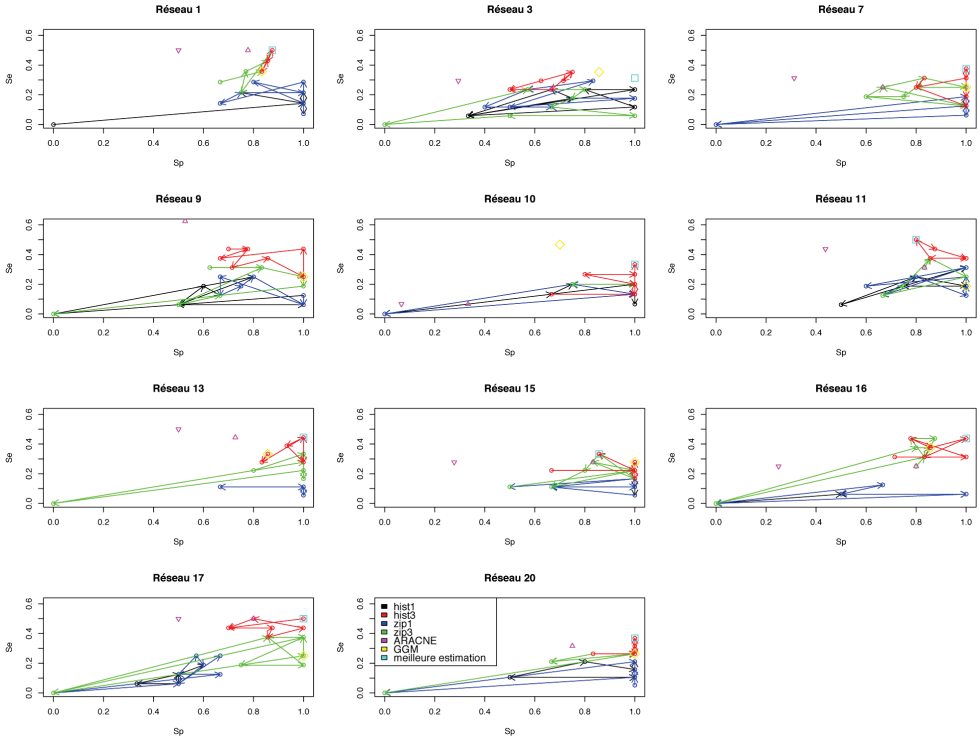


Figure 45. Performances de différentes méthodes d'inférence sur l'ensemble complet des données d'expression à l'équilibre, après perturbation locale des différents réseaux de 10 gènes. Pour chaque estimateur, IGOIM a été utilisée avec différents nombres d'intervalles pour la discrétisation (de 3 à 10), les flèches vont dans le sens croissant du nombre d'intervalles. Ne sont présentés que les résultats pour lesquels les trois différentes méthodes (IGOIM, MGG et ARACNE) ont pu être testées. Nous avons gardé pour ce calcul l'estimateur IM^{MV}, et comme on l'avait supposé, il échoue dans très nombreux cas, (on ne voit pas les flèches noires car la spécificité et la sensibilité, sont nulles pour cette méthode).

Les résultats de la **Figure 46** soutiennent les conclusions que nous avions avancées avec le jeu de données complet après perturbations locales à savoir qu'IGOIM (IM^{MV3}) est la méthode la plus performante et que IGOIM avec l'estimateur IM^{ZIP3} obtient de bons résultats, mais nous n'avons pas une idée précise des paramètres optimaux pour ce cas.

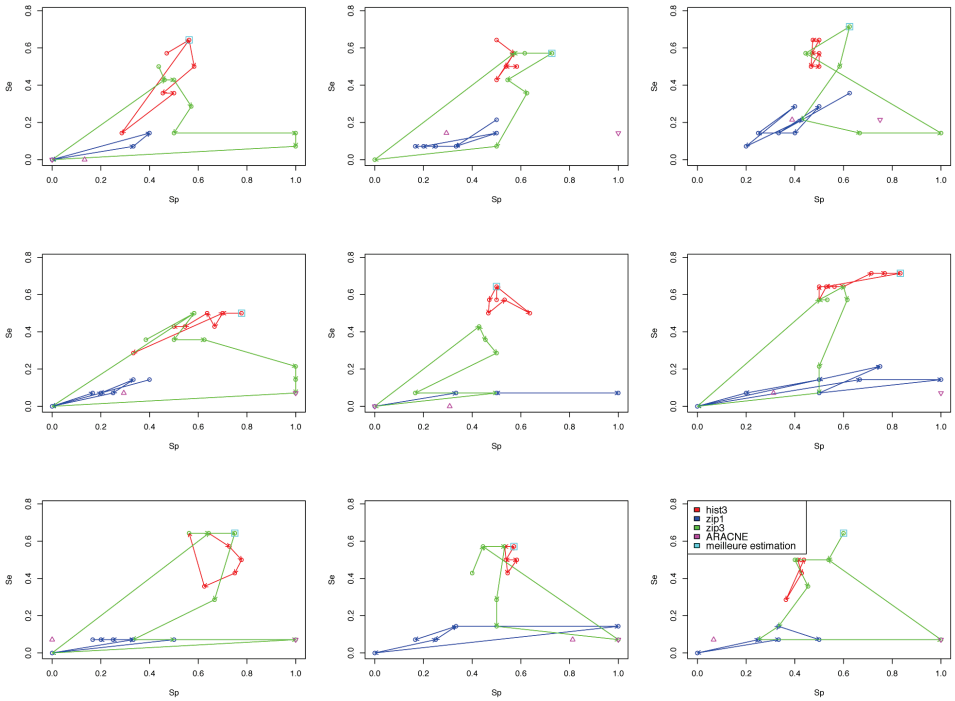


Figure 46. Performances des différentes méthodes d'inférence sur des ensembles des données d'expression à l'équilibre, après 20 différentes perturbations globales, ces données ont été générées avec un même réseau de 10 gènes.

Analyse des réseaux linéaires de 100 gènes

Pour les réseaux de 100 gènes nous avons utilisé un seul nombre d'intervalles pour la discrétisation qui est $\sqrt{n}$. D'après les résultats présentés dans le **Tableau 16**, la méthode IGOIM, avec l'estimateur IM^{MV3} est encore la plus performante. Le choix de discrétisation en 10 intervalles semble être « correct » dans le cas de l'estimateur IM^{ZIP3}, il est possible que pour un autre nombre d'intervalles, des résultats encore meilleurs que ceux-ci soient obtenus avec IGOIM, (IM^{ZIP3}).

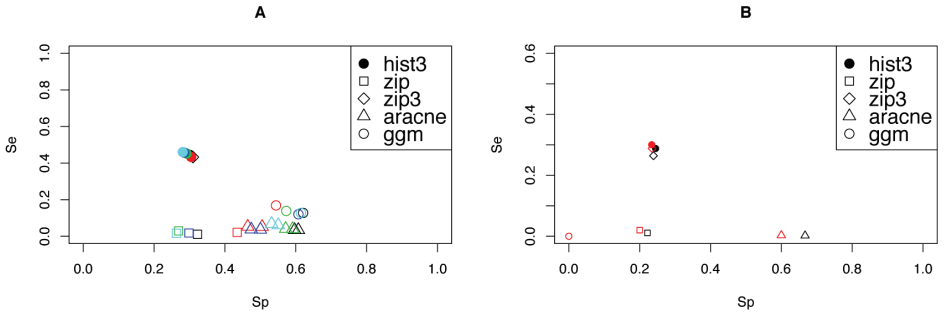


Figure 47. Performances des différentes méthodes d'inférence sur (A) cinq échantillons de 20 perturbations locales et (B) deux échantillons de 20 perturbations globales, générées avec un réseau de 100 gènes.

Les résultats présentés dans la **Figure 47** mettent en valeur la méthode IM^{MV3}. La méthode IM^{ZIP3} peut donner de bons résultats, mais il est probable que le choix des paramètres que nous avons faits ne soit pas optimal pour cet estimateur.

Tableau 16. Performances des différentes méthodes d'inférence sur l'ensemble complet des données d'expression à l'équilibre, après perturbation locale, d'un réseau de 100 gènes.

	Sensibilité	Spécificité
MV3	0.646	0.452
ZIP	0.047	0.346
ZIP3	0.640	0.450
ARACNE	0.091	0.647
MGG	0.128	0.615

4.1.3.2 Analyse de réseaux de gènes non-linéaires (dataset5)

Pour ce jeu de données (modèle non-linéaire), comme nous l'avions supposé, les performances sont très modestes (**Tableau 17**).

La cause principale de la mauvaise qualité des prédictions des méthodes est certainement la forme des données. De plus, dans ces réseaux on a un nombre réduit d'interactions (200 en moyenne) par rapport au nombre de gènes (100), sans oublier qu'il faut soustraire les autorégulations. Pour

la méthode IGOIM (IM^{MV3}) la sensibilité est souvent très bonne, ce qui veut dire qu'elle arrive à détecter une grande partie des interactions, mais le problème est qu'elle infère beaucoup de faux positifs, car la spécificité est de l'ordre de 0.1. La méthode IGOIM (IM^{MV3}) donne les meilleurs résultats pour les données générées par un réseau de type "petit monde", des résultats intermédiaires pour le "scale-free", et finalement les moins bons résultats, pour le réseau aléatoire.

Tableau 17. Performances des différentes méthodes d'inférence sur l'ensemble complet des données d'expression à l'équilibre, après perturbations locales, de trois réseaux de 100 gènes, un réseau avec une topologie aléatoire (RND), un autre avec une topologie "petit-monde" (SW), et enfin un dernier avec une topologie "scale-free" (SF). Dans chaque case du tableau, les résultats sont présentés sous la forme $\frac{Se}{Sp}$.

	MV3	ZIP	ARACNE
RND	0.765	0.102	0.019
	0.055	0.080	0.544
SF	0.644	0.021	0.320
	0.125	0.400	0.525
SW	0.990	0.060	0.211
	0.100	0.120	0.215

Nous avons aussi essayé d'utiliser des sous-ensembles de données du jeu complet de données de perturbations locales de chaque type de réseau, en rajoutant aux profils d'expression du bruit (loi normale de moyenne 0 et d'écart-type égal à 5, 10 et 15 % de la valeur maximale du profil d'expression, données non-présentées). Pour IGOIM, avec l'estimateur IM^{MV3} les performances sur le jeu de données bruitées sont de même ordre que pour le jeu de données non-bruitées (les MGG et ARACNE échouent constamment), voire légèrement meilleures, ce qui confirme que le principal problème de nos résultats est la forme des données.

4.2 Discussion sur la méthode et la validation d'IGOIM

Deux jeux de données (*dataset4* et *dataset5*) ont été utilisés pour étudier l'influence (1) de la taille des réseaux, (2) du type de données (après perturbations locales ou globales), (3) de la topologie des réseaux, (4) du type

des réseaux (linéaire, non-linéaire) et enfin (5) du bruit des données, sur les performances des trois méthodes (IGOIM, MGG et ARACNE).

(1) *Analyse de la taille des réseaux*. Nous avons pu remarquer qu'en appliquant les différentes méthodes sur les ensembles complets de données, après perturbations locales, des réseaux linéaires de 10 gènes et d'un réseau de 100 gènes, la sensibilité était supérieure dans le deuxième cas, mais on obtenait moins de faux positifs dans le cas des réseaux de 10 gènes. Ce résultat n'est pas très net, mais nous nous attendions à ce que les performances d'IGOIM ne changent pas beaucoup avec la taille des réseaux, dans le cas des jeux de données complets après perturbations locales, car la quantité d'information, par gène est constante.

(2) *Analyse du type des données*. Pour les réseaux linéaires de 100 gènes, nous avons pu comparer les résultats obtenus avec des échantillons après perturbations locales et des échantillons après perturbations globales, de même taille (n = 20). On suppose que les données globales quoique plus riches en information (plus d'interactions mises en jeu) sont moins informatives pour les méthodes d'inférence. Pour ce cas précis, nous avons confirmé cette hypothèse. Nous avons aussi remarqué que IGOIM avec l'estimateur MV3 se distingue bien des autres méthodes, en ayant des résultats nettement meilleurs.

L'analyse des différentes méthodes sur ce premier ensemble de données générées à partir de réseaux linéaires semble privilégier la méthode IGOIM avec l'estimateur MV3. Le nombre d'intervalles $\sqrt{n}$ semble être un bon choix, pour IGOIM (estimateur MV3). L'estimateur ZIP3 semble très intéressant mais une analyse plus approfondie de ses propriétés devrait être faite, pour connaître quels sont les paramètres optimaux pour cet estimateur (*e.g.* nombre d'intervalles pour la discrétisation). Dans tous nos essais, les performances des MGG et la méthode ARACNE sont moins bonnes que celles de IGOIM. Cependant il ne faut pas oublier que les MGG construisent un graphe complet, alors que nous nous restreignons aux 0-1 graphes. En effet, si l'on avait utilisé le coefficient de corrélation partiel d'ordre 1, les résultats n'auraient pu être que meilleurs dans ce cas précis où on est certain que les interactions sont exclusivement linéaires.

La réalité du mécanisme de régulation (impliquant très souvent des "cohortes" de régulateurs parfois antagonistes) laisse penser que la relation de dépendance entre gènes est certainement non-linéaire et même non-monotone, c'est pourquoi nous avons testé les trois méthodes sur les jeux de données non-linéaires de Mendes et al. (Mendes 1993, 1997; Mendes *et al.,* 2003). En tenant compte des résultats obtenus sur le jeu de données *dataset3* (échantillons de 2 variables couplées par une fonction sigmoïdale),

nous avons décidé d'appliquer la méthode IGOIM avec seulement deux estimateurs (qui se sont avérés plus adaptés à cette situation), IM^{MV3} et IM^{ZIP}.

(3,4) *Analyse du type de réseau*. Les performances des trois méthodes sont mauvaises sur le jeu de données non-linéaires, par rapport à celles obtenues avec le jeu linéaire. La méthode IGOIM obtient les meilleurs résultats pour les données générées par un réseau de type "small world", des résultats intermédiaires pour le "scale-free", et finalement les moins bons résultats, pour le réseau aléatoire.

(5) *Analyse du bruit*. Nous avons aussi essayé d'appliquer notre méthode et les deux autres sur des sous-ensembles du jeu de données complet de chaque type de réseau non-linéaire, en introduisant différents types de bruit (loi normale plus ou moins dispersée). Les MGG et ARACNE, ainsi que IGOIM avec l'estimateur IM^{ZIP}, échouent constamment dans ce cas. IGOIM avec l'estimateur IM^{MV3} obtient des résultats qui restent constants pour les différents types de bruits introduits. Les performances de IGOIM avec IM^{MV3} sont moins bonnes sur les ensembles de données incomplets. Par contre, lorsqu'on applique cette méthode sur le même jeu de données incomplet sans bruit et ensuite avec différents types de bruit, elle perd en spécificité (ce à quoi nous nous attendions : plus de faux positifs), mais pas en sensibilité. Au contraire, la sensibilité est parfois légèrement meilleure sur les données bruitées. Nous pensons que ceci confirme notre supposition quant à la mauvaise qualité des données. C'est justement pour cette raison que nous avons préféré ne pas tirer de conclusion définitive de cette première analyse, en conditions non-linéaires.

Cette étude nous a permis d'analyser les différentes méthodes d'inférences basées sur trois estimateurs différents de l'IM avec une autre approche issue de la théorie de l'information et avec les MGG. Pour le cas linéaire, on peut dire que notre méthode est nettement meilleure que les deux autres, et que l'estimateur IM^{MV3} est celui qui doit être gardé par la suite. L'estimateur ZIP3 pourrait être intéressant pour l'analyse de profils d'expression très variables, à condition de faire une étude plus approfondie de ses propriétés.

Étant donné que nous avons utilisé le même jeu de données et les mêmes définitions de la sensibilité et de la spécificité, nous pouvons indirectement comparer notre méthode avec la méthode BANJO (méthode d'inférence de réseaux bayésiens) et avec la méthode NIR (méthode d'inférence basée sur des systèmes d'équations linéaires). Dans leurs études, Bansal et al. (2007) avaient observé que pour les données de perturbations locales, ARACNE était meilleur que BANJO, mais pas aussi bon que NIR. On peut alors dire que IGOIM (avec l'estimateur MV3) est probablement meilleure que BANJO. Bien que IGOIM (MV3) n'obtienne pas des résultats aussi per-

formants que NIR, elle reste bien plus adaptée pour l'inférence des réseaux de gènes pour deux raisons : (1) NIR obtient des bons résultats parce que le modèle utilisé pour la simulations des données correspond à son modèle, et enfin (2) parce que NIR a un gros désavantage, elle a besoin de connaître tous les gènes qui ont été perturbés lors de chaque expérience.

Notre choix de l'IM pour détecter des interactions était initialement justifié par la non-linéarité des relations entre les gènes, et pourtant nous avons testé notre méthode sur un jeu de données linéaires. La première justification de cette démarche est qu'il est très difficile de trouver des données non-linéaires fiables dans la littérature. D'autre part, bien que ses estimateurs sont moins puissants que ceux de la corrélation linéaire, l'IM peut détecter des relations de dépendance plus complexes et plus proches de la réalité biologique, relations qui ne sont pas forcément monotones et qui ne seraient sûrement pas détectées par la corrélation. Une méthode hybride utilisant la corrélation et l'IM, profitant des qualités de chacun, serait probablement une perspective de travail à envisager.

Partie IV

Discussion générale

Ce travail est le premier à s'intéresser à la régulation de la transcription chez *Buchnera APS* d'une façon systémique. Les précédents travaux ayant étudié la régulation transcriptionnelle de *Buchnera* étaient tous expérimentaux, ne touchant souvent qu'une seule des facettes du processus de régulation (Nakabachi and Ishikawa 1997; Baumann *et al.,* 1999; Moran *et al.,* 2003; Wilcox *et al.,* 2003; Moran *et al.,* 2005b; Reymond *et al.,* 2006). Nous avons structuré notre analyse en 4 parties. La première partie visait à dresser l'inventaire de la machinerie transcriptionnelle de *Buchnera APS*, réalisée principalement à l'aide d'études comparatives (*Buchnera* vs. *E. coli*). La deuxième partie a consisté à étudier l'architecture génomique de *Buchnera APS*, *i.e.* l'organisation et l'évolution de sa carte opéronique, l'agencement des fragments synthéniques et non-synthéniques et également les forces d'évolution ayant amené à l'agencement des gènes de *Buchnera* le long de son chromosome. La troisième partie a analysé les propriétés structurelles séquence-dépendantes du chromosome de *Buchnera APS* qui ont ensuite été mises en relation avec le profil de périodicité de l'expression des gènes le long du chromosome. Les résultats obtenus à l'issue de cette approche ascendante (*bottom-up*), nous ont amenés à construire un modèle de réseau de la régulation transcriptionnelle chez *Buchnera APS*. Enfin, la quatrième partie, descendante (*top-down*), a consisté à développer une méthode d'inférence de réseau de régulation à partir de données d'expression, appelée IGOIM. Cette méthode a été validée sur des jeux de données simulées mais n'a finalement pas été utilisée pour inférer le réseau de *Buchnera APS* car actuellement nous ne disposons pas de données d'expression qualitativement et quantitativement suffisantes.

L'inventaire des éléments protéiques de la machinerie de transcription de *Buchnera APS*, réalisé dans la première partie de ce travail compte deux facteurs σ (σ^{70} et σ^{32}), trois facteurs de transcription spécifiques (AlaS, BolA et PepA), quatre facteurs bifonctionnels (DksA, CspC, CspE et CsrA), deux facteurs hypothétiques (YchA et YrbA) et huit toporégulateurs (DnaA, FIS, H-NS, HU_a, IHF_{ab}, YbaB, TopA, Gyr_{ab}), codés par 10 gènes.

Buchnera APS n'a conservé que deux facteurs σ, le facteurs σ constitutif, σ^{70} et le facteur σ du choc thermique périplasmique, σ^{32}, ayant perdu quatre autres (σ^{28}, σ^{38}, σ^{54} et σ^{24}) au cours de son évolution symbiotique avec le puceron. Nous avons effectué une analyse des régulons associés à ces facteurs σ, ainsi qu'une recherche systématique de leurs promoteurs en amont des séquences codantes de *Buchnera APS*.

Des promoteurs σ^{70} ont été trouvés en 5' de la majorité des gènes de *Buchnera APS* (plus de 90 % des gènes et 96% des UT prédites). Les longueurs des 5'UTR prédites ont des distributions similaires entre *Buchnera APS* et *E. coli*. Il semble donc qu'en dépit du fort taux de mutation, *Buchnera APS* a conservé des promoteurs constitutifs (σ^{70}) avec un architecture semblable aux autres bactéries (cf. Résultats, §2.2.5). Une différence significative a été notée entre les distributions des scores de prédictions des promoteurs (qui indirectement reflètent la ressemblance entre ces promoteurs et le modèle utilisé pour la recherche) détectés à l'intérieur des unités de transcription d'*E. coli*, par rapport aux promoteurs prédits en amont de ces unités de transcription. Les premiers sont moins proches du modèle de promoteurs utilisé pour la recherche. La même différence a été retrouvée pour les promoteurs prédits de *Buchnera APS*. Etant donnée cette différence de scores entre les promoteurs intra- et inter - unités de transcription nous avons réalisé a postériori que le score des promoteurs σ^{70} prédits aurait pu être utilisé comme variable discriminante des unités de transcription dans notre outil DisTer développé dans ce travail. Par ailleurs, nous avons noté la présence de promoteurs multiples à l'amont de nombreux gènes de *Buchnera APS*. Nous n'avons pour l'instant pas intégré cette variable mais ceci représente une perspective d'amélioration intéressante, d'autant plus que la présence des promoteurs reste une caractéristique structurelle des unités de transcription et des gènes, tout comme la distance intergénique et la présence de terminateurs de transcription.

Le profil thermodynamique (SIDD) des promoteurs semble également avoir une bonne capacité prédictive des unités de transcription. Comme le score Bprom, le SIDD n'a pas été intégré dans notre prédicteur DisTer (cf. Matériels et Méthodes, §4.1), mais ceci représente encore une perspective d'amélioration intéressante de notre travail.

Mendoza-Vargas et al. (2009) ont montré expérimentalement la présence de plusieurs sites d'initiation de la transcription, et donc la présence de multiples promoteurs par gène. La présence de ces multiples promoteurs σ^{70}, en supposant que la sélection agit de façon à éviter l'apparition des sites non-fonctionnels, pourrait s'expliquer par le fait que la présence des multiples promoteurs reconnus par l'holoenzyme, qui s'associe à ces promoteurs de façon transitoire, permet de maintenir une concentration locale de la polymérase.

Lorsqu'un facteur σ est perdu au cours de l'évolution, on peut s'attendre soit à la perte des gènes cibles correspondant, soit à l'évolution du promoteur de ces gènes vers un autre type de promoteur parmi les facteurs σ conservés. Si les régulons présentent des similarités fonctionnelles (*e.g.* σ^{24} par σ^{32}), on peut également imaginer éventuellement leur intégra-

tion dans une unité de transcription ayant un promoteur reconnu par un des facteurs σ conservés. En effet, les gènes dont les orthologues chez *E. coli* ont un promoteur d'un facteur σ perdu chez *Buchnera APS* possèdent systématiquement un promoteur σ^{70} dans la bactérie symbiotique. Nous avons aussi regardé la proportion des régulons σ d'*E. coli* conservée chez *Buchnera APS*. En dehors du régulons σ^{32} dont 28% des gènes ont été conservés chez *Buchnera* APS, les deux autres régulons dont les gènes ont été le plus conservés chez *Buchnera APS* sont le σ^{28} et le σ^{24}. En ce qui concerne le σ^{38} (le facteur de la phase stationnaire) et le σ^{54} (le facteur du métabolisme de l'azote), nous pouvons supposer que les gènes de ces régulons ainsi que leur facteur σ ont été perdus lors de la réduction drastique du génome de *Buchnera APS*, leur action n'étant plus nécessaire dans l'environnement intracellulaire. Les orthologues des gènes du régulon σ^{24} ont probablement été conservés pour compléter la réponse au choc thermique, leur transcription étant contrôlée par l'action du σ^{70} ou du σ^{32}, des promoteurs σ^{32} ayant été trouvés en amont de certains gènes appartenant au régulons σ^{24}. Par contre, ce qui est plus étonnant est la perte du facteur σ^{28}, le facteur σ flagellaire, alors qu'une proportion importante de ses gènes cibles a été conservée chez *Buchnera APS*. Il est probable que les gènes de ce régulon aient changé de programme de transcription dans la lignée de *Buchnera*, le flagelle n'assurant plus sa fonction de motilité mais plutôt une nouvelle fonction de transport (Kubori *et al.,* 1998; Young *et al.,* 1999). Le facteur σ^{38} (ou σ^{S}) est connu pour être très proche de σ^{70} et un grand nombre de promoteurs sont reconnus et transcrits par les deux facteurs (Paget and Helmann 2003), par conséquent sa perte a pu être compensée grâce au facteur σ^{70}.

Nous avons recensé seulement quatre facteurs de transcription spécifiques (AlaS, BolA, PepA et MetR), dont un, MetR, n'est pas fonctionnel chez *Buchnera APS* (pseudogène). Parmi les trois facteurs spécifiques restants nous comptons deux enzymes, AlaS et PepA. Nous pensons que la conservation de ces protéines est due à la forte pression de conservation, liée à leur fonction métabolique et peut-être en moindre mesure à leur fonction régulatrice. Néanmoins, plusieurs enzymes ont récemment été décrites comme impliquées dans la régulation de la transcription (Bachler *et al.,* 2005; Shen *et al.,* 2006; Commichau *et al.,* 2007; Hullo *et al.,* 2007). Chez les bactéries, les systèmes de régulation à deux composants représentent des senseurs majeurs pour la régulation des activités métaboliques de la cellule (Stock *et al.,* 2000). Ces systèmes étant absents chez *Buchnera*, les enzymes elles-mêmes, directement sensibles à la disponibilité des métabolites, et représentant une partie considérable du génome de la bactérie pourraient jouer ce rôle dual de senseur/régulateur à l'instar de PepA et

AlaS. En effet, ces deux protéines sont capables de se lier à l'ADN chez *E. coli* et pourraient avoir étendu leur activité régulatrice chez *Buchnera APS*. Néanmoins, l'association à l'ADN de ces protéines se fait grâce à un domaine qui ne ressemble à aucun des domaines de liaison déjà décrit et l'étude de leur fonction régulatrice ne peut se faire qu'au niveau expérimental. L'exploration des fonctions régulatrices de PepA et AlaS chez *Buchnera APS* représente une perspective de développement intéressante.

Le dernier facteur spécifique énuméré est BolA. Peu d'interactions BolA – gène cible régulé ont été décrites chez *E. coli*, raison pour laquelle il apparaît comme régulateur spécifique dans le réseau d'*E. coli*. Néanmoins, indirectement il est connu que cette protéine influence la transcription de beaucoup plus de gènes impliqués dans le maintien de la morphologie et de la membrane cellulaire d'*E. coli* (Santos *et al.,* 2002; Freire *et al.,* 2009). BolA devrait donc être considéré plutôt comme un facteur de transcription généraliste. Chez *Buchnera APS*, bien que conservé dans les quatre espèces séquencées, c'est le régulateur qui semble avoir le plus évolué et le domaine hélice-tour-hélice présent chez son orthologue chez *E. coli* n'a pas pu être trouvé avec l'algorithme HTH (Dodd and Egan 1990) chez *Buchnera*. Le gène *bolA* est conservé chez les quatre espèces alors que des pertes de gènes se sont produites dans les différentes lignées de *Buchnera* en amont comme en aval de ce gène. Nous pensons donc que BolA possède une fonction de régulation importante chez *Buchnera* et que son évolution lui a permis de changer de gènes cibles, la régulation des gènes impliqués dans la morphologie de la cellule étant de moindre importance dans la cadre symbiotique. Ici encore l'analyse fonctionnelle expérimentale de ce gène représente une perspective intéressante.

MetR est un régulateur des gènes impliqués dans la phase finale de la biosynthèse de la méthionine, il n'a été conservé que chez *Buchnera Sg*. La perte du gène *metR* semble être récente dans la lignée *Buchnera* puisqu'il est situé dans le fragment *yigL-metE*, les deux gènes étant conservés et adjacents chez *Buchnera Bp*, séparés par un seul gène chez *Buchnera APS* et par deux gènes chez *E. coli*. La fonctionnalité de MetR comme régulateur de MetE a été montré expérimentalement par Moran et al. (2005b). Cette étude a montré que le milieu de vie constitue une force d'évolution puisque MetR n'est fonctionnel que chez *Buchnera Sg*, dont l'hôte (*S. graminum*) se nourrit de la sève phloémienne du blé caractérisée par une composition très riche en cystéine. *Buchnera Sg* a ainsi perdu sa capacité de biosynthèse de la cystéine et est devenu dépendante de son hôte pour cette fourniture (contrairement aux autres *Buchnera*). La contre-partie de cette dépendance semble avoir été la nécessité de conserver une régulation de la biosynthèse de la méthionine à partir de la cystéine alimentaire

(fournie sous forme d'homocystéine) présente vraisemblablement en quantité très variable au cours du temps ou du développement de la plante hôte.

Deux autres protéines ont été considérées comme des facteurs de transcription hypothétiques (YchA et YrbA). Il s'agit de protéines sur lesquelles nous n'avons que très peu d'information, même chez *E. coli*. Leurs séquences protéiques primaires ont le plus faible taux d'identité avec leurs orthologues respectifs chez *E. coli*, parmi les protéines constituant l'inventaire de la machinerie de régulation de la transcription de *Buchnera APS*. On pourrait penser que les gènes codant ces protéines, sous faible pression de sélection, devraient être perdues à terme par pseudogénisation. Néanmoins, les deux protéines sont retrouvées intégralement chez les quatre espèces de *Buchnera* séquencées. On peut donc penser que ces séquences sont en train de diverger afin de s'adapter aux nécessités de la régulation chez *Buchnera*. Il est notable que le gène *yrbA* fait partie des gènes fortement exprimés à vitesse d'évolution rapide (faible taux de GC), alors que *ychA* se situe dans la zone diamétralement opposée, faiblement exprimé et plutôt faiblement évolué (riche en bases G et C, données non présentées). Une analyse du rapport Ka/Ks serait nécessaire pour trancher sur la question de sélection positive de ces gènes. Des analyses de type ChipSeq, visant à détecter les gènes cibles de ces deux régulateurs potentiels constitue également une perspective de travail intéressante sur ces deux gènes candidats.

Enfin, nous avons recensé chez *Buchnera APS* huit toporégulateurs. Ce sont des protéines dont l'activité de régulation de la transcription, s'exerce le plus souvent indirectement, en modifiant et en maintenant la structure des régions promotrices ou plus généralement la topologie de la molécule d'ADN (Schneider *et al.,* 1999; Auner *et al.,* 2003). Parmi ces toporégulateurs, nous comptons 6 protéines NAP, dont les membres les plus classiquement étudiés (FIS, HU, H-NS et IHF) sont aussi considérés comme des facteurs de transcription généraux, ainsi que deux topoisomérases (TopA et l'hétérodimère GyrAGyrB).

Les quatre NAP : HU, IHF, FIS et H-NS participent au maintien de la structure du nucléoïde bactérien en s'associant à l'ADN principalement de façon non-spécifique. IHF, FIS et H-NS possèdent également des sites de fixation spécifiques leur conférant des rôles de régulation spécifique.

L'abondance relative des NAP dépend étroitement de l'état cellulaire (de développement ou de conditions environnementales) et définit l'état de condensation et de surenroulement du chromosome, par exemple en phase stationnaire le chromosome est plus relâché, alors que durant la phase exponentielle, il est plus surenroulé.

Enfin, malgré leur rôle fondamental pour le maintien de la structure du chromosome, les NAP ne sont pas individuellement essentielles. Cette propriété est sans doute liée aux homologies entre ces protéines (HU et IHF sont des paralogues) qui leur permet de se complémenter.

Bien que chez *E. coli* HU agit sous forme d'hétérodimère codé par les deux gènes *hupA* et *hupB*, chez la plupart des bactéries cette protéine est présente sous forme homodimérique : HU-α2. Chez les *Buchnera* séquencées également, on note que le gène *hupA* est conservé dans trois des quatre espèces (*APS*, *Sg* et *Bp*), alors que le gène *hupB* n'a été retrouvé dans aucune espèce. IHF devient important dans la mesure où il substitue certaines des fonctions de HU dans les mutants Δ*hupA* chez *E. coli*. Par ailleurs les doubles mutants Δ*hupA*Δ*ihfAB* sont difficiles à obtenir et leur phénotype est sévèrement compromis (Kano and Imamoto 1990). Les gènes *himA* et *himD* codant pour l'hétérodimère sont conservés chez les trois des quatre espèces *Buchnera* séquencées (*APS*, *Sg* et *Cc*). Rappelons que *Buchnera Cc* est la seule espèce n'ayant pas conservé le gène *hupA*. On observe donc que les doubles mutants difficiles à construire ne sont pas retrouvés *in vivo*. Chez ces mêmes espèces, nous retrouvons le gène *fis*.

Enfin, le gène *hns* a été conservé seulement chez l'espèce *Buchnera APS*, parmi les 4 espèces de *Buchnera*. Il est d'ailleurs intriguant que *Buchnera Sg* n'ait pas conservé H-NS alors qu'elle est très proche de *Buchnera APS*. Chez *Buchnera Bp* une large région intergénique (597 pb) entre *cls* et *ribA* semble attester la présence ancienne du gène *hns* chez *Buchnera APS*. A l'inverse l'espace intergénique chez *Buchnera Sg* est très court (184pb). Hershberg al. (2006) expliquent que l'absence d'un répresseur a plus d'impact que l'absence d'un activateur. Comment expliquer alors la perte d'un répresseur universel, très important pour la compaction par ailleurs ? Nous savons que H-NS se lie préférentiellement à des séquence AT-riches courbées, vu la composition des génomes de *Buchnera*, on pourrait penser que H-NS pourrait avoir un effet trop inhibiteur, qui ne pourrait pas être rétrocontrôlé si les protéines comme FIS ou HU sont également absentes du génome. Mais cette proposition reste entièrement hypothétique.

Buchnera APS est la seule des quatre espèces de *Buchnera* ayant conservé l'ensemble complet des NAP les plus caractérisées chez *E. coli* (FIS, HU, H-NS et IHF). Les séquences primaires de NAP ont chez *Buchnera APS* un fort taux de conservation et les sites essentiels à leur fonction de NAP, identifiés par mutation ponctuelle chez *E. coli*, sont conservés.

Buchnera Bp n'a conservé que HU et perdu les autres NAP. Chaque gène des NAP perdu est situé entre des gènes conservés et ayant la même disposition chez les trois espèces (*APS*, *Sg* et *Bp*). Si on suivait le modèle de réduction du génome proposé par Moran (2001), on peut suppo-

ser qu'il s'agit de gènes éliminés ponctuellement, car il n'y a plus de pression de sélection sur ces locus. On peut alors penser que la compaction et l'inhibition ne soient plus essentielles dans le cas de *Buchnera Bp*, idée suggérée également par la faible conservation de H-NS.

Les double mutants Δ*hupA*Δ*mukB* de *E. coli* ne sont pas viables (Jaffe *et al.,* 1997). Sachant que *mukB* est absent dans toutes les espèces de *Buchnera*, il semblerait que *hupA* soit un gène essentiel chez la bactérie symbiotique. Nous avons néanmoins constaté que *Buchnera Cc* n'a pas ce gène, en même temps cette espèce a conservé les deux gènes codant pour l'hétérodimère IHF. Les mutants Δ*hupAB*Δ*ihfAB* et Δ*hupAB*Δ*hns* ont des phénotypes synthétiques et les mutants simultanés de HU, IHF et H-NS ne peuvent pas être obtenus (Kano and Imamoto 1990; Yasuzawa *et al.,* 1992). Tous ces résultats confirment le caractère essentiel des NAP (ou plutôt d'un sous-ensemble de NAP), et tout particulièrement de HU. Même lorsque HU est perdue, son paralogue est présent et récupère probablement les fonctions de HU.

DnaA et YbaB sont les seules NAP conservées chez les quatre espèces *Buchnera*. DnaA est connue pour sa fonction essentielle dans la réplication du chromosome. YbaB, est une NAP découverte plus récente, qui semble très importante à la viabilité cellulaire mais dont la fonction n'a pas encore été révélée (Dillon and Dorman 2010). Comprendre le rôle de cette protéine dans la structuration du nucléoïde chez *Buchnera* représente une perspective intéressante pour la suite de ce travail.

Aux six toporégulateurs que l'on vient de décrire, s'ajoutent deux topoisomérases, la topoisomérase I permettant de relâcher les supertours négatifs et la gyrase qui permet d'en introduire. Ces deux topoisomérases constituent l'ensemble minimal, permettant d'ajuster le niveau de surenroulement de l'ADN. Chez *E. coli*, le niveau de surenroulement est régulé par le concours de 4 topoisomérases, dont deux ne sont pas trouvées dans les quatre espèces de *Buchnera*. Il s'agit de topoisomérases intervenant dans la décaténation et surtout dans la séparation des chromosomes à la fin de la réplication. Ainsi leur perte dans la lignée de *Buchnera* pourrait être en partie à l'origine de sa forte polyploïdie. Plus étonnant est le fait que le gène codant la topoisomérase I n'est pas conservé chez toutes les espèces de *Buchnera* séquencées, il est présent seulement chez *Buchnera APS* et *Sg*. Or, la topoisomérase I est bien la seule capable de relaxer l'ADN. Néanmoins Gil et al. (2004) la qualifient comme non indispensable, en argumentant que probablement la gyrase peut assurer toutes les fonction des topoisomérases, ayant des mécanismes de fonctionnement similaires (liaison à l'ADN, et coupure de un ou deux brins de l'ADN) (Berger *et al.,* 1998). On pourrait aussi supposer que les chromosomes de *Buchnera Bp* et de *Buch-*

nera Cc ont un moindre besoin de relaxation, ces deux espèces étant plus riches en bases A et T que *Buchnera APS* et *Sg* et les séquences plus riches en bases A et T semblent être globalement plus relâchées.

Une autre question qui se pose relativement à la faible conservation du nombre de toporégulateurs dans les espèces de *Buchnera* est la structuration du chromosome en domaines topologiques. Ces domaines, de 10kb en moyenne, représentent un premier niveau de structuration du chromosome d'*E. coli*. Les principaux constituants des barrières topologiques entre ces domaines sont l'ARN polymérase, la gyrase, H-NS, FIS, Pgm, DksA et TktA (Hardy and Cozzarelli 2005). De plus Maurer et al. (2009) ont montré que l'interaction entre l'ADN d'un côté et HU, H-NS et IHF de l'autre, se stabilise sous forme de structures semi-périodiques. D'autres acteurs comme les topoisomérases et le complexe SMC se rajoutent pour mettre en place des structures de plus haut niveau (*e.g.* les macrodomaines). Chez les *Buchnera*, seule la gyrase, la polymérase et l'enzyme TktA sont systématiquement présentes, alors que par rapport à la haute densité d'ADN dans les cellules de cette bactérie, on pourrait s'attendre à un plus grand besoin de structuration et de compaction.

On peut constater qu'à l'exception des deux enzymes à fonction régulatrice potentielle, seuls des facteurs de régulation de la transcription généraux ont été conservés chez *Buchnera*. Ceci peut s'expliquer en partie par le modèle de réduction de son génome. La première étape de réduction s'est faite par l'élimination de grands fragments de génomes, or il est connu que les facteurs de transcription sont localisés au voisinage de leurs cibles de régulation (Hershberg *et al.,* 2005; Hershberg and Margalit 2006). Une autre partie de facteurs de régulation spécifiques a pu être perdue lors de la deuxième étape de réduction du génome de *Buchnera*, quand des plus petits fragments ont été éliminés (notamment par pseudogénisation) à cause d'une pression de conservation plus relâchée, de la perte de certaines cibles et d'un changement de la demande de régulation lié à l'environnement intracellulaire. Sachant le fort impact que l'organisation de la carte opéronique (*i.e.* la structure des unités de transcription) impose sur la régulation de la transcription des gènes chez les bactéries, nous nous y sommes intéressés chez *Buchnera* dans la suite de notre travail, c'est ce qui a constitué la deuxième partie de ce travail de thèse.

Grâce à notre nouvelle méthode de prédiction des unités de transcription, DisTer, nous avons prédit une nouvelle carte opéronique de *Buchnera APS*. Cette carte est constituée de 288 unités de transcription. La nouveauté de notre méthode consiste en l'utilisation des prédictions des terminateurs de transcription. Plus généralement, notre prédiction se base sur des critères entièrement structuraux (distance intergénique et termina-

teurs de transcription), alors que les autres prédictions existantes ont utilisé en plus de la distance intergénique des paramètres comme la similarité fonctionnelle, l'activité métabolique des produits des gènes, ou l'orthologie. Même si, le choix de la similarité fonctionnelle comme critère de prédiction des unités de transcription paraît logique, notre travail montre qu'une telle approche chez *Buchnera APS* ne peut pas détecter les unités de transcription les plus récentes, formées de gènes sans relation fonctionnelle, apparues suite aux réarrangements génomiques et/ou aux mutations des éléments de régulation aux frontières des unités de transcription. En effet, les unités de transcription rassemblant des gènes non-fonctionnellement liés disparaissent très vite dans les bactéries libres (Price *et al.,* 2006), chez *Buchnera APS*, ces unités de transcription sont maintenues à cause de l'absence de la machine de recombinaison dont les éléments ont été perdus durant la réduction drastique de son génome (Tamas *et al.,* 2002; Silva *et al.,* 2003).

La carte opéronique de *Buchnera APS* que nous avons prédite est plus compacte que les trois autres que nous avons pu trouver dans la littérature. La carte opéronique de *Buchnera APS* semble aussi plus compacte que celle *d'E. coli*. En moyenne, chaque unités de transcription de *Buchnera* contient 2.12 gènes, contre 1.63 chez *E. coli* qui a notamment une proportion d'UT monocistroniques plus importante. Enfin, notre analyse de validation expérimentale par RT-PCR a montré que pour toutes les paires prédites comme appartenant à la même unités de transcription (STU) et testées expérimentalement un ARN polycistronique a été amplifié. Par contre, seules deux paires parmi les 10 paires prédites comme appartenant à des unités de transcription différentes (paires UTD) et testées expérimentalement ont été validées comme telles, pour les huit autres un ARNm polycistronique a été amplifié. Ce résultat suggère à notre avis que la vraie carte opéronique de *Buchnera APS* est encore plus compacte que nous l'avons prédit avec des unités de transcription polycistroniques encore plus hétérogènes. La mauvaise prédiction de DisTer (mais aussi des autres méthodes de prédiction) sur les 8 paires prédites UTD mais amplifiées par RT-PCR, peut être expliquée par le fait que DisTer ainsi que les trois autres méthodes ont été entraînées sur le génome *d'E. coli* ou ont utilisé les propriétés communes des opérons bactériens les mieux décrits dans les bases de données. Ces derniers pourraient ne pas refléter toutes les caractéristiques du génome de *Buchnera APS* (biais AT, perte de la capacité de recombinaison, mode de vie intracellulaire).

Notre carte opéronique est bien corrélée avec les promoteurs σ^{70} prédits, ainsi qu'avec les données d'expression (*i.e.*, l'expression des gènes se trouvant dans une même unités de transcription est moins variable

qu'entre des gènes appartenant à des unités de transcription distinctes). Une partie de la corrélation entre les niveaux d'expression des gènes et les unités de transcription peut s'expliquer par leur proximité sur le chromosome (cf. Introduction). Néanmoins, la manière dont nous avons permuté les données d'expression sans détruire la corrélation de l'expression entre les gènes voisins confirme que les structures opéroniques sont significativement impliquées dans la corrélation des niveaux de transcription des gènes colocalisés.

Le génome de *Buchnera* peut être considéré comme un sous-ensemble du génome *d'E. coli,* nous avons ainsi pu comparer leurs cartes opéroniques afin de révéler la dynamique des unités de transcription dans ces deux lignées. Nous avons montré que des changements sont apparus dans les cartes suite aux réarrangements génomiques (recombination, inversion, translocation), mais aussi à l'intérieur des fragments synthéniques, suite aux délétions ponctuelles de gènes. Des évènements de délétion/substitution/insertion de petits fragments ont également entraîné la réorganisation des frontières entre les unités de transcription. Les modifications du contenu des unités de transcription (essentiellement liées aux réarrangements des frontières) ont forcement un impact sur la régulation de l'expression des gènes, ne serait-ce qu'à cause du fait que certains ensembles de gènes co-transcrits changent, alors que les modifications de l'ordre des unités de transcription (liées aux grands réarrangements) sur la carte peuvent avoir une influence plus faible sur le niveau de la transcription des gènes (liée par exemple à l'insertion des unités de transcription dans des régions du chromosome plus ou moins propices thermodynamiquement à la transcription).

La carte opéronique de *Buchnera APS* a été modifiée par la profonde réorganisation génomique qui a eu lieu dans les stades précoces de la symbiose et plus de 45% des unités de transcription ont été affectées par ce processus. Les remaniements locaux de la carte de *Buchnera APS* semblent plus récents et correspondent notamment à l'évolution des séquences intergéniques. Ces régions du génome sont connues pour évoluer plus rapidement que les séquences codantes. Elles contiennent des éléments structuraux comme par exemple les promoteurs et les terminateurs de transcription. Ainsi, l'évolution des ces séquences intergéniques agit localement causant des fluctuation de frontières entre les unités de transcription avoisinantes (*e.g.*, les unités de transcription fragmentées).

Ces deux processus (les réarrangements génomiques et l'évolution locale des séquences) ne sont pas indépendants, et l'évolution de la carte opéronique est généralement le résultat de leur influence conjointe. Chez *Buchnera APS* par contre, des grands réarrangements ne peuvent plus avoir

lieu, la bactérie ayant perdu la capacité de recombinaison. C'est donc l'évolution locale qui prend le dessus, et ainsi, la carte évolue inexorablement vers une compaction de plus en plus forte, en assemblant notamment des gènes non-fonctionnellement liés.

Moran et al. (2009) ont montré récemment que chez *Buchnera APS* l'évolution de la séquence d'ADN n'est pas symétrique. Celle-ci semble n'évoluer que vers un raccourcissement, puisque les seules insertions qu'ils ont pu observer ne représentent que quelques paires de bases et sont principalement dues aux glissements de la polymérase alors que les délétions étaient très nombreuses. Ainsi, les résultats de cette analyse sont en accord avec nos propres résultats puisqu'ils suggèrent que le génome de *Buchnera APS* ne puisse évoluer que vers une carte opéronique plus polycistronique, dans la mesure où les raccourcissements graduels des régions intergéniques sont susceptibles de désintégrer des promoteurs et/ou des terminateurs de transcription et ainsi d'induire la création de nouvelles unités de transcription polycistroniques ou de rallonger des unités de transcription polycistroniques existantes. La désintégration des unités de transcription polycistroniques n'est pas impossible chez *Buchnera APS*, par exemple, une délétion ponctuelle suivie d'une pseudogénisation. Néanmoins cette possibilité semble moins probable dans un génome que nous supposons avoir atteint une certaine stabilité (Tamas *et al.,* 2002). Le nombre d'UT fusionnées que nous avons recensé chez *Buchnera APS* (64) et le nombre de gènes qu'elles rassemblent (231 soit plus de 30% du génome) sont cohérents avec ces hypothèses.

Notre analyse des distances intergéniques (inter- et intra- unités de transcription) chez *Buchnera APS* nous a permis de révéler un autre aspect de sa carte opéronique. Premièrement, nous avons trouvé que les séquences codantes de *Buchnera* se superposent moins que chez *E. coli*. La superposition des séquences codantes est généralement trouvée dans les opérons anciens (Price *et al.,* 2006). Il semble donc que *Buchnera APS* ait perdu certains des opérons anciens et que le contexte de son évolution ne lui permette pas d'évoluer pour faire apparaître des superpositions dans ses nouveaux opérons.

Deuxièmement, les distances intergéniques de *Buchnera APS* sont plus courtes en moyenne que celles d'*E. coli.* Ceci est dû entièrement à la différence significative entre les distributions des distances entre les gènes ayant la même direction de transcription, car les distances entre les gènes adjacents ayant des directions de transcription opposées sont similaires entre les deux bactéries. Il a été observé chez *E. coli* une différence significative entre la distribution des distances intergéniques convergentes et la distribution des distances intergéniques divergentes. Nous avons trouvé

cette même différence significative entre ces deux distributions chez *Buchnera*. L'hypothèse expliquant cette observation, avancée par Rogozin et al. (2002), est que les régions intergéniques divergentes constituent le support physique de deux promoteurs nécessitant plus d'espace (plus de paires de bases) que les régions intergéniques convergentes, contenant seulement des terminateurs de transcription. Cet aspect confirme la présence des promoteurs sous une pression de sélection chez *Buchnera APS*. Les distributions des distances intergéniques divergentes (convergentes) semblent similaires entre *E. coli* et *Buchnera APS*. Dans le cadre de la compaction du génome de *Buchnera APS*, il avait été admis que les régions intergéniques seraient devenues plus courtes (Moran and Mira 2001), notre travail montre que ce n'est pas le cas, du moins pour les régions intergéniques convergentes et divergentes.

Nous avons enfin cherché l'impact de la régulation des gènes sur la dynamique de la carte opéronique de *Buchnera APS*. Dans ce but, nous avons étudié chez *E. coli* les unités de transcription contenant au moins un gène ayant un orthologue chez *Buchnera APS*. Nous avons tout d'abord montré une corrélation entre la régulation des gènes chez *E. coli* et la présence de leurs orthologue chez *Buchnera APS* si un gène est régulé chez *E. coli*, il a plus de chances d'avoir été conservé chez *Buchnera APS* que s'il n'est pas régulé. Ce premier résultat indique que les gènes régulés sont globalement plus importants pour la viabilité que les gènes non-régulés. Néanmoins, nos résultats sur la réorganisation des unités de transcription suggèrent que l'assemblage des gènes de *Buchnera APS* en unités de transcription n'a pas été contraint par la régulation spécifique des gènes chez son ancêtre (sous l'hypothèse contraignante que la régulation de l'expression des gènes chez l'ancêtre était semblable à celle d*'E. coli* actuellement). En effet, des proportions similaires des unités de transcription contenant des gènes dont l'orthologue est spécifiquement régulé chez *E. coli* ont été trouvé dans chacune des cinq classes d'unités de transcription que nous avons définies dans la partie Résultats, §2.2.1.2 (Annexe, A2-3). Deux explications à cette observation sont possibles : soit de nombreuses régulations spécifiques trouvées chez *E. coli* ne se sont mises en place qu'après la divergence des deux lignées ; soit certaines des régulations spécifiques *d'E. coli* sont bien ancestrales, mais ne correspondaient pas aux besoins de *Buchnera APS* et ainsi n'ont pas constitué une pression de conservation suffisante. Néanmoins, la régulation semble avoir été une des forces d'évolution de l'organisation des gènes sur le chromosome de *Buchnera APS*, puisque des nouveaux transcriptons se sont formés suite aux réarrangements génomiques (cf. Résultats, §1.2). Ainsi, c'est donc la deuxième hypothèse qui semble plus probable.

Cette dernière analyse indique que certains des opérons de *Buchnera APS* semblent « accidentels » et passivement maintenus dans le génome de la bactérie. Cette observation est importante pour comprendre la pression de sélection s'exerçant sur la régulation de l'expression des gènes chez *Buchnera APS*. Le maintien d'une co-transcription même sans un lien avec la fonction, ni avec la régulation spécifique suggère que chez *Buchnera APS* c'est la « simple » production de la transcription et non sa régulation finement ajustée, qui est importante.

Après avoir fait l'inventaire des régulateurs protéiques chez *Buchnera APS* et étudié l'organisation et l'évolution de sa carte opéronique, nous nous sommes intéressés à l'organisation du chromosome qui joue également un rôle important dans la régulation de l'expression des gènes (Jeong *et al.,* 2004; Peter *et al.,* 2004; Blot *et al.,* 2006). L'ADN n'est pas seulement un support inerte de l'information, mais c'est aussi un élément actif, qui par ses propriétés chimiques, physiques et structurelles participe aux processus de transcription. C'est l'analyse de ces propriétés physiques et une première tentative de constructions d'un réseau de régulation de *Buchnera APS* qui a constitué notre troisième partie de ce travail de thèse.

Nous avons analysé chez *Buchnera APS* comparativement à *E. coli* par rapport à quatre propriétés structurelles : la courbure intrinsèque, l'énergie d'empilement des bases, l'angle de torsion et le SIDD. L'analyse a été faite de trois façon différentes : globalement, la totalité du chromosome étant considéré ; par région géniques, la séquence du chromosome étant séparée en quatre groupes de séquences (régions intergéniques tandem, régions intergéniques convergentes, régions intergéniques divergentes et régions codantes) ; et par région promotrice (en analysant globalement les propriétés structurelles séquence-dépendantes, des 150pb en amont du site d'initiation de la traduction des séquences codantes).

La comparaison globale de la courbure, de l'énergie d'empilement et de l'angle de torsion, des chromosomes *d'E. coli* et de *Buchnera APS* reflète principalement la différence de composition en bases des deux génomes. Le génome d'*E. coli*, le plus GC-riche, est intrinsèquement moins courbé, plus stable et moins flexible. Néanmoins la composition des génomes ne suffit pas à elle seule pour caractériser les distributions des variables observées. Lorsque nous construisons un premier génome « aléatoire » en permutant les bases du génome de *Buchnera APS*, un génome ayant donc la même taille et la même composition que le génome de *Buchnera APS*, ces distributions ne sont pas toujours centrées autour de la même valeur que la distribution de *Buchnera APS*. De plus ces distributions sont plus étroites (intervalle des valeurs prises plus petit). Nous avons alors construit un deuxième type de génome « aléatoire » de *Buchnera APS*, en

permutant les bases à l'intérieur de chaque région intergénique et de chaque région codante, c'est donc un génome ayant la même taille et la même composition, mais également la même composition par région génique et la même alternance de régions géniques et non-génique que le génome de *Buchnera APS*. La distribution de ce deuxième modèle aléatoire se rapproche considérablement de la distribution de *Buchnera APS*. Les distributions de l'énergie d'empilement des bases et celle de l'angle de torsion du deuxième modèle aléatoire sont plus proches de la vraie distribution de *Buchnera APS* que de la distribution du premier modèle aléatoire, ce qui signifie que l'alternance des régions géniques ainsi que leur composition interviennent dans la forme de la distribution.

Dans le cadre de cette discussion, l'analyse comparative a été étendue aux trois autres espèces de *Buchnera* (données non présentées). Les mêmes propriétés sont alors globalement trouvées avec *Buchnera Sg*, *Buchnera Bp* et *Buchnera Cc,* le positionnement relatif des distributions reflétant majoritairement leurs compositions en bases.

A l'inverse, les distributions du SIDD ne reflètent pas la composition des génomes ou dans une moindre mesure (*i.e.* le positionnement des distributions le long de l'axe du SIDD n'est pas situé dans l'ordre des compositions en bases A et T des génomes), dans ce cas, les distributions de *Buchnera* sont similaires et diffèrent d'*E. coli*, qui a une distribution plus étalée.

On peut supposer que la courbure intrinsèque et la flexibilité plus importante des chromosomes de *Buchnera* (par rapport à *E. coli*) pourrait avoir un rôle dans leur compaction en association avec les NAP, même si celles-ci, et notamment H-NS qui se lie aux séquences les plus courbées, ne sont pas systématiquement présentes chez toutes les *Buchnera* (H-NS n'est présente que chez *Buchnera APS* par exemple).

Comme nous l'avons vu lors des comparaisons précédentes, l'alternance des régions géniques et intergéniques joue un rôle important dans la dynamique du chromosome de *Buchnera APS* vue à travers les distributions des variables de courbure et d'élasticité mesurées. Nous avons voulu tester si ces différents paramètres permettent de distinguer les différents types de régions géniques (régions intergéniques tandem, convergentes, divergentes et les régions codantes). A cette fin, nous avons construit les distributions des mêmes paramètres pour chaque type de région génique chez *Buchnera APS*. Les aspects des distributions des différents types de régions intergéniques reflètent le type de motifs structurels contenus dans ces régions. Le SIDD est la seule propriété ayant des distributions différentes en fonction de la région génique. La distribution du SIDD des régions divergentes, qui doivent contenir au moins un promoteur, est carac-

térisée par des valeurs faibles du SIDD (ADN s'ouvrant plus facilement). Ces régions sont donc plus instables et donc plus favorables à l'initiation de la transcription. La distribution des régions convergentes ne contenant pas de promoteurs mais exclusivement des régions terminatrices, a un comportement opposé, étant caractérisée par des valeurs de SIDD plutôt fortes (ADN très stable). Les régions tandem qui sont susceptibles de contenir des promoteurs et des régions terminatrices ont des aspects intermédiaires, bimodales, entre les deux distributions des régions convergentes et divergentes. La distribution des régions tandem inter-unités de transcription a un aspect bi-modal, ayant des valeurs faibles regroupées autour de 3 (unité arbitraire) et des valeurs fortes regroupées autour de 9 (unité arbitraire). Les régions tandem intra-unités de transcription ont également une distribution bi-modale, avec plus de valeurs faibles, regroupées autour de 2 (ADN ayant une plus forte capacité à s'ouvrir). Les régions tandem intra-unités de transcription ont donc plus de régions stables (valeurs fortes du SIDD) que des régions instables (valeurs faibles du SIDD). Ce résultat peut être interprété (1) par le fait que ces régions sont transcrites et donc plus stables que les autres régions intergéniques, ou (2) que les régions tandem intra-unités de transcription contiennent potentiellement moins de promoteurs (les promoteurs alternatifs des unités de transcription) et ceux-ci sont plus stables.

Wang et al. (2008) ont trouvé que chez *E. coli* les régions promotrices des gènes associés aux fonctions « régulation » et « protéines membranaires » ont des valeurs SIDD significativement plus faibles que la moyenne globale. Nous avons cherché une telle association chez *Buchnera APS* entre le type de région promotrice (instable – faible SIDD/stable - fort SIDD) et la classe fonctionnelle (en utilisant la classification de Riley et al. (1998)) ou la classe métabolique du gène (cf. Matériels et méthodes, §1.2). Le résultat de Wang et al. (2008) a été retrouvé chez *Buchnera APS* puisque nous avons montré que seuls les régulateurs et les transporteurs (protéines membranaires) sont significativement associés avec des régions promotrices instables chez *Buchnera APs*. Les auteurs comme nous-mêmes, n'ont pas pu proposer d'interprétation biologique de ce résultat.

Il est connu que le SIDD varie localement en fonction du niveau de surenroulement de l'ADN (Benham 1996; Benham and Bi 2004; Wang and Benham 2008). Nous avons pensé que les gènes dont la transcription est sensible aux variations du niveau du surenroulement devraient avoir des régions promotrices à SIDD plus faibles que le reste des gènes. Nous avons d'abord testé cette hypothèse chez *E. coli* en utilisant la liste des gènes sensibles aux changements du niveau de surenroulement de Peter et al. (2004), puis ensuite chez *Buchnera APS* en utilisant la liste des orthologues des gènes d'*E. coli*. L'utilisation de la liste des orthologues chez *Buchnera APS*

semble légitime, car les gènes à transcription sensible au niveau de surenroulement, en plus d'avoir des promoteurs à composition spécifique (plus riches ou plus pauvres en bases G et C que la moyenne (Cashel *et al.,* 1996)) sont également régulés chez *E. coli* par l'action des NAP et/ou des topoisomérases, qui comme nous l'avons vu plus haut (cf. Résultats, §2.1.2.4) sont conservés chez *Buchnera APS*. Aucune association significative n'a pourtant été trouvée ni chez *E. coli*, ni chez *Buchnera APS*. Une autre liste de gènes de *Buchnera APS*, potentiellement régulés par les variations du niveau de surenroulement de l'ADN, pouvait être la liste de gènes différentiellement exprimés dans l'expérience de Bermignham et al. (2009). Ici c'est l'hypothèse du taux de croissance de *Buchnera* qui rentre en jeu. Chez la bactérie libre, l'équilibre relatif des NAP varie en fonction de la croissance avec une différence marquée entre la phase exponentielle et la phase stationnaire correspondant à des niveaux de surenroulement de l'ADN très différents (Dorman *et al.,* 1988). Chez *Buchnera*, incultivable en dehors des pucerons, il est difficile d'appréhender le taux de croissance de la bactérie. Néanmoins, il a été montré que dans les stades embryonnaires les plus jeunes, les *Buchnera* sont en phase de croissance « rapide » (la croissance exponentielle n'ayant pas été démontrée à proprement parler), contrairement aux *Buchnera* des bactériocytes maternels, qui semblent plutôt en croissance très lente voire stationnaire (Bermingham *et al.,* 2009). Des données d'expression ont été ainsi acquises et des listes de gènes différentiellement exprimés entre ces deux tissus ont été établies. Dans ce travail, nous avons recherché si les gènes différentiels étaient caractérisés par des promoteurs différents des autres (car plus ou moins sensibles aux NAP et aux changements du niveau de surenroulement). Les régions promotrices de ces gènes n'ont pas de valeurs de SIDD significativement différentes des autres gènes de *Buchnera APS*. La liste de gènes de *Buchnera APS* établie par Bermigham et al. (2009), n'est pas significativement enrichie en gènes orthologues des gènes sensibles au changement du surenroulement chez *E. coli* ((Peter *et al.,* 2004)). Il est très probable que l'ensemble de gènes ayant des promoteurs sensibles au changement du surenroulement ait changé de composition chez *Buchnera APS*, car cette sensibilité est en partie donnée par la composition du promoteur des gènes, qui chez *Buchnera APS* a été modifiée par le fort biais en bases A et T. Les nombreux réarrangements génomiques ont pu entraîner le changement du contexte génomique de ces promoteurs, et donc de leur sensibilité aux changements du niveau de surenroulement.

Buchnera APS a essentiellement conservé des gènes métaboliques dans son génome, le métabolisme étant sa fonction principale dans le cadre de l'association symbiotique. Par conséquent, le système de régulation de la

bactérie devrait être dédié à la régulation de ces gènes. Nous avons cherché à savoir si les promoteurs en amont de gènes métaboliques avaient des SIDD différents de ceux des autres gènes chez *Buchnera APS* et chez *E. coli*. Nous avons montré que les gènes cataboliques de *Buchnera APS* sont significativement associés aves des régions promotrices stables (SIDD fort) laissant supposer une sensibilité différentielle aux variations de surenroulement, alors que chez *E. coli* aucune association significative entre la classe métabolique des gènes et le type de région promotrice n'est constatée. Nous observons donc une différence entre *Buchnera APS* et *E. coli* sur la classe des gènes cataboliques, qui est la classe de gènes métaboliques la plus régulées chez *E. coli* mais également la moins conservée chez *Buchnera APS*.

Le but de nos analyses était de construire un modèle de régulation de la transcription des gènes chez *Buchnera APS*. Dans un premier temps, un réseau de régulation a été construit par orthologie à celui d'*E. coli*, en utilisant l'inventaire de la machinerie de transcription de *Buchnera APS*. Ce réseau a été ensuite étendu grâce aux UT de BAp que nous avons prédites. A cette étape notre réseau comptait 194 interactions entre 140 gènes. Ni les recherches des sites de fixation des facteurs généraux de la transcription (rendue inexploitable par le biais en bases A et T des régions intergéniques), ni les propriétés structurelles des régions promotrices (trop hypothétiques et trop peu corrélées aux données fonctionnelles ou d'expression) ne nous ont permis d'enrichir le réseau avec de nouvelles cibles de régulation.

Sachant que ces interactions étaient principalement dirigées par des toporégulateurs, nous avons tenté de valider notre réseau sur les données de Bermingham et al. (2009) censées impliquer un niveau de surenroulement variable du chromosome et/ou des rapports variables en composition en NAP (Dorman *et al.,* 1988). Néanmoins, les gènes différentiellement exprimés dans cette expérience ne sont pas représentés de façon significative dans le réseau de *Buchnera APS* que nous avons construit.

Compte-tenu des études faîtes sur les sites de fixation des facteurs de la transcription montrant que ce sont des structures très labiles entre des espèces très proches, voire même entre des souches de la même espèce (Rodionov *et al.,* 2004), le réseau de *Buchnera APS*, que nous avions construit par orthologie avec *E. coli* contient surement beaucoup d'interactions hypothétiques. La partie moins hypothétique de ce réseau représente les acteurs de la régulation qui sont dans ce cas principalement les toporégulateurs. Cet aspect du réseau de *Buchnera APS* combiné aux propriétés structurelles globalement différentes de son chromosome relativement à *E. coli* ainsi que ces plus longues unités de transcription, suggèrent que chez

Buchnera APS la régulation se fait par une machinerie généraliste agissant par le biais de la conformation du chromosome pour coordonner l'expression de ses gènes entre quelques régimes d'expression peu nombreux. Plusieurs observations semblent cohérentes avec cette hypothèse. Les rapports d'expression différentielle des gènes de *Buchnera APS* montrent une faible variation en amplitude lors de changements de conditions de vie du puceron (stress nutritionnel, stades de développement). Ce « lissage » de la régulation de l'expression des gènes chez *Buchnera* est cohérent avec un réseau de régulation gouverné par la topologie du chromosome et par les toporégulateurs, car il a été montré que les protéines associées au nucléoïde et le surenroulement du chromosome induisent des changements continus du niveau de transcription des gènes, alors que les facteurs de transcription spécifiques provoquent plutôt des changements binaires de type présence/absence de l'ARNm (Blot *et al.,* 2006; Marr *et al.,* 2008; Balleza *et al.,* 2009). Par ailleurs, le génome très riche en bases A et T de *Buchnera APS* (et donc très homogène) semble peu propice au maintien de sites de fixation très spécifiques, par contre, cette richesse en bases A et T semble plus favorable à des sites dégénérés correspondant à des facteurs de transcription généralistes et pléïotropiques, comme les NAP. Il se peut qu'à cause de la présence du grand nombre des ces sites de fixation peu spécifiques, les facteurs de transcription se lient plus souvent à l'ADN et de façon moins forte. Ainsi, leur action pourrait être ralentie par rapport à celle des bactéries libres dont la composition du génome est plus riche en bases G et C.

Enfin, le profil périodique de l'expression des gènes est sans doute une conséquence de la structuration en domaines et macro-domaines du chromosome de *Buchnera APS*, comme il a déjà été démontré chez les bactéries à forme de vie libre (Jeong *et al.,* 2004; Carpentier *et al.,* 2005). Sachant que cette structuration est due en grande partie à l'interaction des NAP avec l'ADN et que les NAP sont également des régulateurs importants dans la cellule, nous pouvons supposer l'existence d'un lien indirect entre la périodicité du profil de transcription et la régulation de l'expression. Néanmoins, nous n'avons pas réussi à relier cette périodicité avec le profil des sites de liaison des NAP, à cause de la dégénérescnece des sites de fixation des NAP et de la composition très riche en bases A et T du génome de *Buchnera APS*. Nous avons cherché à approfondir cette analyse du profil de transcription notamment en travaillant à pas constant sur le chromosome (et non pas à l'échelle du gène comme dans l'étude initiale), ainsi qu'en enrichissant l'étude par l'analyse spectrale des propriétés structurelles du chromosome. Ainsi, nous avons pu montrer le comportement périodique de l'expression des gènes, de la courbure de l'ADN et du SIDD, la période la

plus importante étant de l'ordre de 100 kb. Le taux de GC apparaît comme un facteur confondant dans cette analyse. En effet, la corrélation entre la courbure et le taux de GC est intrinsèque (les séquences AT-riches sont plus courbées que les séquences GC-riches). De même, la corrélation entre l'expression des gènes et le taux de GC est liée directement à la conservation des gènes fortement exprimés qui résistent au biais mutationnel vers les bases A et T (Viñuelas *et al.,* 2007). Chez *E. coli* (qui ne présente pas de biais compositionnel aussi marqué), un profil périodique de 100 kb avait déjà été observé et expliqué par la liaison de la gyrase à l'ADN et par le repliement du chromosome (Jeong *et al.,* 2004) et dans une moindre mesure par le niveau de surenroulement du chromosome qui semble être à l'origine de la structuration locale, à plus petite échelle, du chromosome. On peut supposer que cette périodicité est due à la localisation régulière des gènes les plus fortement exprimés. Comme la transcription intense d'un gène peut être à l'origine de domaines topologiques (caractérisés par un certain niveau de surenroulement) qui peuvent à leur tour étendre la région chromosomique transcrite à un taux fort. Afin de vérifier cette hypothèse on pourrait faire l'analyse du profil d'expression des gènes de *Buchnera APS* le long du chromosome en éliminant par exemple les 50 gènes les plus exprimés. Une autre explication possible de cette régularité du profil d'expression pourrait être la présence de longues unités de transcription chez *Buchnera APS*. L'étude de la relation entre les unités de transcription prédites et/ou confirmées de *Buchnera APS*, les variations coordonnées des gènes dans les différentes expériences réalisées sur *Buchnera APS* et la périodicité du profil d'expression, constitue une perspective intéressante.

Le milieu intracellulaire symbiotique sélectionne vraisemblablement les *Buchnera* ayant le profil d'expression le plus adapté et donc indirectement ce contexte évolutif doit imposer des contraintes sur la machinerie de régulation de l'expression de *Buchnera APS*. Comme nous l'avons vu, cette sélection se traduit chez *Buchnera APS* par (1) la conservation des promoteurs σ^{70} ayant une architecture similaire aux promoteurs σ^{70} chez les bactéries libres, (2) par des séquences intergéniques divergentes (contenant des promoteurs) plus longues, avec une composition distincte et plus instable (SIDD plus faible) que les régions convergentes (contenant principalement des terminateurs de transcription). Ainsi, en dépit du fort biais mutationnel vers les bases A et T de *Buchnera APS*, qui est par ailleurs plus fort dans les régions non-codantes de son génome, des éléments structuraux nécessaires à l'expression et à sa régulation sont conservés. Il existe donc chez *Buchnera APS* une sélection sur les séquences non-codantes, ce qui nous amène à nous demander quelle est la participation des propriétés des ces régions non-codantes, comparée à la participation des propriétés des

séquences codantes, à la périodicité du niveau d'expression des gènes. Des périodicités du même ordre (environ 100 kb) on été trouvées pour le profil d'expression des gènes, le taux de bases G et C des séquences codantes et la courbure des régions promotrices. Néanmoins, aucune corrélation significative entre le niveau d'expression des gènes et la courbure moyenne des régions promotrices n'a été trouvée. On peut en déduire que les périodes du profil de l'expression et celles des courbures des régions promotrices doivent être décalées. A l'avenir, l'analyse de la périodicité devrait être complétée en intégrant les unités de transcription.

La quatrième et dernière partie de notre travail a consisté à essayer de construire le réseau de régulation de *Buchnera APS* selon une méthode descendante en utilisant les données d'expression acquises expérimentalement dans l'UMR BF2I (Reymond *et al.,* 2006; Viñuelas *et al.,* 2007; Bermingham *et al.,* 2009). Pour cela nous avons tout d'abord dû développer une méthode d'inférence de réseau (appelée IGOIM) adaptée à notre modèle d'étude ; c'est à dire adaptée au sous-échantillonage, car même si cette condition est valable pour tous les modèles, elle est encore plus cruciale chez *Buchnera APS* pour laquelle l'exploration des régimes d'expression est un verrou expérimental très difficile. D'autre part, nous avons fait le choix de limiter notre inférence à des graphes d'ordre 0-1 (et non pas à des graphes complets) car nous avions observé (1) expérimentalement des faibles différentiels d'expression associés à nos jeux de données et (2) *in silico* une densité faible de la matrice de dépendance (peu de régulateurs présents) lors de la construction ascendante du réseau (parties I à III de cette thèse).

IGOIM a été développée et validée sur des jeux de données simulés issus de la littérature (cf. Résultats, Discussion sur la méthode et la validation d'IGOIM), mais au final nous ne l'avons pas appliquée aux données de *Buchnera APS* au cours de cette thèse, et ceci pour deux raisons principales liées à la qualité et à la quantité des données disponibles. En effet, les premières données d'expression disponibles pour *Buchnera APS* ont été acquises dans des conditions de stress nutritionnel : déplétion spécifique en acides aromatiques, déplétion globales en acides aminés essentiels et stress osmotique (Reymond *et al.,* 2006). A l'issue de cette étude et pour tenter d'expliquer des différentiels d'expression relativement faibles, les auteurs ont posé la question de l'aspect cinétique de la régulation et du mélange des populations de *Buchnera APS* issues des compartiments maternel et embryonnaire du puceron. Viñuelas et al. (2007) ont alors montré, qu'en effet, le facteur cinétique est primordial puisqu'une déplétion spécifique de la leucine dans l'alimentation du puceron induit une réponse transcriptionnelle forte de *Buchnera APS* entre un et trois jours pour la plupart des

gènes de la voie de biosynthèse de cet acide aminé essentiel, alors qu'au bout de sept jours (temps établi dans l'expérience de Reymond et al. (2006)) l'expression des gènes de *Buchnera APS* s'éteint. Parallèlement, les études menées sur le compartiment embryonnaire par Bermighnam et al. (2009) ont montré très clairement un différentiel d'expression entre les *Buchnera* des embryons en début de développement (croissance plus rapide) par rapport aux *Buchnera* des embryons en fin de développement et aux *Buchnera* des bactériocytes maternels. Ainsi, pour obtenir des données contrastées et de qualité suffisante, il apparaît nécessaire de réaliser des expériences cinétiques et de séparer les populations maternelles et embryonnaires des *Buchnera*. De telles données (couplées à une analyse transcriptomique de l'hôte) sont actuellement en cours d'acquisition dans le laboratoire (F. Calevro et S. Colella). Elles pourront peut-être ensuite être exploitées par IGOIM. La question de la dimension, c'est-à-dire du nombre de conditions contrastées testables reste un des points les plus difficiles de ces analyses futures.

Partie V

Conclusions et perspectives

Buchnera aphidicola est une bactérie qui intrigue aussi bien par son apparente simplicité que par son statut intermédiaire entre cellule autonome et organite intracellulaire. Façonné par des forces évolutives encore non entièrement décryptées, son système de régulation de l'expression des gènes ne semble pas dominé par les mêmes mécanismes que chez les bactéries à forme de vie libre. Comprendre l'évolution de la machinerie de régulation de *Buchnera* permet de mieux comprendre les contraintes évolutives qui se sont exercées sur les autres sous-systèmes moléculaires de la bactérie (*e.g.* réseau métabolique, réseau d'interaction protéine-protéine). Par ailleurs, dans la mesure où la bactérie doit répondre aux sollicitations nutritionnelles variables de son hôte (notamment au cours du développement du puceron), comprendre les mécanismes de régulation de *Buchnera* permet aussi de mieux comprendre les liens qui se tissent entre les deux partenaires de cette association alimentaire. Plus généralement, le décryptage de l'évolution du système de régulation chez *Buchnera* peut aussi enrichir le savoir qu'on a aujourd'hui sur les scenarios possibles du devenir des réseaux de régulation dans un contexte de réduction drastique de génome.

Ce travail de thèse, fait suite aux travaux expérimentaux ayant débuté la saga du décryptage des mécanismes par lesquels *Buchnera* adapte sa fourniture d'acides aminés essentiels en réponse à la variation en besoins du puceron ((Douglas and Prosser 1992; Febvay *et al.,* 1999) pour plus de détail voir Brinza et al. (2009)). Cette saga compte peu d'études portant sur la régulation transcriptionnelle (Nakabachi and Ishikawa 1997; Baumann *et al.,* 1999; Moran *et al.,* 2003; Wilcox *et al.,* 2003; Moran *et al.,* 2005b; Reymond *et al.,* 2006) et c'est dans cette direction que nous nous sommes investis en présentant une analyse bioinformatique approfondie des capacités régulatrices de *Buchnera APS*. Cette analyse se base sur un inventaire des acteurs protéiques de la régulation, sur une analyse des éléments génomiques, de leur organisation sur le chromosome de la bactérie et de leur évolution. *Buchnera APS* possède seulement deux facteurs σ de l'ARN polymérase (σ^{70} et σ^{32}) ayant perdu les quatre autres présents dans la majorité des entérobactéries et dans l'ancêtre commun de *Buchnera APS* et d'*E. coli*, seulement trois facteurs de transcription spécifiques potentiels et huit topo-régulateurs. L'analyse des éléments génomiques de *Buchnera APS* nous a permis de révéler (1) une carte opéronique originale, avec des unités de transcription plus longues que celles d'*E. coli* ; (2) des tailles de régions non-codantes entre les unités de transcription similaires à celles d'*E. coli* ; (3) une taille significativement plus courte des régions intergéniques convergentes par rapport aux régions divergentes ; (4) des promoteurs σ^{70} pour

95% des gènes et des promoteurs internes aux unités de transcription avec des séquences significativement moins conservées ; (5) une distribution des tailles des régions 5'UTR similaire à celles d'*E. coli* ; (6) des promoteurs σ^{32} pour 244 gènes de *Buchnera APS* suggérant comme chez *Candidatus Blochmania floridanus* que ce facteur pourrait devenir constitutif, au même titre que le σ^{70} (Stoll *et al.,* 2009) ; (7) un génome plus courbé intrinsèquement, plus flexible et moins stable principalement à cause du biais compositionnel très fort vers les bases A et T ; (8) des régions promotrices instables (donc s'ouvrant plus facilement) caractérisant les gènes codant les régulateurs ; (9) des gènes cataboliques avec des régions promotrices plus stables et enfin (10) une périodicité de 100 kb dans les profils d'expression, de courbure intrinsèque des régions promotrices et du taux de GC des régions codantes.

Nous avons ainsi montré que chez *Buchnera APS* l'architecture des unités de transcription, leur taille mise à part, est similaire à celle d'*E. coli*, avec des promoteurs qui semblent fonctionnels, contrairement à ce qui a été dit dans la littérature (Moran and Mira 2001) et contrairement à d'autres bactéries intracellulaires (Himmelreich *et al.,* 1997; Clark *et al.,* 2001; Nakabachi *et al.,* 2006). La périodicité de 100 kb du profil d'expression, de la courbure et du taux de GC suggère une structuration du chromosome à cette échelle, structuration qui semble cohérente avec la composition et les propriétés structurelles de son chromosome. Naturellement, nous nous sommes demandés si la périodicité de l'expression ne peut s'expliquer par celle de la courbure des régions promotrices, sachant que généralement chez les bactéries les régions promotrices des gènes les plus exprimés sont plus courbées (Olivares-Zavaleta *et al.,* 2006; Nov Klaiman *et al.,* 2009). Cependant, nous n'avons pas retrouvé chez *Buchnera APS* de corrélation entre le niveau d'expression des gènes (qui sont fortement corrélés avec le taux de GC de leurs séquences codantes (Viñuelas *et al.,* 2007)) et la courbure de leurs régions promotrices. Ce résultat reste donc à être étudié plus précisément et interprété à l'avenir. Toutefois comme l'échelle de structuration (accessible par la périodicité du profil d'expression) est du même ordre de grandeur que l'échelle de composition en bases nucléotidiques et que celle des propriétés structurelles, ceci nous amène à nous poser des questions sur les contraintes responsables du biais compositionnel en bases A et T, qui jusque là avaient toujours été vues plutôt soit comme une conséquence de la dégénérescence de certains appareils de réparation (Tamas *et al.,* 2008) soit comme une contrainte énergétique liée à la centralité de l'ATP (Rocha and Danchin 2002). Nous avons émis l'hypothèse que ce biais fort pourrait représenter un avantage sélectif par les propriétés conformationnelles et d'expression qu'il provoque au niveau

du chromosome (le rendant plus flexible et plus facilement « régulable » par un petit ensemble de toporégulateurs globaux).

Un travail méthodologique a également été réalisé dans le cadre de cette thèse. Nous avons conçu une méthode de prédiction d'opérons (DisTer) ainsi qu'une méthode d'inférence de réseaux de gènes à partir de données d'expression (IGOIM). Dans les deux cas, nous avons tenu compte de l'originalité de notre modèle biologique.

Les opérons sont des structures qui évoluent vite (Price *et al.,* 2006). Certaines propriétés des groupes de gènes, comme la similarité des fonctions ou l'implication dans la même voie de biosynthèse, permettent de détecter des opérons très conservés, mais pas ceux qui sont relativement « jeunes » généralement transitoires dans les lignées. Or, nous avons vu que chez *Buchnera* à moins de les éliminer, la destruction des opérons, ne semble plus possible et la lignée semble avoir accumulé ces opérons transitoires. Pour cette raison, nous avons décidé de n'utiliser que des propriétés structurelles pour la prédiction des opérons de *Buchnera APS*. Ainsi, DisTer utilise la distance intergénique et les terminateurs de transcription prédits avec TransTermHP (Kingsford *et al.,* 2007). La carte opéronique de *Buchnera APS* construite avec DisTer est plus en accord avec les données d'expression et les expériences de RT-PCR que nous avons réalisées, par rapport aux trois autres cartes trouvées dans la littérature. Nous avons constaté a posteriori que le SIDD et les scores de promoteurs σ^{70} prédits par Bprom sont des variables discriminatives des paires de gènes intra-opéroniques relativement aux paires de gènes inter-opéroniques. L'intégration de ces variables dans le modèle de DisTer, de même que l'application de DisTer aux autres *Buchnera* séquencées (et à d'autres bactéries à génome réduit) représentent des perspectives d'amélioration du prédicteur intéressantes.

La méthode IGOIM (Inférence de Graphes de premier Ordre avec l'Information Mutuelle conditionnelle) est adaptée aux conditions de sous-échantillonnage et aux relations de dépendance non-linéaires entre les niveaux d'expression des gènes. C'est surtout l'adaptation au sous-échantillonnage qui est intéressante pour le modèle *Buchnera*, les données d'expression et les conditions expérimentales variées à tester étant difficiles à réaliser sur ce modèle. En perspective, IGOIM pourrait être améliorée en utilisant l'information mutuelle conditionnelle d'ordre 2 (ou éventuellement d'un ordre encore supérieur), avec le choix de l'ordre en fonction de la quantité de données d'expression. Dans cette optique, il serait nécessaire de réaliser une analyse de puissance grâce à d'autres données simulées plus complètes et plus riches afin de déterminer, par exemple, la taille de la matrice d'expression minimale nécessaire.

La simulation de données d'expression nécessaire pour étudier, calibrer et comparer entre elles les méthodes d'inférence de réseaux de gènes, bien que indispensable, reste une question ardue et un verrou méthodologique. La difficulté de ce problème vient notamment de la nécessité de s'approcher par la simulation des données d'expression réelles. Or, nous ne connaissons pas encore suffisamment la nature même de ces données. Même en restant dans un cadre « statique », il faudrait d'abord savoir quel type de réseau utiliser : réseau petit-monde ou échelle-invariant sachant que ni l'un ni l'autre ne caractérisent entièrement les réseaux biologiques connus aujourd'hui. De plus, lorsqu'on essaye de faire varier plusieurs paramètres liés à la structure de tels graphes, on constate qu'on ne peut pas s'approcher simultanément des paramètres des vrais réseaux (Van den Bulcke *et al.,* 2006). Naturellement, les propriétés de structures des réseaux dépendent étroitement du type d'entité que l'on met dans les nœuds de ces réseaux. Il faut donc commencer par analyser les paramètres de structures pertinents pour caractériser les réseaux de gènes. Vient ensuite l'aspect « dynamique » de la simulation, quel type de modèle mathématique intégrant la stochasticité du processus d'expression devrait-on utiliser ? Plusieurs outils de simulation de données sont actuellement proposés dans la littérature (Mendes 1993, 1997; Kierzek 2002; You *et al.,* 2003; Adalsteinsson *et al.,* 2004; Van den Bulcke *et al.,* 2006; Ribeiro and Lloyd-Price 2007). Très souvent ces outils ne prennent pas en charge l'aspect statique de la simulation laissant l'utilisateur choisir son réseau, et utilisent un seul modèle dynamique de l'expression. Dans les premiers mois de la thèse, nous avons réfléchi au développement d'un outil de simulation plus performant et permettant de simuler des données d'expression plus complètes. Néanmoins nous n'avons finalement pas concrétisé ce projet car nous l'avons jugé trop coûteux en temps de travail.

Après ces trois ans d'étude *in silico* sur la régulation de l'expression des gènes chez *Buchnera APS*, nous pensons maintenant qu'il faudrait se diriger vers des analyses expérimentales. C'est dans cette optique que nous avons envisagé les plus grandes perspectives associées à ce travail.

L'analyse expérimentale des régulateurs potentiels de la transcription que nous avons recensés (incluant le facteur σ^{32}), ainsi que la détermination de leurs régulons, nous semble très importante. Cette étude pourrait se faire par les techniques dites « Chip-Chip » (ou Chip-Seq), comme cela a été fait précédemment pour *E. coli* (Cho *et al.,* 2009). Grâce à cette technique combinant l'immunoprécipitation de protéines associées à l'ADN *in vivo*, et les puces à ADN (ou le séquençage haut débit), nous aurons accès aux séquences d'ADN sur lesquelles les différents régulateurs se fixent (à

condition d'avoir les anticorps spécifiques correspondant aux protéines que l'on veut étudier). Cette expérience permettrait ainsi de vérifier notre hypothèse sur le facteur σ^{32} comme facteur constitutif, de valider les promoteurs σ^{32} que nous avons prédits, et également de vérifier et compléter les régulons correspondants aux autres facteurs de transcription recensés.

Un autre grand projet expérimental serait d'analyser, par d'autres moyens expérimentaux, l'échelle de structuration du chromosome que nous avons déterminée avec la périodicité du profil d'expression. Une étude envisageable chez *Buchnera APS* serait de tester la toporégulation, à la manière de Muskhelishvili et collaborateurs grâce à des agents intercalants capables de changer le niveau de surenroulement du chromosome (Schneider *et al.,* 1999; Blot *et al.,* 2006). Ces intercalants (la norfloxacine et la coumermicine) sont des antibiotiques bactériens et ne devraient pas présenter une toxicité forte pour le puceron. Un traitement *in vivo* de *Buchnera APS* (via l'alimentation du puceron ou en injection intra-abdominale) semble réalisable. Il restera à vérifier le passage de la barrière intestinale et/ou des membranes symbiosomales pour arriver à *Buchnera*. L'analyse transcriptomique différentielle des *Buchnera* traitées (à surenroulement variable) relativement à des *Buchnera* non-traitées permettrait de mesurer directement l'effet du changement de la conformation du chromosome sur le profil d'expression, mais également de tester l'influence de ce changement sur la périodicité de ce profil (à savoir si le profil est toujours périodique, et dans ce cas, si la période est toujours de 100kb). Nous pouvons également imaginer l'étude des toporégulateurs sur la structuration du chromosome, ainsi que sur leur activité en tant que facteurs de transcription par des expériences de mutations et/ou d'ARN antisens. Par exemple, nous pourrions construire un ARN à séquence complémentaire à l'ARNm du H-NS qui devrait conduire à la dégradation des ARNm de H-NS de la cellule, diminuant ainsi son expression. Néanmoins, chez *Buchnera APS*, la transformation bactérienne ou l'inactivation génique nécessiterait au moins de pouvoir cultiver les bactériocytes, voire les *Buchnera AS* hors de leur hôte. Si la culture de *Buchnera APS* ne semble a priori pas réalisable (la bactérie est trop dépendante de son hôte), on peut plus facilement imaginer la culture des bactériocytes en utilisant des milieux de culture d'insectes éventuellement complémentés. Ainsi, dans l'UMR BF2I des milieux de culture d'embryons de pucerons ont déjà été définis (A. Rabatel et F. Calevro, non publiées) et un programme de recherche sur la culture des bactériocytes est prévu dans les années à venir.

Parallèlement, des analyses bioinformatiques devront accompagner ce travail. Par exemple, un des mécanismes de régulation de l'expression des gènes chez les procaryotes que nous n'avons pas étudiés

pendant cette thèse, sont les petits ARN. L'analyse des petits ARN de *Buchnera APS* est envisagée dans le cadre du projet Sisyphe (projet ERC Marie-France Sagot) dans l'équipe Bamboo auquel l'UMR BF2I est associée. Un des objectifs de ce projet est l'analyse bioinformatique et la prédiction des petits ARN du couple symbiotique (chez le puceron et chez *Buchnera*) et de leurs cibles, ainsi que l'investigation expérimentale de ce mécanisme de régulation par l'analyse du séquençage à haut débit des miARN. Une autre question majeure de ce projet est l'allorégulation de l'expression des gènes chez *Buchnera APS*, *i.e.*, si des molécules de l'hôte (*e.g.* miARN, facteurs de transcription eucaryotes) interfèrent et coordonnent la machinerie de transcription de *Buchnera* en traversant la barrière symbiosomale.

L'étude globale que nous avons faîte pourrait être replacée dans un contexte beaucoup plus général de génomique comparative à l'image du travail réalisé par L. Cottret (2009) sur le réseau métabolique de *Buchnera*. Une collaboration étroite avec l'équipe du Pr. Guillaume Beslon (INSA, LIRIS), notamment dans le cadre de son projet « Evolution de la complexité moléculaire des organismes bactériens : approches couplées en génomique comparative et vie artificielle » devrait nous permettre de déceler les principales forces évolutives ayant modelé le génome de *Buchnera APS*, grâce aux simulations réalisées à l'aide d'un modèle génétique digital (*Raevol* (Knibbe *et al.,* 2007a; Knibbe *et al.,* 2007b)) et à des analyses de génomique comparative avec d'autres bactéries libres à génome réduit caractérisées par des tailles de populations et des taux de mutations très différents des bactéries endocytobiotes intracellulaires.

Partie VI

Publications et communications

Communications :

Brinza L., C. Armenise, E. Leproult, A. Paun, P. Da Silva, F. Calevro and H. Charles (2006). Bioinformatics analysis of the nucleoid associated protein FIS in *Buchnera aphidicola*, the intracellular symbiotic bacteria of aphids. Integrative Post-Genomics - IPG'06, Lyon (France), novembre-décembre 2006.

Brinza L., F. Calevro and H. Charles (2007). IGOIM, une méthode d'inférence de réseaux génétiques utilisant l'information mutuelle. Journées Ouvertes Biologie Informatique Mathématiques (JOBIM) - Journée satellite "Analyse statistique du transcriptome", Marseille (France), juillet 2007.

Brinza L., F. Calevro and H. Charles (2009). Fate of transcription units in the reductive evolution of the genome of *Buchnera aphidicola*. 6th International Symbiosis Society (ISS) Congress, Madison (USA), août 2009.

Brinza L., F. Calevro, G. Duport, Y. Rahbé, P. Da Silva, J. P. Gauthier and H. Charles (2010). Evolution of the regulatory network in *Buchnera aphidicola*, the obligate intracellular symbiotic bacteria of aphids. Annual Meeting of the Society for Molecular Biology and Evolution (SMBE 2010), Lyon (France), juillet 2010.

Brinza L., F. Calevro, J. Vinuelas, C. Gautier and H. Charles (2009). Chromosome organisation in *Buchnera*: a dynamic active structure involved in gene expression regulation. Journées Ouvertes Biologie Informatique Mathématiques (JOBIM), Nantes (France), juin 2009.

Calevro F., H. Charles, J. M. Fayard, G. Febvay, Y. Rahbé, G. Duport, J. Vinuelas, A. Rabatel, S. Colella, S. Laroche, et al. (2008). Interactions trophiques dans la symbiose puceron - *Buchnera* : Régulations transcriptionnelles de la réponse à la carence en acides aminés essentiels. Réseau Français de Biologie Adaptative des Pucerons, Paris (France), janvier 2008.

Charles H. and L. Brinza (2010). Exploring the regulatory network of *Buchnera aphidicola*: a case-study of genome evolution in a symbiotic context. Séminaire Rhônalpin de Modélisation du Vivant (SEMOVI), Lyon (France), juin 2010.

Charles H., F. Calevro, S. Colella, J. M. Fayard, G. Febvay, Y. Rahbé, G. Duport, S. Laroche, J. Vinuelas, L. Brinza, et al. (2008). Caractérisation et modélisation de la « fonction symbiotique » de *Buchnera aphidicola* chez le puceron du pois *Acyrthosiphon pisum*. Séminaire INRA AgroBI, Bordeaux (France), décembre 2008.

Cottret L., L. Brinza, J. Vinuelas, G. Duport, F. Calevro, G. Febvay, Y. Rahbé, J. M. Fayard and H. Charles (2006). Régulation génétique et métabolique chez *Buchnera aphidicola* (AgroBi 2006). Journée INRA-INRIA "Insertion des mathématiques dans les approches biologiques : du gène à l'organisme", Lyon (France), décembre 2006.

Publications :

Brinza L., J. Vinuelas, L. Cottret, F. Calevro, Y. Rahbé, G. Febvay, G. Duport, S. Colella, A. Rabatel, C. Gautier, et al. (2009c). "Systemic analysis of the symbiotic function of *Buchnera aphidicola*, the primary endosymbiont of the pea aphid *Acyrthosiphon pisum*." CR Biol (Acad Sci Paris) **332**: 1034-1049.

Deniaud E., J. Baguet, R. Chalard, B. Blanquier, L. Brinza, J. Meunier, M.-C. Michallet, A. Laugraud, C. Ah-Soon, A. Wierinckx, et al. (2009). "Overexpression of transcription factor Sp1 leads to gene expression perturbations and cell cycle inhibition." PLoS ONE **4**(9): e7035, 13p.

Brinza L., F. Calevro, G. Duport, C. Gautier and H. Charles (2010). "Structure and dynamics of the operonic map in the genome of *Buchnera aphidicola sp. strain APS*." BMC Genomics. (*accepté*)

Brinza L. and H. Charles "IGOIM, a mutual information based method for predicting gene networks." (*en préparation*)

Brinza L., H. Charles, F. Calevro and J. P. Gauthier "Transcriptional regulation in *Buchnera aphidicola*, a genome reduced bacteria." (*en préparation*)

Partie VII

Références bibliographiques

ADALSTEINSSON, D., MCMILLEN, D. and ELSTON, T. C. Biochemical Network Stochastic Simulator (BioNetS): software for stochastic modeling of biochemical networks. *BMC Bioinformatics*, 2004, **5**: 24.

AL-SHAHROUR, F., DÍAZ-URIARTE, R. and DOPAZO, J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 2004, **20**(4): 578-80.

ALDEA, M., GARRIDO, T., HERNANDEZ-CHICO, C., VICENTE, M. and KUSHNER, S. R. Induction of a growth-phase-dependent promoter triggers transcription of bolA, an *Escherichia coli* morphogene. *EMBO J*, 1989, **8**(12): 3923-31.

ALI, B. M., AMIT, R., BRASLAVSKY, I., OPPENHEIM, A. B., GILEADI, O. and STAVANS, J. Compaction of single DNA molecules induced by binding of integration host factor (IHF). *Proc Natl Acad Sci U S A*, 2001, **98**(19): 10658-63.

ARAVIND, L., ANANTHARAMAN, V., BALAJI, S., BABU, M. M. and IYER, L. M. The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiology Reviews*, 2005, **29**(2): 231-62.

AUNER, H., BUCKLE, M., DEUFEL, A., KUTATELADZE, T., LAZARUS, L., MAVATHUR, R., MUSKHELISHVILI, G., PEMBERTON, I., SCHNEIDER, R. and TRAVERS, A. Mechanism of transcriptional activation by FIS: role of core promoter structure and DNA topology. *J Mol Biol*, 2003, **331**(2): 331-44.

AZAM, T. and ISHIHAMA, A. Twelve Species of the Nucleoid-associated Protein from *Escherichia coli*. *J Biol Chem*, 1999, **274**: 33105-33113.

BABU, M., BALAJI, S. and ARAVIND, L. General trends in the evolution of prokaryotic transcriptional regulatory networks. *Genome dynamics*, 2007.

BACHLER, C., SCHNEIDER, P., BAHLER, P., LUSTIG, A. and ERNI, B. *Escherichia coli* dihydroxyacetone kinase controls gene expression by binding to transcription factor DhaR. *EMBO J*, 2005, **24**(2): 283-93.

BAE, W., XIA, B., INOUYE, M. and SEVERINOV, K. *Escherichia coli* CspA-family RNA chaperones are transcription antiterminators. *Proc Natl Acad Sci U S A*, 2000, **97**(14): 7784-9.

BALLEZA, E., LOPEZ-BOJORQUEZ, L. N., MARTINEZ-ANTONIO, A., RESENDIS-ANTONIO, O., LOZADA-CHAVEZ, I., BALDERAS-MARTINEZ, Y. I., ENCARNACION, S. and COLLADO-VIDES, J. Regulation by transcription factors in bacteria: beyond description. *FEMS Microbiol Rev*, 2009, **33**(1): 133-51.

BANSAL, M. and DI BERNARDO, D. Inference of gene networks from temporal gene expression profiles. *IET Syst Biol*, 2007, **1**(5): 306-12.

BARKER, M. M., GAAL, T., JOSAITIS, C. A. and GOURSE, R. L. Mechanism of regulation of transcription initiation by ppGpp. I. Effects of ppGpp on transcription initiation *in vivo* and *in vitro*. *J Mol Biol*, 2001, **305**(4): 673-88.

BAUMANN, L., BAUMANN, P. and THAO, M. L. Detection of messenger RNA transcribed from genes encoding enzymes of amino acid biosynthesis in *Buchnera aphidicola* (Endosymbiont of aphids). *Current Microbiology*, 1999, **38**(2): 135-136.

BAUMANN, P., BAUMANN, L., LAI, C. Y., ROUHBAKHSH, D., MORAN, N. A. and CLARK, M. A. Genetics, physiology and evolutionary relationships of the genus *Buchnera*: Intracellular symbionts of aphids. *Annual Review of Microbiology*, 1995, **49**: 55-94.

BAUMANN, P., MUNSON, M. A., LAI, C. Y., CLARK, M. A., BAUMANN, L., MORAN, N. A. and CAMPBELL, B. C. Origin and properties of the bacterial endosymbionts of aphids, whiteflies, and mealybugs. *American Society of Microbiology News*, 1993, **59**(1): 21-24.

BAUMBACH, J., RAHMANN, S. and TAUCH, A. Reliable transfer of transcriptional gene regulatory networks between taxonomically related organisms. *BMC Syst Biol*, 2009, **3**: 8.

BEIRLANT, J., DUDEWICZ, E., GYÖRFI, L. and VAN DER ..., E. Nonparametric entropy estimation: An overview. *International Journal of ...*, 1997.

BENDAT, S. J. and PIERSOL, G. A. Measurement and analysis of random data. Wiley, 1968.

BENHAM, C. J. Duplex destabilization in superhelical DNA is predicted to occur at specific transcriptional regulatory regions. *J Mol Biol*, 1996, **255**(3): 425-34.

BENHAM, C. J. and BI, C. The analysis of stress-induced duplex destabilization in long genomic DNA sequences. *J Comput Biol*, 2004, **11**(4): 519-43.

BERGER, J. M., FASS, D., WANG, J. C. and HARRISON, S. C. Structural similarities between topoisomerases that cleave one or both DNA strands. *Proc Natl Acad Sci USA*, 1998, **95**(14): 7876-81.

BERMINGHAM, J., RABATEL, A., CALEVRO, F., VINUELAS, J., FEBVAY, G., CHARLES, H., DOUGLAS, A. and WILKINSON, T. Impact of host developmental age on the transcriptome of the symbiotic bacterium *Buchnera aphidicola* in the pea aphid (*Acyrthosiphon pisum*). *Appl Environ Microbiol*, 2009, **75**(22): 7294-7.

BERMINGHAM, J. and WILKINSON, T. L. Embryo nutrition in parthenogenetic viviparous aphids. *Physiological Entomology*, 2009, **34**(2): 103-109.

BERTHET, J. Dictionnaire de Biologie De Boeck, 2005.

BIRKLE, L. M., MINTO, L. B., WALTERS, K. F. A. and DOUGLAS, A. E. Microbial genotype and insect fitness in an aphid-bacterial symbiosis. *Functional Ecology*, 2004, **18**(4): 598-604.

BISHOP, C. Pattern recognition and machine learning. 2006.

BLATTNER, F. R., PLUNKETT, G., 3RD, BLOCH, C. A., PERNA, N. T., BURLAND, V., RILEY, M., COLLADO-VIDES, J., GLASNER, J. D., RODE, C. K., MAYHEW, G. F., GREGOR, J., DAVIS, N. W., KIRKPATRICK, H. A., GOEDEN, M. A., ROSE, D. J., MAU, B. and SHAO, Y. The complete genome sequence of *Escherichia coli K-12*. *Science*, 1997, **277**(5331): 1453-62.

BLOCH, V., YANG, Y., MARGEAT, E., CHAVANIEU, A., AUGE, M. T., ROBERT, B., AROLD, S., RIMSKY, S. and KOCHOYAN, M. The H-NS dimerization domain defines a new fold contributing to DNA recognition. *Nat Struct Biol*, 2003, **10**(3): 212-8.

BLOT, N., MAVATHUR, R., GEERTZ, M., TRAVERS, A. and MUSKHELISHVILI, G. Homeostatic regulation of supercoiling sensitivity coordinates transcription of the bacterial genome. *EMBO Rep*, 2006, **7**(7): 710-5.

BOCKHORST, J., CRAVEN, M., PAGE, D., SHAVLIK, J. and GLASNER, J. A Bayesian network approach to operon prediction. *Bioinformatics*, 2003a, **19**(10): 1227-35.

BOCKHORST, J., QIU, Y., GLASNER, J., LIU, M., BLATTNER, F. and CRAVEN, M. Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*, 2003b, **19 Suppl 1**: i34-43.

BOLSHOY, A., MCNAMARA, P., HARRINGTON, R. E. and TRIFONOV, E. N. Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proc Natl Acad Sci USA*, 1991, **88**(6): 2312-6.

BOYER, F., MORGAT, A., LABARRE, L., POTHIER, J. and VIARI, A. Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data. *Bioinformatics*, 2005, **21**(23): 4209-15.

BRACHO, A. M., MARTINEZTORRES, D., LATORRE, A. and MOYA, A. Plasmid-Borne genes for leucine synthesis localized in bacterial symbionts of aphid mycetocytes. *J Mol Evol*, 1994, **prepri**: x-x.

BRAENDLE, C., MIURA, H., BICKEL, R., SHINGLETON, A. W., KAMBHAMPATHI, S. and STERN, D. L. Developmental origin and evolution of bacteriocytes in the aphid-*Buchnera* symbiosis. *Plos Biol*, 2003, **1**(1): 70-76.

BRAZHNIK, P., DE LA FUENTE, A. and MENDES, P. Gene networks: how to put the function in genomics. *TRENDS in Biotechnology*, 2002.

BRILLINGER, D. Some data analyses using mutual information. *Brazilian J Prob and Statist*, 2005.

BRINZA, L., VIÑUELAS, J., COTTRET, L., CALEVRO, F., RAHBÉ, Y., FEBVAY, G., DUPORT, G., COLELLA, S., RABATEL, A., GAUTIER, C., FAYARD, J. M., SAGOT, M. F. and CHARLES, H. Systemic analysis of the symbiotic function of *Buchnera aphidicola*, the primary endosymbiont of the pea aphid *Acyrthosiphon pisum*. *Cr Biol*, 2009, **332**(11): 1034-1049.

BROWNING, D. F. and BUSBY, S. J. The regulation of bacterial transcription initiation. *Nat Rev Microbiol*, 2004, **2**(1): 57-65.

BRUKNER, I., JURUKOVSKI, V. and SAVIC, A. Sequence-dependent structural variations of DNA revealed by DNase I. *Nucleic Acids Res*, 1990, **18**(4): 891-4.

BRYNNEL, E. U., KURLAND, C. G., MORAN, N. A. and ANDERSSON, S. G. E. Evolutionary rates for *tuf* genes in endosymbionts of aphids. *Mol Biol Evol*, 1998, **15**(5): 574-582.

BUCHNER, P. Aphids. In: P. BUCHNER. Endosymbiosis of animals with plant microorganisms. New York (USA), Interscience, 1965, 297-332.

BURR, T., MITCHELL, J., KOLB, A., MINCHIN, S. and BUSBY, S. DNA sequence elements located immediately upstream of the -10 hexamer in *Escherichia coli* promoters: a systematic study. *Nucleic Acids Res*, 2000, **28**(9): 1864-70.

BUTTE, A. and KOHANE, I. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, 2000.

CALEVRO, F., CHARLES, H., REYMOND, N., DUGAS, V., CLOAREC, J. P., BERNILLON, J., RAHBE, Y., FEBVAY, G. and FAYARD, J. M. Assessment of 35mer amino-modified oligonucleotide based microarray with bacterial samples. *J Microbiol Meth*, 2004, **57**(2): 207-18.

CARPENTIER, A.-S., TORRÉSANI, B., GROSSMANN, A. and HÉNAUT, A. Decoding the nucleoid organisation of *Bacillus subtilis* and *Escherichia coli* through gene expression data. *Bmc Genomics*, 2005, **6**(1): 84.

CASHEL, M., GENTRY, D. R., HERNANDEZ, V. J. and VINELLA, D. The stringent response. In: F. C. NEIDHARDT, J. L. INGRAHAM, K. B. LOWet al. *Escherichia coli* and *Salmonella*: cellular and molecular biology. Washington, D.C., ASM Press. **2**, 1996, 1458–1496.

CAVANAGH, A. T., KLOCKO, A. D., LIU, X. and WASSARMAN, K. M. Promoter specificity for 6S RNA regulation of transcription is determined by core promoter sequences and competition for region 4.2 of sigma70. *Mol Microbiol*, 2008, **67**(6): 1242-56.

CHAMPOUX, J. J. DNA topoisomerases: structure, function, and mechanism. *Annu Rev Biochem*, 2001, **70**: 369-413.

CHARANIYA, S., MEHRA, S., LIAN, W., JAYAPAL, K. P., KARYPIS, G. and HU, W. S. Transcriptome dynamics-based operon prediction and verification in *Streptomyces coelicolor*. *Nucleic Acids Res*, 2007, **35**(21): 7222-36.

CHARIF, D., LOBRY, J., NECSULEA, A., PALMEIRA, L. and PENEL, S. Package 'seqinr'. *Citeseer*, 2010.

CHARLES, H., CALEVRO, F., VINUELAS, J., FAYARD, J. M. and RAHBÉ, Y. Codon usage bias and tRNA over-expression in *Buchnera aphidicola* after aromatic amino acid nutritional stress on its host *Acyrthosiphon pisum*. *Nucleic Acids Res*, 2006, **34**: 4583-4592.

CHARLES, H., HEDDI, A. and RAHBÉ, Y. A putative insect intracellular endosymbiont stem clade, within the *Enterobacteriaceae*, infered from phylogenetic analysis based on a heterogeneous model of DNA evolution. *Comptes Rendus de l'Académie des Sciences*, 2001, **324**: 489-494.

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

CHARLES, H., MOUCHIROUD, D., LOBRY, J., GONCALVES, I. and RAHBÉ, Y. Gene size reduction in the bacterial aphid endosymbiont, *Buchnera*. *Mol Biol Evol*, 1999, **16**(12): 1820-1822.

CHARLIER, D., KHOLTI, A., HUYSVELD, N., GIGOT, D., MAES, D., THIA-TOONG, T. L. and GLANSDORFF, N. Mutational analysis of *Escherichia coli* PepA, a multifunctional DNA-binding aminopeptidase. *J Mol Biol*, 2000, **302**(2): 411-26.

CHATTERJI, D. and OJHA, A. K. Revisiting the stringent response, ppGpp and starvation signaling. *Curr Opin Microbiol*, 2001, **4**(2): 160-5.

CHATTOPADHYAY, R. and ROY, S. DnaK-sigma 32 interaction is temperature-dependent. Implication for the mechanism of heat shock response. *J Biol Chem*, 2002, **277**(37): 33641-7.

CHEN, X., SU, Z., DAM, P., PALENIK, B., XU, Y. and JIANG, T. Operon prediction by comparative genomics: an application to the *Synechococcus sp. WH8102* genome. *Nucleic Acids Res*, 2004, **32**(7): 2147-57.

CHENG, Y. S., YANG, W. Z., JOHNSON, R. C. and YUAN, H. S. Structural analysis of the transcriptional activation region on Fis: crystal structures of six Fis mutants with different activation properties. *J Mol Biol*, 2000, **302**(5): 1139-51.

CHEUNG, K. J., BADARINARAYANA, V., SELINGER, D. W., JANSE, D. and CHURCH, G. M. A microarray-based antibiotic screen identifies a regulatory role for supercoiling in the osmotic stress response of *Escherichia coli*. *Genome Res*, 2003, **13**(2): 206-15.

CHO, B. K., ZENGLER, K., QIU, Y., PARK, Y. S., KNIGHT, E. M., BARRETT, C. L., GAO, Y. and PALSSON, B. O. The transcription unit architecture of the *Escherichia coli* genome. *Nat Biotechnol*, 2009, **27**(11): 1043-9.

CLARET, L. and ROUVIERE-YANIV, J. Variation in HU composition during growth of *Escherichia coli*: the heterodimer is required for long term survival. *J Mol Biol*, 1997, **273**(1): 93-104.

CLARK, M. A., BAUMANN, L., THAO, M. L., MORAN, N. A. and BAUMANN, P. Degenerative minimalism in the genome of a psyllid endosymbiont. *J Bacteriol*, 2001, **183**(6): 1853-61.

CLARK, M. A., MORAN, N. A. and BAUMANN, P. Sequence evolution in bacterial endosymbionts having extreme base composition. *Mol Biol Evol*, 1999, **16**(11): 1586-1598.

COMBET, C., BLANCHET, C., GEOURJON, C. and DELEAGE, G. NPS@: network protein sequence analysis. *Trends Biochem Sci*, 2000, **25**(3): 147-50.

COMMICHAU, F. M., HERZBERG, C., TRIPAL, P., VALERIUS, O. and STULKE, J. A regulatory protein-protein interaction governs glutamate biosynthesis in *Bacillus subtilis*: the glutamate dehydrogenase RocG moonlights in controlling the transcription factor GltC. *Mol Microbiol*, 2007, **65**(3): 642-54.

COOLEY, A. E., RILEY, S. P., KRAL, K., MILLER, M. C., DEMOLL, E., FRIED, M. G. and STEVENSON, B. DNA-binding by *Haemophilus influenzae* and *Escherichia coli* YbaB, members of a widely-distributed bacterial protein family. *BMC Microbiol*, 2009, **9**: 137.

COTTRET, L. (2009). Analyse systémique de la symbiose intracellulaire : évolution et organisation du réseau métabolique des endocytobiotes. Thèse de doctorat de l'Université Claude Bernard, Lyon 1**:** 221.

CRAVEN, M., PAGE, D., SHAVLIK, J., BOCKHORST, J. and GLASNER, J. A probabilistic learning approach to whole-genome operon prediction. *Proc Int Conf Intell Syst Mol Biol*, 2000, **8**: 116-27.

CROZAT, E., PHILIPPE, N., LENSKI, R. E., GEISELMANN, J. and SCHNEIDER, D. Long-term experimental evolution in *Escherichia coli*. XII. DNA topology as a key target of selection. *Genetics*, 2005, **169**(2): 523-32.

DAGAN, T., BLEKHMAN, R. and GRAUR, D. The "domino theory" of gene death: gradual and mass gene extinction events in three lineages of obligate symbiotic bacterial pathogens. *Mol Biol Evol*, 2006, **23**(2): 310--316.

DAM, P., OLMAN, V., HARRIS, K., SU, Z. and XU, Y. Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Res*, 2007, **35**(1): 288-98.

DAME, R. T. and GOOSEN, N. HU: promoting or counteracting DNA compaction? *FEBS Lett*, 2002, **529**(2-3): 151-6.

DAME, R. T., VAN MAMEREN, J., LUIJSTERBURG, M. S., MYSIAK, M. E., JANICIJEVIC, A., PAZDZIOR, G., VAN DER VLIET, P. C., WYMAN, C. and WUITE, G. J. Analysis of scanning force microscopy images of protein-induced DNA bending using simulations. *Nucleic Acids Res*, 2005, **33**(7): e68.

DAME, R. T., WYMAN, C. and GOOSEN, N. Structural basis for preferential binding of H-NS to curved DNA. *Biochimie*, 2001, **83**(2): 231-4.

DAUB, C. O., STEUER, R., SELBIG, J. and KLOSKA, S. Estimating mutual information using B-spline functions--an improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, 2004, **5**: 118.

DE BARY, H. De la symbiose. *Revue Internationale des Sciences*, 1879, **3**: 301-309.

DE HOON, M. J., IMOTO, S., KOBAYASHI, K., OGASAWARA, N. and MIYANO, S. Predicting the operon structure of *Bacillus subtilis* using operon length, intergene distance, and gene expression information. *Pac Symp Biocomput*, 2004: 276-87.

DEDRYVER, C. A. (2007). Pucerons : des dégâts et des hommes. Biofutur. **26:** 22-25.

DELMOTTE, F., RISPE, C., SCHABER, J., SILVA, F. J. and MOYA, A. Tempo and mode of early gene loss in endosymbiotic bacteria from insects. *BMC Evol Biol*, 2006, **6**: 56.

DENNIS, P. P., EHRENBERG, M., FANGE, D. and BREMER, H. Varying rate of RNA chain elongation during *rrn* transcription in *Escherichia coli*. *J Bacteriol*, 2009, **191**(11): 3740-6.

DILLON, S. C. and DORMAN, C. J. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nat Rev Microbiol*, 2010, **8**(3): 185-95.

DIXON, G. Aphid ecology, an optimization approach. London, Chapman and Hall,, 1998.

DODD, I. B. and EGAN, J. B. Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. *Nucleic Acids Res*, 1990, **18**(17): 5019-26.

DORMAN, C. Regulation of Transcription in Bacteria by DNA Supercoiling. *Bacterial Physiology*, 2008: 155-178.

DORMAN, C. J. H-NS: a universal regulator for a dynamic genome. *Nat Rev Microbiol*, 2004, **2**(5): 391-400.

DORMAN, C. J., BARR, G. C., NI BHRIAIN, N. and HIGGINS, C. F. DNA supercoiling and the anaerobic and growth phase regulation of *tonB* gene expression. *J Bacteriol*, 1988, **170**(6): 2816-26.

DORMAN, C. J. and DEIGHAN, P. Regulation of gene expression by histone-like proteins in bacteria. *Curr Opin Genet Dev*, 2003, **13**(2): 179-84.

DORMAN, C. J., HINTON, J. C. and FREE, A. Domain organization and oligomerization among H-NS-like nucleoid-associated proteins in bacteria. *Trends Microbiol*, 1999, **7**(3): 124-8.

DOUGLAS, A. E. Sulphate utilization in an aphid symbiosis. *Insect Biochemistry*, 1988, **18**(6): 599-605.

DOUGLAS, A. E. Phloem-sap feeding by animals: problems and solutions. *J Exp Bot*, 2006a, **57**(4): 747-54.

DOUGLAS, A. E. Phloem-sap feeding by animals: problems and solutions. *Journal of Experimental Botany*, 2006b.

DOUGLAS, A. E. Conflict, cheats and the persistence of symbioses. *New Phytologist*, 2008, **177**(4): 849-858.

DOUGLAS, A. E. and PROSSER, W. A. Synthesis of the essential amino acid tryptophan in the pea aphid (*Acyrthosiphon pisum*) symbiosis. *Journal of Insect Physiology*, 1992, **38**(8): 565-568.

DRLICA, K. and ROUVIERE-YANIV, J. Histonelike proteins of bacteria. *Microbiol Rev*, 1987, **51**(3): 301-19.

DRUMMOND, A. J., ASHTON, B., BUXTON, S., CHEUNG, M., COOPER, A., HELED, J., KEARSE, M., MOIR, R., STONES-HAVAS, S., STURROCK, S., THIERER, T. and WILSON, A. Geneious v5.1 http://www.geneious.com. 2010.

EFRON, B. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 1981, **68** (3): 589-599.

ELOWITZ, M. B., LEVINE, A. J., SIGGIA, E. D. and SWAIN, P. S. Stochastic gene expression in a single cell. *Science*, 2002, **297**(5584): 1183-6.

ENZ, S., MAHREN, S., MENZEL, C. and BRAUN, V. Analysis of the ferric citrate transport gene promoter of *Escherichia coli*. *J Bacteriol*, 2003, **185**(7): 2387-91.

ERMOLAEVA, M. D., WHITE, O. and SALZBERG, S. L. Prediction of operons in microbial genomes. *Nucleic Acids Res*, 2001, **29**(5): 1216-21.

ESPOSITO, D., PETROVIC, A., HARRIS, R., ONO, S., ECCLESTON, J. F., MBABAALI, A., HAQ, I., HIGGINS, C. F., HINTON, J. C., DRISCOLL, P. C. and LADBURY, J. E. H-NS oligomerization domain structure reveals the mechanism for high order self-association of the intact protein. *J Mol Biol*, 2002, **324**(4): 841-50.

ESTREM, S. T., GAAL, T., ROSS, W. and GOURSE, R. L. Identification of an UP element consensus sequence for bacterial promoters. *Proc Natl Acad Sci U S A*, 1998, **95**(17): 9761-6.

FEBVAY, G., LIADOUZE, I., GUILLAUD, J. and BONNOT, G. Analysis of energetic aminoacid metabolism in *Acyrthosiphon pisum*: a multidimensional aproach to amin acid metabolism in aphids. *Archives of Insect Biochemistry and Physiology*, 1995, **29**: 45-69.

FEBVAY, G., RAHBÉ, Y., RYNKIEWICZ, M., GUILLAUD, J. and BONNOT, G. Fate of dietary sucrose and neosynthesis of amino acids in the pea aphid, *Acyrthosiphon pisum*, reared on different diets. *Journal of Experimental Biology*, 1999, **202**(19): 2639-2652.

FLASHNER, Y. and GRALLA, J. D. DNA dynamic flexibility and protein recognition: differential stimulation by bacterial histone-like protein HU. *Cell*, 1988, **54**(5): 713-21.

FRANK, A. C., AMIRI, H. and ANDERSSON, S. G. E. Genome deterioration: loss of repeated sequences and accumulation of junk DNA. *Genetica*, 2002, **115**(1): 1-12.

FRANK, A. C. and LOBRY, J. R. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene*, 1999, **238**(1): 65-77.

FREDERICO, L. A., KUNKEL, T. A. and SHAW, B. R. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry*, 1990, **29**(10): 2532-7.

FREIRE, P., MOREIRA, R. N. and ARRAIANO, C. M. BolA inhibits cell elongation and regulates MreB expression levels. *J Mol Biol*, 2009, **385**(5): 1345-51.

FRIEDMAN, N. Inferring cellular networks using probabilistic graphical models. *Science*, 2004, **303**(5659): 799-805.

FRONTIER, S. and PICHOD-VIALE, D. Ecosystèmes - structure, fonctionnement, évolution. Masson, 1991.

FUKATSU, T. and ISHIKAWA, H. A novel eukaryotic extracellular symbiont in an aphid, *Astegopteryx styraci* (Homoptera, Aphididae, Hormaphidinae). *Journal of Insect Physiology*, 1992, **38**(10): 765-773.

FUNK, D. J., WERNEGREEN, J. J. and MORAN, N. A. Intraspecific variation in symbiont genomes: Bottlenecks and the aphid-*Buchnera* association. *Genetics*, 2001, **157**(2): 477-489.

GAMA-CASTRO, S., JIMENEZ-JACINTO, V., PERALTA-GIL, M., SANTOS-ZAVALETA, A., PENALOZA-SPINOLA, M. I., CONTRERAS-MOREIRA, B., SEGURA-SALAZAR, J., MUNIZ-RASCADO, L., MARTINEZ-FLORES, I., SALGADO, H., BONAVIDES-MARTINEZ, C., ABREU-GOODGER, C., RODRIGUEZ-PENAGOS, C., MIRANDA-RIOS, J., MORETT, E., MERINO, E., HUERTA, A. M., TREVINO-QUINTANILLA, L. and COLLADO-VIDES, J. RegulonDB (version 6.0): gene regulation model of *Escherichia coli K-12* beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res*, 2008, **36**(Database issue): D120-4.

GARDNER, T. S. and FAITH, J. J. Reverse-engineering transcription control networks. *Phys Life Rev*, 2010, **2**(1): 65-88.

GIL, R., SABATER-MUNOZ, B., LATORRE, A., SILVA, F. J. and 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, **99**(7): 4454-8.

GIL, R., SABATER-MUNOZ, B., PEREZ-BROCAL, V., SILVA, F. J. and LATORRE, A. Plasmids in the aphid endosymbiont *Buchnera aphidicola* with the smallest genomes. A puzzling evolutionary story. *Gene*, 2006.

GIL, R., SILVA, F. J., PERETO, J. and MOYA, A. Determination of the core of a minimal bacterial gene set. *Microbiology and Molecular Biology Reviews*, 2004, **68**(3): 518-37.

GOODRICH, J. A., SCHWARTZ, M. L. and MCCLURE, W. R. Searching for and predicting the activity of sites for DNA binding proteins: compilation and analysis of the binding sites for *Escherichia coli* integration host factor (IHF). *Nucleic Acids Res*, 1990, **18**(17): 4993-5000.

GOTTESMAN, S. Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet*, 2005, **21**(7): 399-404.

GRANTHAM, R., GAUTIER, C., GOUY, M., JACOBZONE, M. and MERCIER, R. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res*, 1981, **9**(1): r43-74.

GROSS, C. A. Function and regulation of the heat shock proteins. In: F. C. NEIDHARDT. *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. Washington, DC, American Society for Microbiology Press,. **1**, 1996, 1382-1399.

HALLIN, P., STÆRFELDT, H., ROTENBERG, E. and BINNEWIES, T. GeneWiz browser: An Interactive Tool for Visualizing Sequenced Chromosomes. *Standards in Genomic Sciences*, 2009, **1**(2): 204-215.

HARADA, H., OYAIZU, H. and ISHIKAWA, H. A consideration about the origin of aphid intracellular symbiont in connection with gut bacterial flora. *Journal of General and Applied Microbiology*, 1996, **42**(1): 17-26.

HARDY, C. D. and COZZARELLI, N. R. A genetic selection for supercoiling mutants of *Escherichia coli* reveals proteins implicated in chromosome structure. *Mol Microbiol*, 2005, **57**(6): 1636-1652.

HASSAN, M. and CALLADINE, C. Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *Journal of molecular biology*, 1996, **259**: 95-103.

HAUSSER, J. (2006). Improving Entropy Estimation and the Inference of Genetic Regulatory Networks. Munich (Germany), INSA Dépt Biosciences/Department of Statistics, University of Munich**:** 33 pages.

HEDDI, A., CHARLES, H., KHATCHADOURIAN, C., BONNOT, G. and NARDON, P. Molecular characterization of the principal symbiotic bacteria of the weevil *Sitophilus oryzae*: a peculiar G + C content of an endocytobiotic DNA. *J Mol Evol*, 1998, **47**: 52-61.

HENGEN, P. N., BARTRAM, S. L., STEWART, L. E. and SCHNEIDER, T. D. Information analysis of Fis binding sites. *Nucleic Acids Res*, 1997, **25**(24): 4994-5002.

HERBECK, J. T., DEGNAN, P. H. and WERNEGREEN, J. J. Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the enterobacteriales (γ-Proteobacteria). *Mol Biol Evol*, 2005, **22**(3): 520-32.

HERSHBERG, R. and MARGALIT, H. Co-evolution of transcription factors and their targets depends on mode of regulation. *Genome Biol*, 2006, **7**(7): R62.

HERSHBERG, R., YEGER-LOTEM, E. and MARGALIT, H. Chromosomal organization is shaped by the transcription regulatory network. *Trends Genet*, 2005, **21**(3): 138-42.

HIETER, P. and BOGUSKI, M. Functional genomics: it's all how you read it. *Science*, 1997, **278**(5338): 601-2.

HIMMELREICH, R., PLAGENS, H., HILBERT, H., REINER, B. and HERRMANN, R. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. *Nucleic Acids Res*, 1997, **25**(4): 701-12.

HOPFNER, K. P., KARCHER, A., SHIN, D. S., CRAIG, L., ARTHUR, L. M., CARNEY, J. P. and TAINER, J. A. Structural biology of Rad50 ATPase: ATP-driven conformational control in DNA double-strand break repair and the ABC-ATPase superfamily. *Cell*, 2000, **101**(7): 789-800.

HULLO, M. F., AUGER, S., SOUTOURINA, O., BARZU, O., YVON, M., DANCHIN, A. and MARTIN-VERSTRAETE, I. Conversion of methionine to cysteine in *Bacillus subtilis* and its regulation. *J Bacteriol*, 2007, **189**(1): 187-97.

HUMPHREYS, N. J. and DOUGLAS, A. E. Partitioning of symbiotic bacteria between generations of insect: a quantitative study of a *Buchnera* sp. in the pea aphid (*Acyrthosiphon pisum*) reared at different temperatures. *Appl Environ Microbiol*, 1997, **63**(8): 3294-3296.

IAGC. Genome sequence of the pea aphid *Acyrthosiphon pisum*. *Plos Biol*, 2010, **8**(2): e1000313.

ISHIKAWA, H. Insect Symbiosis : An introduction. In: K. BOURTZIS and T. A. MILLER. Insect Symbiosis. Boca Raton, Florida, CRC Press. **Chapter 1**, 2003, 1-21.

JAFFE, A., VINELLA, D. and D'ARI, R. The *Escherichia coli* histone-like protein HU affects DNA initiation, chromosome partitioning via MukB, and cell division via MinCDE. *J Bacteriol*, 1997, **179**(11): 3494-9.

JEONG, K. S., AHN, J. and KHODURSKY, A. B. Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. *Genome Biol*, 2004, **5**(11): R86.

JOHNSON, R. C., JOHNSON, L. M., SCHMIDT, J. W. and GARDNER, J. F. Major nucleoid proteins in the structure and function of the *Escherichia coli* chromosome. In: N. P. HIGGINS. The Bacterial Chromosome Washington, ASM Press, 2005, 65-131.

JORDAN, M. and WEISS, Y. Graphical models: Probabilistic inference. *The Handbook of Brain Theory and Neural Network*, 2002.

KANO, Y. and IMAMOTO, F. Requirement of integration host factor (IHF) for growth of *Escherichia coli* deficient in HU protein. *Gene*, 1990, **89**(1): 133-7.

KAWULA, T. H. and ORNDORFF, P. E. Rapid site-specific DNA inversion in *Escherichia coli* mutants lacking the histonelike protein H-NS. *J Bacteriol*, 1991, **173**(13): 4116-23.

KIDO, M., YAMANAKA, K., MITANI, T., NIKI, H., OGURA, T. and HIRAGA, S. RNase E polypeptides lacking a carboxyl-terminal half suppress a mukB mutation in Escherichia coli. *J Bacteriol*, 1996, **178**(13): 3917-25.

KIERZEK, A. M. STOCKS: STOChastic Kinetic Simulations of biochemical systems with Gillespie algorithm. *Bioinformatics*, 2002, **18**(3): 470-81.

KINGSFORD, C. L., AYANBULE, K. and SALZBERG, S. L. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol*, 2007, **8**(2): R22.

KLASSON, L. and ANDERSSON, S. G. Strong asymmetric mutation bias in endosymbiont genomes coincide with loss of genes for replication restart pathways. *Mol Biol Evol*, 2006, **23**(5): 1031-9.

KNIBBE, C., COULON, A., MAZET, O., FAYARD, J.-M. and BESLON, G. A long-term evolutionary pressure on the amount of noncoding DNA. *Mol Biol Evol*, 2007a, **24**(10): 2344-53.

KNIBBE, C., MAZET, O., CHAUDIER, F., FAYARD, J.-M. and BESLON, G. Evolutionary coupling between the deleteriousness of gene mutations and the amount of non-coding sequences. *J Theor Biol*, 2007b, **244**(4): 621-30.

KOBRYN, K., LAVOIE, B. D. and CHACONAS, G. Supercoiling-dependent site-specific binding of HU to naked Mu DNA. *J Mol Biol*, 1999, **289**(4): 777-84.

KOCH, C., NINNEMANN, O., FUSS, H. and KAHMANN, R. The N-terminal part of the E.coli DNA binding protein FIS is essential for stimulating site-specific DNA inversion but is not required for specific DNA binding. *Nucleic Acids Res*, 1991, **19**(21): 5915-22.

KOGA, R., TSUCHIDA, K. and FUKATSU, T. Changing partners in an obligate symbiosis: a facultative endosymbiont can compensate for loss of the essential endosymbiont *Buchnera* in an aphid. *Proceedings of the Royal Society of London, Series B Biological Sciences*, 2003, **270**(1533): 2543-50.

KOMAKI, K. and ISHIKAWA, H. Intracellular bacterial symbionts of aphids possess many genomic copies per bacterium. *J Mol Evol*, 1999, **48**(6): 717-722.

KOMAKI, K. and ISHIKAWA, H. Genomic copy number of intracellular bacterial symbionts of aphids varies in response to developmental stage and morph of their host. *Insect Biochem Molec Biol*, 2000, **30**(3): 253-258.

KUBORI, T., MATSUSHIMA, Y., NAKAMURA, D., URALIL, J., LARA-TEJERO, M., SUKHAN, A., GAL·N, J. E. and AIZAWA, S. I. Supramolecular structure of the *Salmonella typhimurium* type III protein secretion system. *Science*, 1998, **280**(5363): 602--605.

LACOUR, S., KOLB, A. and LANDINI, P. Nucleotides from -16 to -12 determine specific promoter recognition by bacterial sigmaS-RNA polymerase. *J Biol Chem*, 2003, **278**(39): 37160-8.

LAI, C. Y., BAUMANN, L. and BAUMANN, P. Amplification of *trpEG* - adaptation of *Buchnera aphidicola* to an endosymbiotic association with aphids. *Proceedings of the National Academy of Sciences of the USA*, 1994, **91**(9): 3819-3823.

LANG, B., BLOT, N., BOUFFARTIGUES, E., BUCKLE, M., GEERTZ, M., GUALERZI, C. O., MAVATHUR, R., MUSKHELISHVILI, G., PON, C. L., RIMSKY, S., STELLA, S., BABU, M. M. and TRAVERS, A. High-affinity DNA binding sites for H-NS provide a molecular basis for selective silencing within proteobacterial genomes. *Nucleic Acids Res*, 2007, **35**(18): 6330-7.

LATORRE, A., GIL, R., SILVA, F. J. and MOYA, A. Chromosomal stasis versus plasmid plasticity in aphid endosymbiont *Buchnera aphidicola*. *Heredity*, 2005, **95**(5): 339-47.

LAWRENCE, J. G. Catalyzing bacterial speciation: correlating lateral transfer with genetic headroom. *Syst Biol*, 2001, **50**(4): 479-96.

LAWRENCE, J. G. and ROTH, J. R. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, 1996, **143**(4): 1843-60.

LEWIS, P. J., THAKER, S. D. and ERRINGTON, J. Compartmentalization of transcription and translation in *Bacillus subtilis*. *EMBO J*, 2000, **19**(4): 710-8.

LI, S. and WATERS, R. *Escherichia coli* strains lacking protein HU are UV sensitive due to a role for HU in homologous recombination. *J Bacteriol*, 1998, **180**(15): 3750-6.

LIADOUZE, I., FEBVAY, G., GUILLAUD, J. and BONNOT, G. Metabolic fate of energetic amino acids in the aposymbiotic pea aphid *Acyrthosiphon pisum* (Harris) (Homoptera: Aphididae). *Symbiosis*, 1996, **21**: 115-127.

LINDOW, J. C., BRITTON, R. A. and GROSSMAN, A. D. Structural maintenance of chromosomes protein of *Bacillus subtilis* affects supercoiling in vivo. *J Bacteriol*, 2002, **184**(19): 5317-22.

LOBRY, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol*, 1996, **13**(5): 660-5.

LOBRY, J. R. and SUEOKA, N. Asymmetric directional mutation pressures in bacteria. *Genome Biol*, 2002, **3**(10): RESEARCH0058.

LOZADA-CHÁVEZ, I., ANGARICA, V. E., COLLADO-VIDES, J. and CONTRERAS-MOREIRA, B. The role of DNA-binding specificity in the evolution of bacterial regulatory networks. *J Mol Biol*, 2008, **379**(3): 627-43.

LUIJSTERBURG, M., NOOM, M., WUITE, G. and DAME, R. The architectural role of nucleoid-associated proteins in the organization of bacterial chromatin: a molecular perspective. *Journal of Structural Biology*, 2006, **156**(2): 262-272.

MADAN BABU, M., TEICHMANN, S. A. and ARAVIND, L. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol*, 2006, **358**(2): 614-33.

MAEZAWA, K., SHIGENOBU, S., TANIGUCHI, H., KUBO, T., AIZAWA, S. and MORIOKA, M. Hundreds of flagellar basal bodies cover the cell surface of the endosymbiotic bacterium *Buchnera aphidicola sp. strain APS*. *Journal of Bacteriology*, 2006, **188**(18): 6539-43.

MAGWENE, P. M. and KIM, J. Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol*, 2004, **5**(12): R100.

MAJANDER, K., ANTON, L., ANTIKAINEN, J., LANG, H., BRUMMER, M., KORHONEN, T. K. and WESTERLUND-WIKSTROM, B. Extracellular secretion of polypeptides using a modified *Escherichia coli* flagellar secretion apparatus. *Nature Biotechnology*, 2005, **23**(4): 475-81.

MARGOLIN, A. A., NEMENMAN, I., BASSO, K., WIGGINS, C., STOLOVITZKY, G., DALLA FAVERA, R. and CALIFANO, A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 2006, **7 Suppl 1**: S7.

MARGULIS, L. and FESTER, R. Symbiosis as a source of evolutionnary innovation. Speciation and Morphogenesis. Cambridge, MIT Press, 1991.

MARR, C., GEERTZ, M., HUTT, M. T. and MUSKHELISHVILI, G. Dissecting the logical types of network control in gene expression profiles. *BMC Syst Biol*, 2008, **2**: 18.

MAURER, S., FRITZ, J. and MUSKHELISHVILI, G. A systematic in vitro study of nucleoprotein complexes formed by bacterial nucleoid-associated proteins revealing novel types of DNA organization. *J Mol Biol*, 2009, **387**(5): 1261-76.

MAYO, A. E., SETTY, Y., SHAVIT, S., ZASLAVER, A. and ALON, U. Plasticity of the cis-regulatory input function of a gene. *Plos Biol*, 2006, **4**(4): e45.

MCLEOD, S. M., XU, J., CRAMTON, S. E., GAAL, T., GOURSE, R. L. and JOHNSON, R. C. Localization of amino acids required for Fis to function as a class II transcriptional activator at the RpoS-dependent proP P2 promoter. *J Mol Biol*, 1999, **294**(2): 333-46.

MELBY, T. E., CIAMPAGLIO, C. N., BRISCOE, G. and ERICKSON, H. P. The symmetrical structure of structural maintenance of chromosomes (SMC) and MukB proteins: long, antiparallel coiled coils, folded at a flexible hinge. *J Cell Biol*, 1998, **142**(6): 1595-604.

MEMBRILLO-HERNÁNDEZ, J., KWON, O., DE WULF, P., FINKEL, S. E. and LIN, E. C. Regulation of *adhE* (encoding ethanol oxidoreductase) by the Fis protein in *Escherichia coli*. *J Bacteriol*, 1999, **181**(23): 7390-3.

MENDES, P. GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Comput Appl Biosci*, 1993, **9**(5): 563-71.

MENDES, P. Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem Sci*, 1997, **22**(9): 361-3.

MENDES, P., SHA, W. and YE, K. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 2003, **19 Suppl 2**: ii122-9.

MENDOZA-VARGAS, A., OLVERA, L., OLVERA, M., GRANDE, R., VEGA-ALVARADO, L., TABOADA, B., JIMENEZ-JACINTO, V., SALGADO, H., JUÁREZ, K., CONTRERAS-MOREIRA, B., HUERTA, A. M., COLLADO-VIDES, J. and MORETT, E. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. *PLoS ONE*, 2009, **4**(10): e7526.

MONTLLOR, C. B., MAXMEN, A. and PURCELL, A. H. Facultative bacterial endosymbionts benefit pea aphids *Acyrthosiphon pisum* under heat stress. *Ecological Entomology*, 2002, **27**(2): 189-195.

MORAN, N. A. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences of the USA*, 1996, **93**(7): 2873-2878.

MORAN, N. A. and BAUMANN, P. Bacterial endosymbionts in animals. *Current Opinion in Microbiology*, 2000, **3**(3): 270-275.

MORAN, N. A. and DEGNAN, P. H. Functional genomics of *Buchnera* and the ecology of aphid hosts. *Molecular Ecology*, 2006, **15**(5): 1251-1261.

MORAN, N. A., DEGNAN, P. H., SANTOS, S. R., DUNBAR, H. E. and OCHMAN, H. The players in a mutualistic symbiosis: insects, bacteria, viruses, and virulence genes. *Proceedings of the National Academy of Sciences of the USA*, 2005a, **102**(47): 16919-26.

MORAN, N. A., DUNBAR, H. E. and WILCOX, J. L. Regulation of Transcription in a Reduced Bacterial Genome: Nutrient-Provisioning Genes of the Obligate Symbiont *Buchnera aphidicola*. *Journal of Bacteriology*, 2005b, **187**(12): 4229-37.

MORAN, N. A., MCLAUGHLIN, H. J. and SOREK, R. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science*, 2009, **323**(5912): 379-382.

MORAN, N. A. and MIRA, A. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biology*, 2001, **2**(12): 54.1-54.12.

MORAN, N. A., MUNSON, M. A., BAUMANN, P. and ISHIKAWA, H. A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. *Proc R Soc Lond B Biol Sci*, 1993, **253**(1337): 167-171.

MORAN, N. A., PLAGUE, G. R., SANDSTROM, J. P. and WILCOX, J. L. A genomic perspective on nutrient provisioning by bacterial symbionts of insects. *Proceedings of the National Academy of Sciences of the USA*, 2003, **3**: 3.

MORAN, N. A., RUSSELL, J. A., KOGA, R. and FUKATSU, T. Evolutionary relationships of three new species of Enterobacteriaceae living as symbionts of aphids and other insects. *Appl Environ Microbiol*, 2005c, **71**(6): 3302-10.

MORENO-HAGELSIEB, G. and COLLADO-VIDES, J. A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, 2002, **18 Suppl 1**: S329-36.

MOSTELLER, F. and WILDER TUKEY, J. Data analysis and regression: a second course in statistics. Addison-Wesley Pub. Co., 1977.

MUNSON, M. A., BAUMANN, P., CLARK, M. A., BAUMANN, L., MORAN, N. A., VOEGTLIN, D. J. and CAMPBELL, B. C. Evidence for the establishment of

aphid-eubacterium endosymbiosis in an ancestor of four aphid families. *Journal of Bacteriology*, 1991a, **173**(20): 6321-6324.

MUNSON, M. A., BAUMANN, P. and KINSEY, M. G. *Buchnera* gen. nov. and *Buchnera aphidicola* sp. nov., a taxon consisting of the mycetocyte-associated, primary endosymbionts of aphids. *International Journal of Systematic Bacteriology*, 1991b, **41**(4): 566-568.

MURAKAMI, K. S., MASUDA, S., CAMPBELL, E. A., MUZZIN, O. and DARST, S. A. Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex. *Science*, 2002a, **296**(5571): 1285-90.

MURAKAMI, K. S., MASUDA, S. and DARST, S. A. Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 A resolution. *Science*, 2002b, **296**(5571): 1280-4.

MURPHY, L. D. and ZIMMERMAN, S. B. Isolation and characterization of spermidine nucleoids from *Escherichia coli*. *Journal of Structural Biology*, 1997a, **119**(3): 321-35.

MURPHY, L. D. and ZIMMERMAN, S. B. Stabilization of compact spermidine nucleoids from *Escherichia coli* under crowded conditions: implications for in vivo nucleoid structure. *Journal of Structural Biology*, 1997b, **119**(3): 336-46.

NAÏM , P. Réseaux bayésiens. Paris, Eyrolles, 2004.

NAKABACHI, A. and ISHIKAWA, H. Differential display of mRNAs related to amino acid metabolism in the endosymbiotic system of aphids. *Insect Biochem Molec Biol*, 1997, **27**(12): 1057-1062.

NAKABACHI, A., YAMASHITA, A., TOH, H., ISHIKAWA, H., DUNBAR, H. E., MORAN, N. A. and HATTORI, M. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science*, 2006, **314**(5797): 267.

NARDON, P., DE LA CHAPELLE, B. and HEDDI, A. Symbiosis in the weevil *Sitophilus oryzae*: physiological integration and evolutionary impact. In: S. SATO, ISHIDA, M., ISHIKAWA, H. Endocytobiology V Proc 5th Int Coll on Endocytobiology & Symbiosis Philips Univ, Uji-Kyoto (JAP), June 23-27 1992. Tubingen (GER), Tubingen University Press, 1993, 95-102.

NARDON, P. and GRENIER, A. M. Symbiose et évolution. *Ann Soc Entomol Fr*, 1993, **29**(2): 113-140.

NARDON, P., LEFÈVRE, C., DELOBEL, B., CHARLES, H. and HEDDI, A. Occurrence of endosymbiosis in Dryophthoridae weevils: Cytological insights into bacterial symbiotic structures. *Symbiosis*, 2002, **33**: 227-241.

NASH, H. A. The HU and IHF proteins: accessory factors for complex protein—DNA assemblies of special interest. In: E. L. A. A. LYNCH. Regulation of Gene Expression. Austin TX, RG Landes Company, 1996, 150-179.

NEMENMAN, I. Information theory, multivariate dependence, and genetic network inference. *Arxiv preprint q-bio*, 2004.

NEMENMAN, I., SHAFEE, F. and BIALEK, W. (2002). Entropy and inference, revisited. Neural Information Processing Systems (NIPS). T. G. DIETTERICH, S. BECKER and Z. GHAHRAMAN, MIT Press.

NIMWEGEN, E. Scaling laws in the functional content of genomes. *Power Laws, Scale-Free Networks and Genome Biology*, 2006: 236-253.

NOV KLAIMAN, T., HOSID, S. and BOLSHOY, A. Upstream curved sequences in *E. coli* are related to the regulation of transcription initiation. *Comput Biol Chem*, 2009, **33**(4): 275-82.

NYSTROM, T. Glucose starvation stimulon of *Escherichia coli*: role of integration host factor in starvation survival and growth phase-dependent protein synthesis. *J Bacteriol*, 1995, **177**(19): 5707-10.

OGATA, H., FUJIBUCHI, W., GOTO, S. and KANEHISA, M. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res*, 2000, **28**(20): 4021-8.

OLIVARES-ZAVALETA, N., JAUREGUI, R. and MERINO, E. Genome analysis of *Escherichia coli* promoter sequences evidences that DNA static curvature plays a more important role in gene transcription than has previously been anticipated. *Genomics*, 2006, **87**(3): 329-37.

OLIVER, K. M., MORAN, N. A. and HUNTER, M. S. Costs and benefits of a superinfection of facultative symbionts in aphids. *Proc Biol Sci*, 2006, **273**(1591): 1273-80.

OLSON, W. K., GORIN, A. A., LU, X. J., HOCK, L. M. and ZHURKIN, V. B. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci USA*, 1998, **95**(19): 11163-8.

ORNSTEIN, R., REIN, R., BREEN, D. and MACELROY, R. An optimized potential function for the calculation of nucleic acid interaction energies I. Base stacking. *Biopolymers*, 1978, **17**(10): 2341-2360.

OSUNA, R., FINKEL, S. E. and JOHNSON, R. C. Identification of two functional regions in Fis: the N-terminus is required to promote Hin-mediated DNA inversion but not lambda excision. *EMBO J*, 1991, **10**(6): 1593-603.

OSUNA, R., LIENAU, D., HUGHES, K. T. and JOHNSON, R. C. Sequence, regulation, and functions of fis in *Salmonella typhimurium*. *J Bacteriol*, 1995, **177**(8): 2021-32.

PAGET, M. S. B. and HELMANN, J. D. Protein family review - The sigma(70) family of sigma factors. *Genome Biology*, 2003, **4**(1): 203.1-203.6.

PAN, C. Q., FINKEL, S. E., CRAMTON, S. E., FENG, J. A., SIGMAN, D. S. and JOHNSON, R. C. Variable structures of Fis-DNA complexes determined by flanking DNA-protein contacts. *J Mol Biol*, 1996, **264**(4): 675-95.

PANINSKI, L. Estimation of entropy and mutual information. *Neural Computation*, 2003.

PANINSKI, L. Estimating entropy on m bins given fewer than m samples. *Information Theory, IEEE Transactions on*, 2004, **50**(9): 2200 - 2203.

PAUL, B. J., BARKER, M. M., ROSS, W., SCHNEIDER, D. A., WEBB, C., FOSTER, J. W. and GOURSE, R. L. DksA: a critical component of the transcription initiation

machinery that potentiates the regulation of rRNA promoters by ppGpp and the initiating NTP. *Cell*, 2004, **118**(3): 311-22.

PAULL, T. T., HAYKINSON, M. J. and JOHNSON, R. C. HU and functional analogs in eukaryotes promote Hin invertasome assembly. *Biochimie*, 1994, **76**(10-11): 992-1004.

PEDERSEN, A. G., JENSEN, L. J., BRUNAK, S., STAERFELDT, H. H. and USSERY, D. W. A DNA structural atlas for *Escherichia coli*. *J Mol Biol*, 2000, **299**(4): 907-30.

PEREZ-BROCAL, V., GIL, R., RAMOS, S., LAMELAS, A., POSTIGO, M., MICHELENA, J. M., SILVA, F. J., MOYA, A. and LATORRE, A. A small microbial genome: the end of a long symbiotic relationship? *Science*, 2006, **314**(5797): 312-313.

PETER, B. J., ARSUAGA, J., BREIER, A. M., KHODURSKY, A. B., BROWN, P. O. and COZZARELLI, N. R. Genomic transcriptional response to loss of chromosomal supercoiling in *Escherichia coli*. *Genome Biol*, 2004, **5**(11): R87.

PHADTARE, S. and INOUYE, M. Sequence-selective interactions with RNA by CspB, CspC and CspE, members of the CspA family of *Escherichia coli*. *Mol Microbiol*, 1999, **33**(5): 1004-14.

PINSON, V., TAKAHASHI, M. and ROUVIERE-YANIV, J. Differential binding of the *Escherichia coli* HU, homodimeric forms and heterodimeric form to linear, gapped and cruciform DNA. *J Mol Biol*, 1999, **287**(3): 485-97.

PLAGUE, G. R., DALE, C. and MORAN, N. A. Low and homogeneous copy number of plasmid-borne symbiont genes affecting host nutrition in *Buchnera aphidicola* of the aphid *Uroleucon ambrosiae*. *Molecular Ecology*, 2003, **12**(4): 1095-1100.

POSTOW, L., HARDY, C. D., ARSUAGA, J. and COZZARELLI, N. R. Topological domain structure of the *Escherichia coli* chromosome. *Genes Dev*, 2004, **18**(14): 1766-79.

PRICE, M. N., ARKIN, A. P. and ALM, E. J. The life-cycle of operons. *PLoS Genet*, 2006, **2**(6): e96.

PRICE, M. N., DEHAL, P. S. and ARKIN, A. P. Orthologous transcription factors in bacteria have different functions and regulate different genes. *PLoS Comput Biol*, 2007, **3**(9): 1739-50.

PRICE, M. N., HUANG, K. H., ALM, E. J. and ARKIN, A. P. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res*, 2005, **33**(3): 880-92.

PRICKETT, M. D., PAGE, M., DOUGLAS, A. E. and THOMAS, G. H. BuchneraBASE: a post-genomic resource for *Buchnera sp. APS*. *Bioinformatics*, 2006, **22**(5): 641-2.

PROVOROV, N., VOROBYOV, N. and ANDRONOV, E. Macro- and microevolution of bacteria in symbiotic systems. *Russian Journal of Genetics*, 2008, **44**(1): 6-20.

PROVOROV, N. A. Molecular basis of symbiogenic evolution: from free-living bacteria towards organelles. *Zh Obshch Biol*, 2005, **66**(5): 371-88.

RAHBÉ, Y., CHARLES, H., CALEVRO, F., SIMON, J. C. and RISPE, C. Unis pour survivre : les pucerons et leurs bactéries symbiotiques. *Biofutur*, 2007, **26**(279): 49-52.

REYMOND, N. (2004). Bioinformatique des puces à ADN et application à l'analyse du transcriptome de *Buchnera aphidicola*. Ecole Doctorale : Evolution, Ecosystèmes, Microbiologie et Modélisation. Lyon, France, Institut National des Sciences Appliquées de Lyon**:** 323 pages.

REYMOND, N., CALEVRO, F., VIÑUELAS, J., MORIN, N., RAHBÉ, Y., FEBVAY, G., LAUGIER, C., DOUGLAS, A. E., FAYARD, J. M. and CHARLES, H. Different levels of transcriptional regulation due to trophic constraints in the reduced genome of *Buchnera aphidicola APS*. *Appl Environ Microb*, 2006, **72**(12): 7760-7766.

RHODIUS, V. A., SUH, W. C., NONAKA, G., WEST, J. and GROSS, C. A. Conserved and variable functions of the sigmaE stress response in related genomes. *Plos Biol*, 2006, **4**(1): e2.

RIBEIRO, A. S. and LLOYD-PRICE, J. SGN Sim, a stochastic genetic networks simulator. *Bioinformatics*, 2007, **23**(6): 777-9.

RICE, P. A., YANG, S., MIZUUCHI, K. and NASH, H. A. Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. *Cell*, 1996, **87**(7): 1295-306.

RICHMOND, C. S., GLASNER, J. D., MAU, R., JIN, H. and BLATTNER, F. R. Genome-wide expression profiling in *Escherichia coli K-12*. *Nucleic Acids Res*, 1999, **27**(19): 3821-35.

RILEY, M. Functions of the gene products of *Escherichia coli*. *Microbiol Rev*, 1993, **57**(4): 862-952.

RILEY, M. Genes and proteins of *Escherichia coli K-12*. *Nucleic Acids Res*, 1998, **26**(1): 54.

RILEY, M., ABE, T., ARNAUD, M. B., BERLYN, M. K., BLATTNER, F. R., CHAUDHURI, R. R., GLASNER, J. D., HORIUCHI, T., KESELER, I. M., KOSUGE, T., MORI, H., PERNA, N. T., PLUNKETT, G., 3RD, RUDD, K. E., SERRES, M. H., THOMAS, G. H., THOMSON, N. R., WISHART, D. and WANNER, B. L. *Escherichia coli K-12*: a cooperatively developed annotation snapshot--2005. *Nucleic Acids Res*, 2006, **34**(1): 1-9.

RISPE, C., DELMOTTE, F., VAN HAM, R. C. and MOYA, A. Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. *Genome Res*, 2004, **14**(1): 44-53.

RISPE, C. and MORAN, N. A. Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection. *The American Naturalist*, 2000, **156**(4): 425-441.

ROCHA, E. P. The replication-related organization of bacterial genomes. *Microbiology*, 2004, **150**(Pt 6): 1609-27.

ROCHA, E. P. C. and DANCHIN, A. Base composition bias might result from competition for metabolic resources. *Trends in Genetics*, 2002, **18**(6): 291-4.

RODIONOV, D. A., GELFAND, M. S. and HUGOUVIEUX-COTTE-PATTAT, N. Comparative genomics of the KdgR regulon in *Erwinia chrysanthemi* 3937 and other gamma-proteobacteria. *Microbiology*, 2004, **150**(Pt 11): 3571-90.

ROGOZIN, I. B., MAKAROVA, K. S., NATALE, D. A., SPIRIDONOV, A. N., TATUSOV, R. L., WOLF, Y. I., YIN, J. and KOONIN, E. V. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Res*, 2002, **30**(19): 4264-71.

ROMEO, T. Global regulation by the small RNA-binding protein CsrA and the non-coding RNA molecule CsrB. *Mol Microbiol*, 1998, **29**(6): 1321-30.

ROMERO, P. R. and KARP, P. D. Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases. *Bioinformatics*, 2004, **20**(5): 709-17.

ROSS, W., ERNST, A. and GOURSE, R. L. Fine structure of *E. coli* RNA polymerase-promoter interactions: alpha subunit binding to the UP element minor groove. *Genes Dev*, 2001, **15**(5): 491-506.

ROUVIÈRE-YANIV, J., YANIV, M. and GERMOND, J. E. *E. coli* DNA binding protein HU forms nucleosomelike structure with circular double-stranded DNA. *Cell*, 1979, **17**(2): 265-74.

RYAN, V. T., GRIMWADE, J. E., NIEVERA, C. J. and LEONARD, A. C. IHF and HU stimulate assembly of pre-replication complexes at *Escherichia coli oriC* by two different mechanisms. *Mol Microbiol*, 2002, **46**(1): 113-24.

SABATTI, C., ROHLIN, L., OH, M. K. and LIAO, J. C. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res*, 2002, **30**(13): 2886-93.

SABNIS, N. A., YANG, H. and ROMEO, T. Pleiotropic regulation of central carbohydrate metabolism in *Escherichia coli* via the gene *csrA*. *J Biol Chem*, 1995, **270**(49): 29096-104.

SAFO, M. K., YANG, W. Z., CORSELLI, L., CRAMTON, S. E., YUAN, H. S. and JOHNSON, R. C. The transactivation region of the fis protein that controls site-specific DNA inversion contains extended mobile beta-hairpin arms. *EMBO J*, 1997, **16**(22): 6860-73.

SALGADO, H., MORENO-HAGELSIEB, G., SMITH, T. F. and COLLADO-VIDES, J. Operons in *Escherichia coli*: genomic analyses and predictions. *Proc Natl Acad Sci U S A*, 2000, **97**(12): 6652-7.

SANDERSON, A., MITCHELL, J. E., MINCHIN, S. D. and BUSBY, S. J. Substitutions in the *Escherichia coli* RNA polymerase sigma70 factor that affect recognition of extended -10 elements at promoters. *FEBS Lett*, 2003, **544**(1-3): 199-205.

SANTOS, J. M., FREIRE, P., VICENTE, M. and ARRAIANO, C. M. The stationary-phase morphogene bolA from *Escherichia coli* is induced by stress during early stages of growth. *Mol Microbiol*, 1999, **32**(4): 789-798.

SANTOS, J. M., LOBO, M., MATOS, A. P., DE PEDRO, M. A. and ARRAIANO, C. M. The gene bolA regulates dacA (PBP5), dacC (PBP6) and ampC (AmpC), promoting normal morphology in *Escherichia coli*. *Mol Microbiol*, 2002, **45**(6): 1729-40.

SASAKI, T. and ISHIKAWA, H. Production of essential amino acids from glutamate by mycetocyte symbionts of the pea aphid, *Acyrthosiphon pisum*. *Journal of Insect Physiology*, 1995, **41**(1): 41-46.

SCHÄFER, J., OPGEN-RHEIN, R. and STRIMMER, K. Reverse engineering genetic networks using the GeneNet package. *The Newsletter of the R Project*, 2006, **6**/**5**: 50-54.

SCHÄFER, J. and STRIMMER, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol*, 2005, **4**(1).

SCHEWEMMLER, W. Symbiogenesis in insects as a model for morphogenesis, cell differentiation, and speciation In: L. MARGULIS and R. FESTER. Symbiosis as a source of evolutionnary innovation Speciation and Morphogenesis. Cambridge, MIT Press, 1991, 178-204.

SCHLITT, T. and BRAZMA, A. Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, 2007.

SCHMIDT, A. K. and BALIGA, N. S. Prokaryotic systems biology. In: M. AL-RUBEAI and M. FUSSENEGGER. Systems biology, Springer. **5**, 2007, 394-405.

SCHNEIDER, E., BLUNDELL, M. and KENNELL, D. Translation and mRNA decay. *Mol Gen Genet*, 1978, **160**(2): 121-9.

SCHNEIDER, R., LURZ, R., LUDER, G., TOLKSDORF, C., TRAVERS, A. and MUSKHELISHVILI, G. An architectural role of the *Escherichia coli* chromatin protein FIS in organising DNA. *Nucleic Acids Res*, 2001, **29**(24): 5107-14.

SCHNEIDER, R., TRAVERS, A., KUTATELADZE, T. and MUSKHELISHVILI, G. A DNA architectural protein couples cellular physiology and DNA topology in *Escherichia coli*. *Mol Microbiol*, 1999, **34**(5): 953-64.

SCHNEIDER, T. D. Information content of individual genetic sequences. *J Theor Biol*, 1997, **189**(4): 427-41.

SELINGER, D. W., SAXENA, R. M., CHEUNG, K. J., CHURCH, G. M. and ROSENOW, C. Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Res*, 2003, **13**(2): 216-23.

SESHASAYEE, A. S. N., FRASER, G. M., BABU, M. M. and LUSCOMBE, N. M. Principles of transcriptional regulation and evolution of the metabolic system in *E. coli*. *Genome Research*, 2009, **19**(1): 79-91.

SHEARWIN, K. E., CALLEN, B. P. and EGAN, J. B. Transcriptional interference--a crash course. *Trends Genet*, 2005, **21**(6): 339-45.

SHEN, A., KAMP, H. D., GRUNDLING, A. and HIGGINS, D. E. A bifunctional O-GlcNAc transferase governs flagellar motility through anti-repression. *Genes Dev*, 2006, **20**(23): 3283-95.

SHERIDAN, S. D., BENHAM, C. J. and HATFIELD, G. W. Activation of gene expression by a novel DNA structural transmission mechanism that requires supercoiling-induced DNA duplex destabilization in an upstream activating sequence. *J Biol Chem*, 1998, **273**(33): 21298-308.

SHERIDAN, S. D., BENHAM, C. J. and HATFIELD, G. W. Inhibition of DNA supercoiling-dependent transcriptional activation by a distant B-DNA to Z-DNA transition. *J Biol Chem*, 1999, **274**(12): 8169-74.

SHIGENOBU, S., WATANABE, H., HATTORI, M., SASAKI, Y. and ISHIKAWA, H. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature*, 2000, **407**(6800): 81-86.

SHINDO, H., IWAKI, T., IEDA, R., KURUMIZAKA, H., UEGUCHI, C., MIZUNO, T., MORIKAWA, S., NAKAMURA, H. and KUBONIWA, H. Solution structure of the DNA binding domain of a nucleoid-associated protein, H-NS, from *Escherichia coli*. *FEBS Lett*, 1995, **360**(2): 125-31.

SHINGLETON, A. W., SISK, G. C. and STERN, D. L. Diapause in the pea aphid (*Acyrthosiphon pisum*) is a slowing but not a cessation of development. *BMC Developmental Biology*, 2003, **3**: 7.

SHPIGELMAN, E. S., TRIFONOV, E. N. and BOLSHOY, A. CURVATURE: software for the analysis of curved DNA. *Comput Appl Biosci*, 1993, **9**(4): 435-40.

SILVA, F. J., LATORRE, A. and MOYA, A. Genome size reduction through multiple events of gene disintegration in *Buchnera APS*. *Trends in Genetics*, 2001, **17**(11): 615-8.

SILVA, F. J., LATORRE, A. and MOYA, A. Why are the genomes of endosymbiotic bacteria so stable? *Trends in Genetics*, 2003, **19**(4): 176-80.

SILVA, F. J., VANHAM, R. C. H. J., SABATER, B. and 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, **168**(1): 43-49.

SIMON, J. C., JAUBERT, S., RISPE, C. and TAGU, D. (2007). La vie sexuée et asexuée des pucerons. Biofutur. **26:** 53-56.

SINDEN, R., PEARSON, C. and POTAMAN, V. DNA: structure and function. *Advances in Genome Biology*, 1998, **5A**: 1-141.

SINDEN, R. R. and PETTIJOHN, D. E. Chromosomes in living *Escherichia coli* cells are segregated into domains of supercoiling. *Proc Natl Acad Sci USA*, 1981, **78**(1): 224-8.

SRIVATSAN, A. and WANG, J. D. Control of bacterial transcription, translation and replication by (p)ppGpp. *Curr Opin Microbiol*, 2008, **11**(2): 100-5.

STEPANOVA, E., LEE, J., OZEROVA, M., SEMENOVA, E., DATSENKO, K., WANNER, B. L., SEVERINOV, K. and BORUKHOV, S. Analysis of promoter targets for *Escherichia coli* transcription elongation factor GreA *in vivo* and *in vitro*. *J Bacteriol*, 2007, **189**(24): 8772-85.

STEUER, R., KURTHS, J., DAUB, C. O., WEISE, J. and SELBIG, J. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 2002, **18 Suppl 2**: S231-40.

STOCK, A. M., ROBINSON, V. L. and GOUDREAU, P. N. Two-component signal transduction. *Annu Rev Biochem*, 2000, **69**: 183-215.

STOLL, S., FELDHAAR, H. and GROSS, R. Promoter characterization in the AT-rich genome of the obligate endosymbiont "*Candidatus Blochmannia floridanus*". *J Bacteriol*, 2009, **191**(11): 3747-51.

STRONG, M., MALLICK, P., PELLEGRINI, M., THOMPSON, M. J. and EISENBERG, D. Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach. *Genome Biol*, 2003, **4**(9): R59.

STRUNNIKOV, A. V. SMC complexes in bacterial chromosome condensation and segregation. *Plasmid*, 2006, **55**(2): 135-44.

SWINGER, K. K., LEMBERG, K. M., ZHANG, Y. and RICE, P. A. Flexible DNA bending in HU-DNA cocrystal structures. *EMBO J*, 2003, **22**(14): 3749-60.

TAMAS, I., KLASSON, L., CANBACK, B., NASLUND, A. K., ERIKSSON, A. S., WERNEGREEN, J. J., SANDSTROM, J. P., MORAN, N. A. and ANDERSSON, S. G. E. 50 million years of genomic stasis in endosymbiotic bacteria. *Science*, 2002, **296**(5577): 2376-2379.

TAMAS, I., WERNEGREEN, J. J., NYSTEDT, B., KAUPPINEN, S. N., DARBY, A. C., GOMEZ-VALERO, L., LUNDIN, D., POOLE, A. M. and ANDERSSON, S. G. Endosymbiont gene functions impaired and rescued by polymerase infidelity at poly(A) tracts. *Proc Natl Acad Sci U S A*, 2008, **105**(39): 14934-9.

TJADEN, B., HAYNOR, D. R., STOLYAR, S., ROSENOW, C. and KOLKER, E. Identifying operons and untranslated regions of transcripts using *Escherichia coli* RNA expression analysis. *Bioinformatics*, 2002, **18 Suppl 1**: S337-44.

TOFT, C. and FARES, M. A. The evolution of the flagellar assembly pathway in endosymbiotic bacterial genomes. *Mol Biol Evol*, 2008, **25**(9): 2069--2076.

TOFT, C. and FARES, M. A. Selection for Translational Robustness in *Buchnera aphidicola*, Endosymbiotic Bacteria of Aphids. *Mol Biol Evol*, 2009.

TOMII, K. and KANEHISA, M. A comparative analysis of ABC transporters in complete microbial genomes. *Genome Res*, 1998, **8**(10): 1048-1059.

TRAVERS, A. and MUSKHELISHVILI, G. Bacterial chromatin. *Curr Opin Genet Dev*, 2005a, **15**(5): 507-14.

TRAVERS, A. and MUSKHELISHVILI, G. DNA supercoiling - a global transcriptional regulator for enterobacterial growth? *Nat Rev Microbiol*, 2005b, **3**(2): 157-69.

TSUCHIDA, T., KOGA, R. and FUKATSU, T. Host plant specialization governed by facultative symbiont. *Science*, 2004, **303**(5666): 1989.

USSERY, D., LARSEN, T. S., WILKES, K. T., FRIIS, C., WORNING, P., KROGH, A. and BRUNAK, S. Genome organisation and chromatin structure in *Escherichia coli*. *Biochimie*, 2001, **83**(2): 201-12.

VALENS, M., PENAUD, S., ROSSIGNOL, M., CORNET, F. and BOCCARD, F. Macrodomain organization of the *Escherichia coli* chromosome. *EMBO J*, 2004, **23**(21): 4330-41.

VAN DEN BULCKE, T., VAN LEEMPUT, K., NAUDTS, B., VAN REMORTEL, P., MA, H., VERSCHOREN, A., DE MOOR, B. and MARCHAL, K. SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 2006, **7**: 43.

VAN HAM, R. C., KAMERBEEK, J., PALACIOS, C., RAUSELL, C., ABASCAL, F., BASTOLLA, U., FERNANDEZ, J. M., JIMENEZ, L., POSTIGO, M., SILVA, F. J., TAMAMES, J., VIGUERA, E., LATORRE, A., VALENCIA, A., MORAN, F. and MOYA, A. Reductive genome evolution in *Buchnera aphidicola*. *Proceedings of the National Academy of Sciences of the USA*, 2003, **100**(2): 581-586.

VAN HIJUM, S. A. F. T., MEDEMA, M. H. and KUIPERS, O. P. Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. *Microbiol Mol Biol Rev*, 2009, **73**(3): 481-509, Table of Contents.

VAN HOEK, M. J. A. and HOGEWEG, P. The role of mutational dynamics in genome shrinkage. *Mol Biol Evol*, 2007, **24**(11): 2485--2494.

VAN NOORT, J., VERBRUGGE, S., GOOSEN, N., DEKKER, C. and DAME, R. T. Dual architectural roles of HU: formation of flexible hinges and rigid filaments. *Proc Natl Acad Sci U S A*, 2004, **101**(18): 6969-74.

VIÑUELAS, J. (2008). Caractérisation des capacités de régulation génétique de la bactérie *Buchnera aphidicola* en liaison avec sa fonction symbiotique chez le puceron *Acyrthosiphon pisum*. Thèse de Doctorat, INSA de Lyon**:** 204.

VIÑUELAS, J., CALEVRO, F., REMOND, D., BERNILLON, J., RAHBE, Y., FEBVAY, G., FAYARD, J. M. and CHARLES, H. Conservation of the links between gene transcription and chromosomal organization in the highly reduced genome of *Buchnera aphidicola*. *Bmc Genomics*, 2007, **8**(1): 143.

VON DASSOW, G. and ODELL, G. M. Design and constraints of the *Drosophila* segment polarity module: robust spatial patterning emerges from intertwined cell state switches. *J Exp Zool*, 2002, **294**(3): 179-215.

VON DOHLEN, C. D. and MORAN, N. A. Molecular data support a rapid radiation of aphids in the Cretaceous and multiple origins of host alternation. *Biological Journal of the Linnean Society*, 2000, **71**(4): 689-717.

WANG, H. and BENHAM, C. J. Superhelical destabilization in regulatory regions of stress response genes. *PLoS Comput Biol*, 2008, **4**(1): e17.

WANG, Y. and DEHASETH, P. L. Sigma 32-dependent promoter activity in vivo: sequence determinants of the groE promoter. *J Bacteriol*, 2003, **185**(19): 5800-6.

WERREN, J. H., ZHANG, W. and GUO, L. R. Evolution and phylogeny of *Wolbachia*: reproductive parasites of arthropods. *Proceedings of the Royal Society of London, Series B Biological Sciences*, 1995, **261**(1360): 55-63.

WILCOX, J. L., DUNBAR, H. E., WOLFINGER, R. D. and MORAN, N. A. Consequences of reductive evolution for gene expression in an obligate endosymbiont. *Mol Microbiol*, 2003, **48**(6): 1491-1500.

WILKINSON, T. L., FUKATSU, T. and ISHIKAWA, H. Transmission of symbiotic bacteria *Buchnera* to parthenogenetic embryos in the aphid *Acyrthosiphon pisum* (Hemiptera: Aphidoidea). *Arthropod Structure and Development*, 2003, **32**: 241-245.

WILLE, A. and BUHLMANN, P. Low-order conditional independence graphs for inferring genetic networks. *Stat Appl Genet Mol Biol*, 2006, **5**: Article1.

XU, J. and JOHNSON, R. C. Identification of genes negatively regulated by Fis: Fis and RpoS comodulate growth-phase-dependent gene expression in *Escherichia coli*. *J Bacteriol*, 1995, **177**(4): 938-47.

YADA, T., NAKAO, M., TOTOKI, Y. and NAKAI, K. Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models. *Bioinformatics*, 1999, **15**(12): 987-93.

YANG, H., LIU, M. Y. and ROMEO, T. Coordinate genetic regulation of glycogen catabolism and biosynthesis in *Escherichia coli* via the CsrA gene product. *J Bacteriol*, 1996, **178**(4): 1012-7.

YANG, W. Z., KO, T. P., CORSELLI, L., JOHNSON, R. C. and YUAN, H. S. Conversion of a beta-strand to an alpha-helix induced by a single-site mutation observed in the crystal structure of Fis mutant Pro26Ala. *Protein Sci*, 1998, **7**(9): 1875-83.

YASUZAWA, K., HAYASHI, N., GOSHIMA, N., KOHNO, K., IMAMOTO, F. and KANO, Y. Histone-like proteins are required for cell growth and constraint of supercoils in DNA. *Gene*, 1992, **122**(1): 9-15.

YOU, L., HOONLOR, A. and YIN, J. Modeling biological systems using Dynetica--a simulator of dynamic networks. *Bioinformatics*, 2003, **19**(3): 435-6.

YOUNG, G. M., SMITH, M. J., MINNICH, S. A. and MILLER, V. L. The *Yersinia enterocolitica* motility master regulatory operon, *flhDC*, is required for flagellin production, swimming motility, and swarming motility. *J Bacteriol*, 1999, **181**(9): 2823--2833.

ZECHIEDRICH, E. L., KHODURSKY, A. B., BACHELLIER, S., SCHNEIDER, R., CHEN, D., LILLEY, D. M. and COZZARELLI, N. R. Roles of topoisomerases in maintaining steady-state DNA supercoiling in *Escherichia coli*. *J Biol Chem*, 2000, **275**(11): 8103-13.

ZHENG, Y., SZUSTAKOWSKI, J. D., FORTNOW, L., ROBERTS, R. J. and KASIF, S. Computational identification of operons in microbial genomes. *Genome Res*, 2002, **12**(8): 1221-30.

ZIENTZ, E., DANDEKAR, T. and GROSS, R. Metabolic interdependence of obligate intracellular bacteria and their insect hosts. *Microbiology and Molecular Biology Reviews*, 2004, **68**(4): 745-70.

ZIENTZ, E., SILVA, F. J. and GROSS, R. Genome interdependence in insect-bacterium symbioses. *Genome Biology*, 2001, **2**(12): 321-326.

Annexes

Tableau A1. La liste des unites de transcription de *Buchnera APS* prédites avec DisTer.

Unité de transcription	Gènes
1	*gidA*
2	*atpB, atpE, atpF, atpH, atpA, atpG, atpD, atpC*
3	*gyrB, dnaN, dnaA*
4	*rpmH, rnpA*
5	*yidC, thdF*
6	*tRNA-Phe-GAA*
7	*groES, mopA*
8	*efp*
9	*dnaC, dnaT, yhhF*
10	*ftsY*
11	*rpoH*
12	*glmS, glmU*
13	*yigL*
14	*metR*
15	*metE*
16	*purH*
17	*hupA*
18	*rpoC, rpoB*
19	*rplL, rplJ*
20	*rplA, rplK, nusG, secE*
21	*tRNA-Thr-GGT, tRNA-Gly-TCC, tRNA-Tyr-GTA, tRNA-Thr-TGT*
22	*murB*
23	*metF*
24	*argE*
25	*argC, argB, argG, argH, yibN*
26	*secB*
27	*cysE*
28	*rpoD*
29	*dnaG, rpsU*
30	*ygjD*
31	*ribB*
32	*rfaE*
33	*cca*
34	*uppP, crr, ptsI*
35	*ptsH*
36	*cysK*
37	*lig*
38	*tRNA-Lys-TTT, tRNA-Val-TAC*
39	*gltX*
40	*tRNA-Ala-GGC*
41	*fliE*
42	*fliF, fliG, fliH, fliI, fliJ, fliK, fliM, fliN, fliOP, fliQ, fliR*
43	*rpmG, rpmB*
44	*ytfN*

45 *ppa*
46 *pmbA*
47 *rnpB, yraL*
48 *fabB*
49 *talA, tktB*
50 *dapE*
51 *dapA*
52 *aroC*
53 *yfcN*
54 *hisG, hisD, hisC, hisB, hisH, hisA, hisF, hisI*
55 *gnd*
56 *dcd*
57 *metG*
58 *mesJ, tRNA-Val-GAC*
59 *ribE, rnfA, rnfB*
60 *rnfC, rnfD*
61 *rnfG, ydgQ, nth, priA*
62 *tyrS*
63 *sufA*
64 *ydiK*
65 *aroH*
66 *thrS, infC, rpmI, rplT*
67 *pheS, pheT, himA, queA, tgt, yajC*
68 *glyS, glyQ*
69 *folE*
70 *nfo, rplY*
71 *yabI*
72 *surA, ksgA, apaH*
73 *folA, carB, carA*
74 *dapB, lytB, lspA, ileS, ribF*
75 *rpsT*
76 *dnaJ, dnaK*
77 *nuoA, nuoB, nuoCD, nuoE, nuoF, nuoG, nuoH, nuoI, nuoJ, nuoK, nuoL, nuoM, nuoN*
78 *folC, cvpA*
79 *prsA*
80 *ispE*
81 *prfA, hemK*
82 *ychA*
83 *nadE*
84 *ackA, pta*
85 *yfaE, nrdB, nrdA, gyrA*
86 *yba2*
87 *ahpC*
88 *ung*
89 *grpE*
90 *nadK*
91 *smpA*
92 *ssrA*

93 *ydhD*
94 *rnt*
95 *sodA*
96 *pth, ychF*
97 *thrC, thrB, thrA*
98 *hpt*
99 *panC, panB, dksA*
100 *truA, mrcB*
101 *secA, mutT*
102 *coaE*
103 *guaC*
104 *aceE, aceF, lpdA*
105 *speD, speE*
106 *pfs, yadR*
107 *ftsZ, ftsA, ddlB*
108 *murC, murG, ftsW, murD, mraY, murF, murE, ftsI, ftsL, yabC, ilvH, ilvI*
109 *apbE*
110 *htrA*
111 *dapD, map*
112 *rpsB, tsf, pyrH, frr, dxr, uppS*
113 *yaeT, dnaE*
114 *proS*
115 *flhB, flhA*
116 *argS*
117 *rrs*
118 *tRNA-Ile-GAT, tRNA-Ala-TGC*
119 *gloB, rnhA*
120 *dnaQ, tRNA-Asp-GTC, lpcA*
121 *gpt*
122 *grpE1*
123 *yfjF*
124 *smpB, yfhC*
125 *acpS, era, rnc*
126 *lepB, lepA*
127 *trmU, ycfC, purB, mltE, fabI*
128 *rnb*
129 *ychE, lipB, lipA*
130 *pyrF, ribA*
131 *hns*
132 *cls*
133 *yciA, yciB, yciC*
134 *trpA, trpB, trpC, trpD*
135 *yedA, rluB, sohB, topA*
136 *suhB*
137 *yfgB, gcpE, hisS, glyA*
138 *bioD*
139 *bioB*
140 *bioA*

141 *pgl, mfd*
142 *lolC, lolD, lolE, gapA*
143 *fldA*
144 *phrB, ybgI, sucA, sucB, gpmA*
145 *pfkA, glpF*
146 *tpiA, himD, rpsA*
147 *cmk, aroA, serC, serS*
148 *trxB*
149 *infA*
150 *aspS*
151 *znuB, znuC*
152 *pykA*
153 *zwf*
154 *htpX*
155 *cspC*
156 *yoaE, yeaZ*
157 *minE, minD, minC*
158 *rsmC*
159 *tRNA-Leu-TAA, tRNA-Cys-GCA*
160 *tRNA-Ser-TGA*
161 *ompA*
162 *mviN*
163 *pyrC*
164 *flgN, flgA*
165 *flgB, flgC, flgD, flgE, flgF, flgG, flgH*
166 *flgI, flgJ*
167 *flgK*
168 *rne*
169 *rluC, rpmF*
170 *fabD, fabG, acpP, tmk, holB, ycfH, ptsG, ycfF, ycfM*
171 *ompF*
172 *asnS*
173 *pncB*
174 *pyrD*
175 *ycbY, uup*
176 *yceA, valS, pepA*
177 *argF, pyrB, pyrI, yhaR*
178 *deaD*
179 *pnp*
180 *rpsO*
181 *truB, rbfA, infB, nusA*
182 *tRNA-Leu-GAG, secG*
183 *mrsA*
184 *hflB, ftsJ, greA*
185 *yrbA, murA*
186 *rplU, rpmA*
187 *yhbZ*
188 *rpsI, rplM*

189 *pheA*
190 *ffh*
191 *rpsP, rimM, trmD, rplS*
192 *tldD*
193 *aroQ*
194 *fis*
195 *rluD*
196 *yfiO*
197 *alaS*
198 *csrA*
199 *tRNA-Ser-GCT, tRNA-Arg-ACG*
200 *gshA*
201 *metK*
202 *endA, yggJ*
203 *rpiA*
204 *tRNA-Gln-TTG, tRNA-Leu-TAG, tRNA-Met-CAT1*
205 *glnS*
206 *pyrG, eno*
207 *nlpD, ispF, ispD, ftsB, cysC, cysN, cysD, cysG*
208 *cysH, cysI, cysJ*
209 *mutS*
210 *dsbA, polA*
211 *yihA*
212 *typA*
213 *gmk, ygfZ*
214 *prfB, lysS, lysA, lgt, thyA, yleA*
215 *ybeY, ybeX*
216 *leuS, holA, nadD, sirA*
217 *asd*
218 *yhgN, pgk, fba*
219 *mscS, recC, recB, recD*
220 *argA*
221 *tRNA-Met-CAT2*
222 *mltA, ribH, thiL, ribD1, ribD2, nusB*
223 *dxs, ispA, yajR*
224 *yccK*
225 *cyoE, cyoD, cyoC, cyoB, cyoA*
226 *bolA*
227 *tig*
228 *clpP, clpX*
229 *lon*
230 *ppiD*
231 *ybaX*
232 *mdl, mdlB*
233 *ffs*
234 *dnaX*
235 *ybaB*
236 *htpG, adk*

237 *tRNA-Arg-TCT*
238 *folD*
239 *cysS*
240 *ybeD*
241 *cspE*
242 *rrf*
243 *rrl*
244 *tRNA-Glu-TTC*
245 *aroE, yrdC, smg*
246 *def, fmt*
247 *rplQ, rpoA, rpsD, rpsK, rpsM, rpmJ, secY, rplO, rpmD, rpsE, rplR, rplF, rpsH, rpsN, rplE, rplX, rplN, rpsQ, rpmC, rplP, rpsC, rplV, rpsS, rplB, rplW, rplD, rplC, rpsJ*
248 *tufA, fusA, rpsG, rpsL*
249 *yheL, yheM, yheN, fkpA*
250 *argD*
251 *tsgA*
252 *trpS, rpe*
253 *aroB, aroK*
254 *tRNA-Ser-GGA*
255 *deoD, deoB, prfC*
256 *yhgI*
257 *ssb*
258 *dnaB*
259 *gshB, yqgF, yggS, yggW*
260 *yggH*
261 *mutY, yggX, murI*
262 *sbcB*
263 *yeeX*
264 *tRNA-Asn-GTT*
265 *tRNA-Met-CAT3, pyrE*
266 *dut, cysQ, rplI, rpsR, rpsF*
267 *vacB, purA, hflC, hflK*
268 *miaA, mutL, mtlD, mtlA*
269 *pgi*
270 *orn, tRNA-Gly-GCC*
271 *amiB*
272 *rpmE*
273 *hslV, hslU*
274 *ibpA, fpr*
275 *poxA*
276 *kdtB*
277 *yba3, yba4*
278 *yhiQ*
279 *pitA*
280 *ynfM, dapF*
281 *cyaY, hemC*
282 *hemD*
283 *tRNA-Pro-TGG, tRNA-His-GTG, tRNA-Arg-CCG*

284 *rho, trxA*
285 *rep, ilvC, ilvD*
286 *tRNA-Trp-CCA*
287 *iscS, iscU, hscB, hscA, fdx*
288 *der, yfgM*

Tableau A2. La distribution des unités de transcription d'*E. coli*, contenant au moins un gène ayant un orthologue chez *Buchnera APS*, dans les classes d'unités de transcription de *Buchnera APS* (identiques, similaires, fusionnées, fragmentées et réorganisées). Cette distribution apparaît comme significativement différente (test de Chi2, p-value = 0.04), ce résultats s'expliquant par la classe des unités de transcription fragmentées, ayant des faibles effectifs, le test de Chi2 n'est pas significatif lorsqu'on élimine cette classe (p-value = 0.3).

	Identiques	Similaires	Fragmentées	Fusionnées	Réorganisées
Unités de transcription d'*E. coli* régulées par des facteurs de transcription connus	48	14	10	49	12
Unités de transcription d'*E. coli* non régulées par des facteurs de transcription connus	72	40	5	101	21

Tableau A3. Effectifs des gènes de l'ancêtre de *Buchnera* et d'*E. coli* conservés ou pas chez *Buchnera*, régulées ou pas chez *E. coli* par des facteurs de transcription (test de Chi2, p-value = 0.008).

	Conservés chez *Buchnera APS*	Pas conservés chez *Buchnera APS*
Régulés chez *E. coli*	247	831
Non régulés chez *E. coli*	344	1488

FOLIO ADMINISTRATIF

THESE SOUTENUE DEVANT L'INSTITUT NATIONAL DES SCIENCES APPLIQUEES DE LYON

NOM : BRINZA DATE de SOUTENANCE : 08/12/2010
Prénoms : Lilia

TITRE : Exploration et inférence du réseau de régulation de la transcription de la bactérie symbiotique intracellulaire à génome réduit, *Buchnera aphidicola*

NATURE : Doctorat Numéro d'ordre : 2010-ISAL-0102
Ecole doctorale : Evolution, Ecosystèmes, Microbiologie, Modélisation (E2M2)
Spécialité : Méthodes en bioinformatique moléculaire

Cote B.I.U. - Lyon : T 50/210/19 / et bis CLASSE :

RESUME :
Cette thèse est une étude systémique de la régulation de la transcription des gènes de la bactérie *Buchnera aphidicola* vivant en symbiose intracellulaire obligatoire avec le puceron du pois, *Acyrthosiphon pisum*. Avec un génome extrêmement réduit, très riche en bases A et T, enrichi en gènes métaboliques et dénué de régulateurs transcriptionnels, *Buchnera* constitue un modèle bactérien très intriguant du point de vue évolutif et fonctionnel. Plusieurs études expérimentales antérieures sur ce modèle de symbiose attestent d'une part que la bactérie fournit à son hôte le complément nutritionnel qu'il ne trouve pas dans son alimentation, et d'autre part, que la bactérie adapte cette fourniture aux variations de la demande de son hôte. Néanmoins, les mécanismes impliqués dans cette régulation demeuraient relativement obscurs. Nous avons structuré notre analyse de la régulation de la transcription chez *Buchnera* en quatre parties. La première, basée principalement sur des études comparatives (*Buchnera* vs. *Escherichia coli*) dresse l'inventaire de la machinerie transcriptionnelle de *Buchnera*. La deuxième partie analyse l'architecture génomique de *Buchnera*, i.e. l'organisation et l'évolution de sa carte opéronique, l'agencement des fragments synthéniques et non-synthéniques et également les forces d'évolution ayant amené à l'agencement des gènes de *Buchnera* le long de son chromosome. Pour cela, nous avons été amenés à développer une méthode bayésienne de prédiction d'opérons adaptée à *Buchnera*, ce qui nous a permis de proposer une nouvelle carte opéronique de la bactérie que nous avons ensuite partiellement validée par RT-PCR. La troisième partie porte sur les propriétés structurelles séquence-dépendantes du chromosome de *Buchnera* : la courbure intrinsèque (curvature), l'énergie d'empilement des bases (stacking energy), l'angle de torsion (propeller twist) et le SIDD (Stress Induced Duplex Destabilization). Ces propriétés ont ensuite été mises en relation avec le profil périodique de l'expression des gènes le long du chromosome. Les résultats obtenus à l'issue de cette approche ascendante, nous ont amené à construire un premier modèle de réseau de la régulation transcriptionnelle chez *Buchnera*. Enfin, la quatrième partie est un travail de modélisation suivant une approche descendante. Il s'agit du développement d'une méthode d'inférence de réseau de régulation à partir de données d'expression que nous avons appelée IGOIM (Inférence de Graphe de premier Ordre avec l'Information Mutuelle conditionnelle). Cette méthode a été validée sur des jeux de données simulées et de la littérature mais n'a finalement pas été utilisée pour inférer le réseau de *Buchnera*, car actuellement nous ne disposons pas de données d'expression qualitativement et quantitativement suffisantes.

MOTS-CLES : *Buchnera aphidicola*, *Acyrthosiphon pisum*, réseau de régulation génétique, réseau de gènes, transcription, symbiose intracellulaire, carte opéronique

Laboratoire (s) de recherche : Laboratoire de Biologie Fonctionnelle, Insectes et Interactions
UMR 0203 INRA / INSA de Lyon (BF2I)

Directeur de thèse: Hubert Charles, Christian Gautier

Président de jury : Hubert Vidal

Composition du jury :

Président :	H. Vidal
Rapporteurs :	G. Fichant, J. Geiselmann
Examinateurs :	F. Calevro, S. Reverchon-Pescheux
Directeurs :	H. Charles, C. Gautier