Determination of the role of certain bacterial peptidases by inference from heterogeneous and incomplete data

Liliana López Kleine

▶ **To cite this version:**

Liliana López Kleine. Determination of the role of certain bacterial peptidases by inference from heterogeneous and incomplete data. Life Sciences [q-bio]. AgroParisTech, 2008. English. NNT : 2008AGPT0056 . pastel-00004768

HAL Id: pastel-00004768
https://pastel.hal.science/pastel-00004768

Submitted on 13 Feb 2009

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

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

N°/__/__/__/__/__/__/__/__/__/

THÈSE

pour obtenir le grade de

Docteur

de

L'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech)

Spécialité : Biologie et Statistiques Appliquées

présentée et soutenue publiquement
par

Liliana López Kleine

Le 7 octobre 2008

DETERMINATION DU RÔLE DE CERTAINES PEPTIDASES BACTERIENNES PAR INFERENCE A PARTIR DE DONNEES HETEROGENES ET INCOMPLETES

Directeurs de thèse : ***Véronique MONNET*** **et** ***Alain TRUBUIL***

Travail réalisé : INRA , UR477, Unité de Biochimie Bactérienne et UR341, Unité de Mathématiques et Informatique Appliquées, F-78350, Jouy-en-Josas

Devant le jury :

M. Anthony PUGSLEY	Professeur	**Institut Pasteur**	**Rapporteur**
M. Stéphane CANU	Professeur	**INSA**	**Rapporteur**
M. Claude GAILLARDIN	Professeur	**AgroParisTech**	**Examinateur**
M. Jean-Jacques DAUDIN	Professeur	**AgroParisTech**	**Examinateur**
Mme. Véronique MONNET	Directeur de Recherche	**INRA**	**Directeur de thèse**
M. Alain TRUBUIL	Ingénieur de Recherche	**INRA**	**Directeur de thèse**

L'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech) est un Grand Etablissement dépendant du Ministère de l'Agriculture et de la Pêche, composé de l'INA PG, de l'ENGREF et de l'ENSIA (décret n° 2006-1592 du 13 décembre 2006)

L'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech) est un Grand Etablissement dépendant du Ministère de l'Agriculture et de la Pêche, composé de l'INA PG, de l'ENGREF et de l'ENSIA
(décret n° 2006-1592 du 13 décembre 2006)

Détermination du rôle des certaines peptidases bactériennes par inférence à partir de données hétérogènes et incomplètes

Résumé

La détermination du rôle des protéines est à l'origine de la plupart des études biologiques. Elle est encore plus d'actualité depuis l'avènement du séquençage des génomes qui fournit des protéines de rôle inconnu en grande quantité. Le séquençage de nouveaux génomes a permis, par ailleurs, l'obtention de données post-génomiques à haut débit comme les données transciptomique. Il a permis également la construction de données comme les profils phylogénétiques. L'approche que nous avons utilisée, mêlant statistiques et biologie, intègre ces données post-génomiques pour prédire puis valider le rôle de protéines protéolytiques chez *Lactococcus lactis.*

Nous avons procédé en trois étapes réalisant, dans un premier temps, une inférence statistique des liens prédits entre protéines du lactocoque basée sur cinq sources de données distinctes : le graphe des voies métaboliques, des données transcriptomiques, des profils phylogénétiques, des distances entre gènes en paires de bases et des données protéomiques issues de gels 2D. Ces liens nous ont permis de émettre des hypothèses biologiques sur le rôle de protéines cibles. Dans un deuxième temps, nous avons réalisé des expériences de laboratoire pour valider les prédictions comparant la souche sauvage aux mutants de délétion de PepF et YvjB. La troisième étape de ce travail a consisté à réaliser une analyse de sensibilité dans le but de souligner les données qui ont contribuées le plus aux prédictions et de trouver un sous-ensemble de données aboutissant aux mêmes prédictions.

Appliquée à deux enzymes protéolytiques potentielles du lactocoque, PepF et YvjB, cette approche nous a permis de vérifier leur participation dans l'export de protéines telle qu'elle avait été prédite par l'analyse statistique. Nous avons montré que PepF participe à l'export de protéines libérées dans le milieu extérieur (protéines sécrétées) en hydrolysant le peptide signal libéré par la signal peptidase I. PepF participe par ailleurs aussi à la synthèse de la paroi cellulaire et au métabolisme du pyruvate, fonctions aussi prédites par l'analyse statistique. YvjB participe à la mise en place et la maturation d'un autre groupe de protéines exportées, les lipoprotéines. Elle est nécessaire pour la localisation correcte de certaines lipoprotéines et participe au clivage de leur peptide signal.
Notre approche s'est avérée très utile dans la mesure où elle permet d'émettre des hypothèses biologiques qui guident les validations expérimentales. Par ailleurs, cette démarche est applicable à tout organisme entièrement séquencé pour lequel suffisamment de données post-génomiques sont disponibles.

L'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech) est un Grand Etablissement dépendant du Ministère de l'Agriculture et de la Pêche, composé de l'INA PG, de l'ENGREF et de l'ENSIA (décret n° 2006-1592 du 13 décembre 2006)

Determination of the role of certain bacterial peptidases by inference from heterogeneous and incomplete data

Abstract

The determination of the role of proteins is at the origin of most biological studies. It becomes currently more important since the advent of genome sequencing, which provides a great number of proteins of unknown roles. The sequencing of new genomes allows at the same time to obtain high-throughput post-genomic data as for example microarrays data. It allows also the construction of phylogenetic profiles. The approach we have used, couples statistics and biology and integrates this post-genomic data in order to predict and further validate the role of putative proteolytic enzymes of *Lactococcus lactis*.

We worked in three different steps. First, we did a statistical inference of the predicted relationships between proteins of lactococci based on five different data sources: the graph of metabolic pathways, microarray data, phylogenetic profiles and the distance between genes in terms of base pairs on the chromosome and proteomic data from 2D gels. These relationships allowed us to pose biological hypotheses on the role of target proteins. Second, we carried out wet-lab experiments to validate the predictions comparing the wild type strain to knock-out mutants of PepF and YvjB. The third step consisted of a sensitivity analysis with the aim to outline the most important data in terms of contribution to the relationship prediction of our target proteins and to find a subset of data leading to the same predictions.

Applied to two proteolytic enzymes of lactococci, PepF and YvjB, this approach allowed us to validate their participation in protein export predicted by the statistical analysis. We have shown that PepF participates in the export of proteins liberated in the supernatant by the hydrolysis of the signal peptide, which is cleaved by the signal peptidase I. PepF participates also in the synthesis of the cell wall and pyruvate metabolism, two functions also predicted by the statistical analysis. YvjB participates in the localization and maturation of lipoproteins, another group of exported proteins. It is necessary for the correct localization of certain lipoproteins and participates in their signal peptide cleavage.
Our approach has shown to be very useful, as it allows to pose biological hypotheses that guide experimental validations. Moreover, this approach is applicable to all completely sequenced organisms for which enough post-genomic data is available.

L'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech) est un Grand Etablissement dépendant du Ministère de l'Agriculture et de la Pêche, composé de l'INA PG, de l'ENGREF et de l'ENSIA (décret n° 2006-1592 du 13 décembre 2006)

A Andrés
A ma mère

Je tiens tout d'abord à remercier les professeurs Anthony Pugsley, Stéphane Canu, Jean- Claude Gaillardin et Jean-Jacques Daudin d'avoir accepté de participer à l'évaluation de ce travail.

Je remercie Véronique Monnet et Alain Trubuil de m'avoir permis de réaliser cette thèse à cheval entre la biologie et la mathématique. Je les remercie d'avoir fait l'effort d'être disponibles malgré leurs multiples obligations en tant que directeurs des unités BioBac et MIA respectivement. Je me suis toujours sentie encadrée par deux pôles opposés, chacun représentant un monde différent, la biologie - la mathématique, un monde appliqué - un monde abstrait... merci d'avoir formé un tout ! C'était une grande chance pour moi d'avoir eu de si bons encadrants dans les deux domaines qui m'ont toujours bien guidée et appris tant de choses !

Véronique, il n'est pas simple d'exprimer ma gratitude pour ces trois ans. Tout d'abord merci pour la confiance que tu m'as accordée, l'indépendance que tu m'as donnée tout en restant disponible quand j'avais besoin de toi. Je te remercie de m'avoir donné l'opportunité de découvrir le monde de la microbiologie et la biochimie en me donnant accès à toutes les techniques et les experts de l'unité (et du centre de Jouy). Merci également de t'être toujours soucié de mon bien-être personnel.

Alain, je te remercie pour ta patience et le dévouement avec lequel tu m'as expliqué tant de choses concernant les mathématiques et la programmation. Il a été toujours très agréable de travailler avec toi, d'apprendre, de voir que tout était plus facile que je ne le croyais. Merci pour tout le temps que tu as passé avec moi.

Je voudrais remercier à tous mes collègues de BioBac pour m'avoir fait sentir partie de l'unité et m'avoir permis de parcticiper aux discussions biologiques et analyses statistiques de tous. Je les remercie pour tous les bons moments que nous avons passés ensemble à BioBac et en dehors du travail. D'abord je remercie Françoise Wessner de m'avoir accompagnée pendant mes premiers jours à BioBac et de m'avoir tout montré. Je remercie Christophe et Stéphane, qui partageainet lbureau et laboratoire avec moi pendant ces trois ans, pour avoir été toujours disponibles pour m'aider, me montrer comment faire les choses et discuter des résultats frais de chaque manip. Je remercie également Emilie, Pascale, Gaëlle, Alain et Céline pour avoir fait des séquençages, des analyses HPLC et de la spectrométrie de masses. Merci à Mickael pour m'avoir donné tant de bons conseils et bons protocoles. Je voudrais remercier tout spécialement Colette pour tout ce qu'elle fait pour le bon fonctionnement du labo, les conseils qu'elle m'a donnés au cours de tout mon travail pratique et pour avoir sauvé quelques manips que j'avais oubliées, bien sûr aussi pour faire le café tous les jours.
Je remercie toutes les chargées de recherche de BioBac pour leurs bons conseils et les discussions intéressantes que nous avons eues. Je remercie Marie-Pierre pour m'avoir aidé dans la caractérisation du peptidoglycane, Rozenn pour toutes les discussions autour de Eep, des techniques de biologie moléculaire et pour la relecture d'abstracts et de la partie de la thèse concernant Eep (article inclus), Mireille pour la relecture de l'article sur PepF, Françoise pour avoir toujours été là pour m'aider avec mon français et pour partager avec moi ses connaissances. Finalement, je remercie les doctorants de BioBac pour les moments que nous avons partagés, aussi pour les discussions qui me rappelaient bien que je n'étais pas la seule dans la même situation. Luciana, gracias por recibirme con los brazos abiertos y por toda tu ayuda. Qué bueno que pude desahogarme en español de vez en cuando. Jasna, merci pour le temps passé ensemble et pour m'avoir fait rire avec tes histoires. Michael, es war schön, dich kennen zu lernen und mit dir über alles sprechen zu können was hier in Frankreich einfach anders ist als sonst wo auf der

Welt. Chez BOG, nos voisins, je remercie Valérie, Lionel, Florence, Mariam et Claire, pour leur aide, les réactifs et les diverses pièces détachées qui m'ont dépannés plusieurs fois.

Je remercie les gens de MIA, mon autre équipe, pour la bonne ambiance qui règne dans l'unité. Je remercie Kien d'avoir participé dans mon comité de thèse et pour la relecture de l'article sur PepF. Je remercie Jean Pierre et Hervé pour leurs explications sur l'analyse de sensibilité. Je remercie Eric pour sa sympathie, sa bonne énergie, son aide pour résoudre mes problèmes informatiques et pour son invitation à faire de la musique sur le centre. Je remercie tous les doctorants, David, Fanny, Ikhlef, Zaher, Najat, Antoine, Mati, Louise et Valérie pour les moments que nous avons passés ensemble. Rafa, gracias por tu amistad, la buena energía y por ser tan buen interlocutor no sólo para hablar de la modelación matemática de sistemas biológicos, sino también de literatura y de todo el resto ! Qué viva Colombia ! Je remercie tout spécialement Anne et David pour leur aide dans la programmation Matlab. Merci aussi à Caroline pour avoir essayé d'intégrer tous mes petits scripts dans un workflow, pour avoir relu ma revue bibliographique sur les méthodes statistiques et pour avoir été toujours disponible.

A propos de relecteurs je voudrais remercier aussi Aude, Emilie, Vincent Juillard et Andrés pour leurs corrections et commentaires.

Etre à l'INRA de Jouy-en-Josas signifie faire partie d'une grande communauté scientifique et d'avoir la chance d'échanger des connaissances, de l'information, des idées, du matériel, etc. Dans ce sens je voudrais remercier les gens de MIG : Vincent Fromion, Phillippe Bessières et Elodie pour les échanges sur la biologie de systèmes ; Pierre et Valentin pour leur aide à chaque fois que j'avais des questions sur l'annotation de génomes, les alignements, etc. Je remercie aussi les gens de UBLO : Isabelle Poquet et Nicolas pour toutes les discussions autour de PepF et la sécrétion de protéines et tout spécialement à Isabelle pour la relecture de l'article sur Eep. Chez UBLO toujours, je remercie Jean-Christophe Piard, Vincent Juillard et Virginie d'avoir partagé avec moi leurs connaissances sur les protéines ancrées et sur OptA, ainsi que plusieurs souches et des anticorps construits dans leur équipe. Merci également aux gens de Génétique Microbienne, Pierre Renault et Eric Guedon de m'avoir donné accès aux données transcriptomiques et pour toutes leurs explications à ce sujet. Merci également à Emanuelle Maguin pour avoir partagé avec moi des données double hybride et à Nicolas pour son aide dans la création d'un fichier maître de mon manuscrit de thèse. Toujours dans l'équipe de Génétique Microbienne je remercie Marion Velten et Sophie Cheruel pour leur aide dans la production hétérologue de PepF. Dans l'unité UEPSD je remercie Phillippe Langella pour ses conseils et pour la souche surproductrice de protéines sécrétées.

A mis amigos hispanoparlantes gracias por darme la oportunidad de hablar mi lengua materna de vez en cuando y por haber estado ahí para los divertidos almuerzos durante los cuales nos reímos tanto de las diferentes expresiones de cada uno y de las cuales se concluye que es en España en donde más raro se habla español. Of course, I cannot forget « the immigrants »: thank you for your freindship and your company during the diverse activities (museums, concerts, bars, picnics, restaurants, opera, etc.) that helped me to recharge energy after a long week spent in the lab.

Andrés, gracias por tu incondicional compañía, tu ayuda, tu paciencia. Como ya lo habías dicho, nuestros doctorados y todo lo que vivimos, viajamos y construimos alrededor son un « proceso realizado con éxito » .

Mami, danke für deine Unterstützung und ganz besonders dafür, dass du dreimal hier warst.

Table de matières

1 Introduction

De nombreuses études biologiques cherchent à déterminer le rôle de protéines dont l'existence potentielle a souvent été découverte lors du séquençage des génomes. Dans ce travail nous avons utilisé une démarche mêlant biologie et statistiques pour déterminer le rôle de certaines protéines de la famille des enzymes protéolytiques chez *Lactococcus lactis*. Notre démarche comporte l'inférence statistique des rôles puis la validation expérimentale des rôles prédits. Cette démarche est générale dans la mesure où elle est applicable à tout organisme entièrement séquencé pour lequel des données post-génomiques sont disponibles.

La protéolyse a été étudiée en détail chez *L. lactis* car elle joue un rôle important dans la nutrition azotée (Kunji *et al.*, 1998; Poolman *et al.*, 1995; Savijoki *et al.*, 2006). Il est bien connu que le lactocoque possède une machinerie protéolytique complexe spécialisée dans la dégradation des protéines du lait. Cependant, peu de choses sont connues sur les autres rôles de la protéolyse chez cette bactérie. Seules HtrA (Poquet *et al.*, 2000) et Clp (Varmanen *et al.*, 2000) ont été décrites comme participant dans la dégradation des protéines mal repliées ; DacA, DacB (Courtin *et al.*, 2006) et YjgB (Redko *et al.*, 2007) dans la maturation du peptidoglycane et ComC (Wydau *et al.*, 2006) dans la compétence.

Nous nous sommes intéressés à deux protéines en particulier, PepF et YvjB de *L. lactis*. PepF est une protéine étudiée auparavant chez le lactocoque, mais dont le rôle précis reste inconnu (Monnet *et al.*, 1994). YvjB n'a jamais été étudiée chez cette bactérie mais des protéines homologues ont été étudiées. C'est le cas de Eep, qui a été décrite chez *Enterococcus faecalis* comme participant à la maturation de phéromones (Antiporta & Dunny, 2002). Nous avons également exploré les prédictions pour la prenyl-peptidase CAAX. Néanmoins, faut de temps, nous n'avons pas réalisé de validations expérimentales pour confirmer les liens prédits pour cette peptidase.

Pour l'inférence statistique, nous avons utilisé des données existantes sur les protéines de *L. lactis* provenant de différentes sources. La séquence disponible de l'organisme nous permet d'utiliser plusieurs types de données dont des données transcriptomiques (microarray) ou des données protéomiques (gels 2D). Ces dernières fournissent une information très riche sur le niveau d'expression de gènes codant pour les protéines d'un organisme et sur les protéines elle mêmes, respectivement. Afin de pouvoir utiliser ces données hétérogènes, nous avons choisi une méthode à noyau qui permet la représentation et l'intégration de données hétérogènes. L'analyse statistique que nous avons réalisée permet d'inférer des liens des protéines de rôle inconnu avec d'autres protéines du lactocoque. Elle implique l'apprentissage de liens connus issus du graphe des voies métaboliques du lactocoque. Les liens prédits ont été validés grâce aux liens métaboliques connus. D'autres types de liens connus entre protéines du lactocoque comme les interactions physiques entre protéines et les régulons ont été comparés aux liens prédits avec notre méthode mais il n'y avait pas correspondance parfaite entre les deux.

Une fois obtenus, les liens prédits nous ont aidé à émettre des hypothèses biologiques sur le rôle des protéines. Nous avons trouvé une implication de PepF et YvjB dans la translocation de protéines et dans d'autres fonctions cellulaires variées. Les expériences de laboratoire, faites en comparant le phénotype de mutants de délétion de ces deux protéines à la souche sauvage, ont confirmé la plupart

des liens prédits. Ces validations ont montré que l'analyse statistique réalisée est pertinente pour la détermination du rôle de protéines.

Lors de l'analyse statistique, un grand nombre de données a été utilisé. Chacune de sources de données est constituée de plusieurs centaines de variables. Ces variables correspondent, par exemple, à des expériences différentes pour les données transcriptomiques ou des organismes pour les profils phylogénétiques. Nous avons donc voulu savoir quelles étaient les données essentielles en regard de la prédiction de protéines cibles. Pour répondre à cette question, nous avons réalisé une analyse de sensibilité sur deux des sources de données utilisées. Elle nous a permis de trouver un sous-ensemble de données stable fournissant les mêmes prédictions que la totalité de données.

2 Revue bibliographique

2.1 La protéolyse bactérienne

La protéolyse joue un rôle clé dans plusieurs processus biologiques chez les bactéries : nutrition azotée, virulence, turn-over protéique, activation de protéines, dégradation de protéines non conformes, sporulation, sécrétion …
La machinerie protéolytique est constituée d'un ensemble d'enzymes qui diffèrent par leur mécanisme catalytique, leur spécificité d'hydrolyse, leur localisation cellulaire, leur rôle.
Dans cette revue, j'aborderai, dans un premier temps, les différents modes de classification de ces enzymes protéolytiques et, dans un second temps, je balayerai les différents rôles identifiés, à ce jour (principalement chez le lactocoque) de ces enzymes.

2.1.1 Classification des protéases et peptidases

Les protéases (qui hydrolysent des protéines) et peptidases (qui hydrolysent des peptides) peuvent être classées selon trois critères différents : A) la nature de leur site catalytique, B) la spécificité du clivage et C) leur besoin en ATP (adénosine triphosphate).

A) Quatre mécanismes catalytiques principaux ont été décrits pour les enzymes protéolytiques, ce qui permet de les regrouper en grandes familles en fonction des acides aminés du site actif :

1) Les **protéases à sérine** possèdent trois acides aminés caractéristiques dans leur site actif : une sérine, une histidine et un aspartate. Les protéases à sérine ont en commun le mécanisme de coupure, basé sur la polarisation de la liaison peptidique par un groupement sérine. Pour atteindre cet état, l'histidine et l'aspartate, formant la triade avec la sérine, doivent être positionnés pour que le groupement OH de la sérine soit très fortement polarisé. Le mécanisme est résumé dans la Figure 2-1.

2) Les **protéases à thiol** possèdent une cystéine dans leur site actif. Il s'agit de protéases ou peptidases dans lesquelles le nucléophile est le soufre de la cystéine (Barrett & Rawlings, 2001). Le mécanisme d'action est montré dans la Figure 2-2.

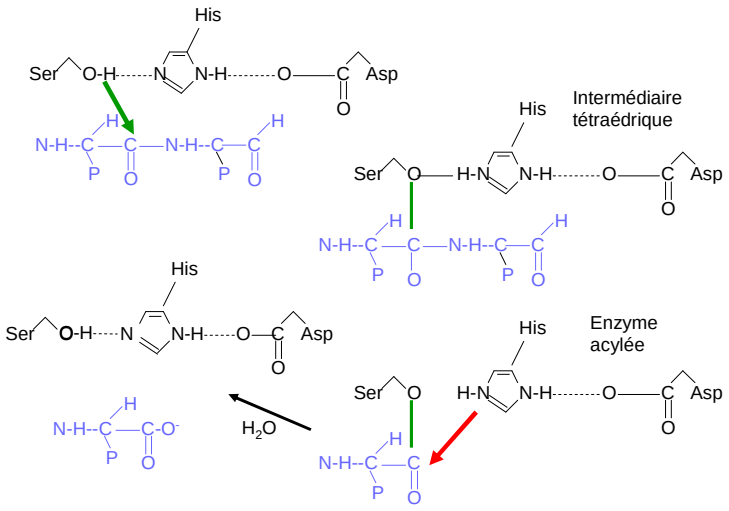


Figure 2-1 : Mécanisme des peptidases à sérine. Le polypeptide (bleu) s'insère dans la protéase (noir) à sérine de telle manière que le groupement carbonyle soit proche de la sérine. Le groupement OH de la sérine attaque le groupement carbonyle et l'azote de l'histidine accepte le OH de la sérine (vert). Une association enzyme-substrat intermédiaire tétraédrique se forme. Puis, l'enzyme est acylée et le premier produit d'hydrolyse est libéré (flèche rouge). Dans une réaction finale l'enzyme est déacylée et l'extrémité C-terminale est libérée (Kraut, 1977).

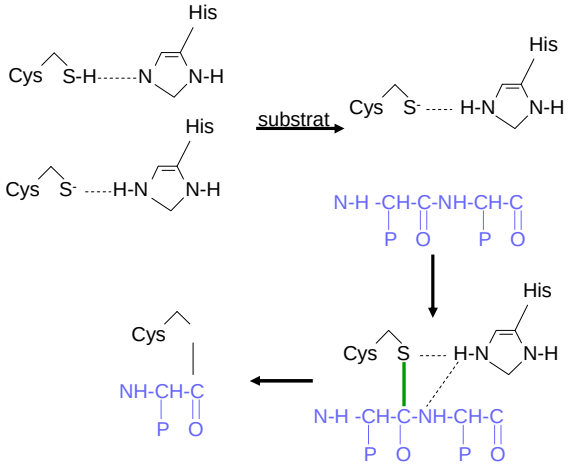


Figure 2-2: Mécanisme des peptidases à cystéine : Le premier pas du mécanisme catalytique est la déprotonisation d'un groupement thiol dans le site actif de l'enzyme (noir) réalisé par un acide aminé adjacent possédant une chaîne latérale basique (souvent c'est une histidine). Le deuxième pas consiste en une attaque nucléophile du soufre anionique de la cystéine sur le groupe carbonyle du substrat (vert). Dans ce pas, un fragment du substrat (bleu) est libéré. L'histidine dans la protéase est restaurée sous sa forme deprotonisé et un intermédiaire thiœsther lie l'extrémité carboxy-terminale du substrat à la cystéine. La liaison thioesther est hydrolysée pour générer un acide carboxy-terminal et l'enzyme est restaurée (Storer & Ménard, 1994).

3) Les **protéases acides** agissent à pH acide et possèdent un acide aspartique dans leur site actif. Plusieurs mécanismes d'action des protéases acides ont été proposés. Celui qui est généralement admis est un mécanisme général acide-base basé sur la coordination d'une molécule d'eau entre les deux aspartates du site actif. Un aspartate active la molécule d'eau par soustraction d'un proton. Cette activation de la molécule d'eau permet une modification du carbone carbonyle (carbone avec double liaison à une molécule d'oxygène) du substrat dans la liaison à cliver générant un oxyanion intermédiaire. C'est le réarrangement de cet oxyanion qui permet le clivage (Suguna *et al.*, 1987).

4) Les **métalloprotéases** possèdent un site de fixation pour un cation métallique. Les métalloprotéases sont le plus grand groupe de protéases et peptidases. Plus de 50 familles y ont été décrites (Figure 2-3). Dans ces enzymes, un cation, généralement le zinc, active une molécule d'eau. L'ion métallique est maintenu en place par trois acides aminés. Les acides aminés qui peuvent remplir cette fonction sont His, Glu, Asp ou Lys. Parmi les métalloprotéases connues, environ la moitié contiennent le motif HEXXH, qui se situe à l'endroit de la liaison du cation métallique (Rawlings & Barrett, 1995). Le groupe de métalloprotéases possédant ce motif est appelé la superfamille des « zincins ». Cette superfamille est divisée en au moins dix autres familles selon la localisation du troisième ligand du zinc. Les métalloprotéases bactériennes appartiennent à trois familles : thermolysine, serralysine et neurotoxine dont les enzymes types sont produites par *Bacillus thermoproteolyticus, Serratia marcescens* et *Clostridium botulinum* ou *C. tetani*, respectivement (Miyoshi & Shinoda, 2000).

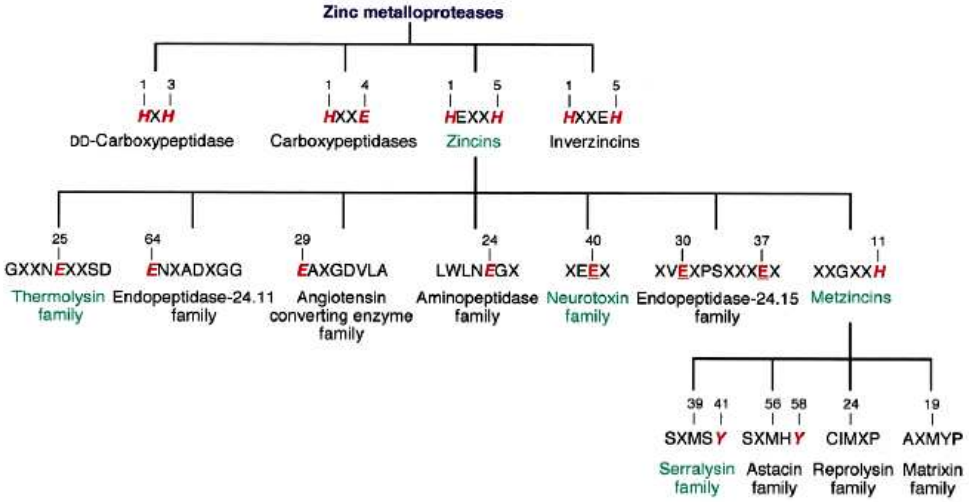


Figure 2-3 : Séquences des sites actifs des familles de métalloprotéases selon (Miyoshi & Shinoda, 2000).

B) Selon leur spécificité de clivage, les protéases peuvent être classées en deux grands groupes : les exopeptidases qui clivent aux extrémités des peptides et les endopeptidases qui le font à l'intérieur des chaînes peptidiques. Il existe donc des endopeptidases et des exopeptidases qui sont soit des metallo-, cystéine- ou sérine-peptidases, les métalloprotéases étant les plus répandues chez les bactéries (Gonzales & Robert-Baudony, 1996).

Les **aminopeptidases** libèrent des acides aminés de l'extrémité N-terminale d'un polypeptide. Ces peptidases ont soit une spécificité très restreinte pour un acide aminé donné ou très large et peuvent ainsi agir sur des polypeptides très différents. Le clivage de la méthionine N-terminale est, par exemple, réalisé par la méthionine aminopeptidase qui est très spécifique (Gonzales & Robert-Baudony, 1996). Les **carboxypeptidases** coupent entre l'avant dernier et le dernier acide aminé de l'extrémité C-terminale d'un polypeptide. Les **endopeptidases** sont capables de couper à l'intérieur de la chaîne protéique.

Parmi les aminopeptidases et les carboxypeptidases, on peut trouver aussi des dipeptidyl-peptidases. Il s'agit de peptidases qui enlèvent des dipeptides d'un polypeptide. La peptidase type est DAP I, aussi connue comme cathepsine C. Il s'agit d'une protéase à sérine de spécificité large (Metrione *et al.*, 1966).

Une nomenclature a été proposée par (Tan *et al.*, 1993) pour les peptidases des bactéries lactiques: elles sont nommés Pep lorsque le gène correspondant a été caractérisé, suivi d'une lettre capitale indiquant soit la spécificité de l'enzyme (« P » pour proline, par exemple) ou l'origine bactérienne (« S » pour *Streptococcus* sp.).

C) Un troisième classement, en plus de ceux liés à la nature du site catalytique et à la spécificité de clivage, est lié au besoin ou non d'ATP pour fonctionner (Hlavacek & Vachoa, 2002).
Ce groupe de protéases comprend des protéases composées de plusieurs sous-unités contenant des domaines d'ATPase et des domaines protéolytiques. Les protéases les plus connues de ce type sont les protéases Lon, Clp, HsIUV et FtsH qui participent à la dégradation de protéines mal repliées (voir section 2.1.2.2) et au *turnover* ou recyclage protéique (section 2.1.2.4).

Il existe des protéases qui ne peuvent être classées dans aucun de ces groupes. Ainsi, par exemple les signal peptidases des lipoprotéines forment une nouvelle famille (Tjalsma *et al.*, 1999b). Un autre exemple est le groupe des endopeptidases phagiques (Loessner, 2006).

2.1.2 Fonctions de la protéolyse chez les bactéries

2.1.2.1 La nutrition azotée

La protéolyse chez les bactéries lactiques est bien connue (Poolman *et al.*, 1995) ; (Kunji *et al.*, 1998; Savijoki *et al.*, 2006). La machinerie protéolytique développée chez ces bactéries est assez complexe du fait de la présence limitée d'acides aminés libres dans le milieu lait et aux auxotrophies de cette espèce pour plusieurs acides aminés. Le système protéolytique joue un rôle clé dans la fermentation du lait et permet l'obtention d'acides aminés à partir des caséines, les protéines les plus abondantes dans le lait.

La protéolyse chez les bactéries lactiques commence par l'action d'une protéase de paroi, enzyme à sérine qui hydrolyse les caséines du lait en oligopeptides. Certaines souches de bactéries lactiques ne possèdent pas de protéase de paroi et dépendent de la protéase de paroi présente chez les autres souches pour se développer dans le lait (Savijoki *et al.*, 2006). Cinq types de protéases de paroi de la même famille mais présentant certaines différences ont été caractérisées chez les bactéries lactiques : PrtP chez *Lactococcus lactis* et *Lactobacillus paracasei*, PrtH chez *L. helveticus*, PrtR chez *Lactobacillus rhalnosus*, PrtS chez *S. thermophilus* et PrtB chez *L. bulgaricus* (Savijoki *et al.*, 2006). Elles contiennent différents domaines. Le domaine correspondant au peptide signal (PP), le domaine catalytique des protéases à sérine (PR), un domaine d'insert (I) qui régule probablement leur spécificité, le domaine A de fonction inconnue, le domaine B participant probablement à la stabilité, le domaine hélix (H) qui positionne A et B à l'extérieur de la cellule et un domaine hydrophobe W (Siezen, 1999), voir Figure 2-4.

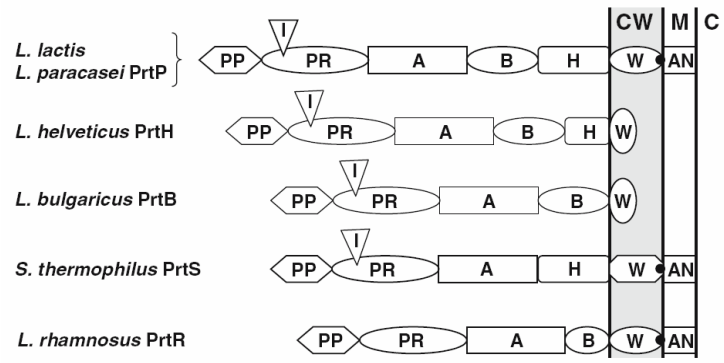


Figure 2-4: Représentation schématique des protéases de parois, selon Siezen *et al* 1999.

Les oligopeptides produits par l'action de la protéase constituent la source principale d'acides aminés pour le lactocoque. Les oligopeptides sont ensuite transportés à l'intérieur de la cellule (Tynkkynen *et al.*, 1993) par trois ou quatre transporteurs selon la souche. Ils appartiennent à deux grand groupes : les PRT (peptide transport) et les ABC (ATP-binding-cassette) transporteurs. Pour les premiers, le transport de peptides dépend de la force proton-motrice, pour les deuxièmes, il dépend de l'ATP. Les systèmes Opp appartiennent à la superfamille des ABC transporteurs. DtpT et DtpP sont des transporteurs proton dépendants. Ces transporteurs se distinguent par leur organisation et leur spécificité. Le système Opp transporte des peptides entre 4 et 18 acides aminés (Lamarque *et al.*, 2004). DtpT et DtpP transportent les di- et tri-peptides (Guedon *et al.*, 2001a; Kunji *et al.*, 1998).

Les peptides internalisés sont ensuite hydrolysés par plusieurs peptidases différentes avec des spécificités chevauchantes dans certains cas : Figure 2-5 (Guedon *et al.*, 2001a; Kunji *et al.*, 1998; Lamarque *et al.*, 2004; Savijoki *et al.*, 2006). Les aminopeptidases générales comme PepC et PepN libèrent des acides aminés de l'extrémité N-terminale d'une grande quantité d'oligopeptides. Les aminopeptidases PepV et PepT ont aussi des spécificités de séquence très larges et hydrolysent des di- et tri-peptides, respectivement. L'endopeptidase PepO, ou encore PepF, hydrolysent des oligopeptides d'une certaine taille. Certaines peptidases hydrolysent spécifiquement certaines séquences, ainsi par exemple PepX et PepQ sont spécialisées dans l'hydrolyse des peptides contenant la proline, très abondante dans les caséines (voir 2.2.1). La construction de mutants simples et multiples des genes codant ces peptidases et le suivi de leur croissance dans le lait a permis de mettre en évidence celles qui étaient importantes pour la nutrition azotée, comme par exemple PepN, PepC, PepO, PepT et PrtP (Mierau *et al.*, 1996).

Les bactéries lactiques répondent à la quantité d'azote disponible en régulant leur système protéolytique. (Guedon *et al.*, 2001a) ont montré que l'expression de six gènes ou groupes de gènes, *prtP, prtM, opp-pepO, pepD, pepN, pepC* et *pepX* est réprimée de 5 à 150 fois suite à l'ajout d'un hydrolysat contenant 80% de peptides et 20% d'acides aminés et qu'il y a expression de ces gènes uniquement en conditions d'azote limitantes. Il est maintenant connu que CodY régule négativement l'expression de certains gènes codant des protéines participant à la protéolyse et que cette répression est modulée par la quantité intracellulaire d'acides aminés branchés (isoleucine, leucine, valine) (Guedon *et al.*, 2001b), voir Figure 2-5. Des expériences *in vitro* ont montré que CodY reconnaît une séquence intérgénique précédant l'opéron *opp* régulé et que les acides aminés branchés stimulent cet accrochage (den Hengst *et al.*, 2005a). Ces résultats confirment les observations réalisées par (Mierau *et al.*, 1996) lors de l'études des mutants de protéases et

peptidases.

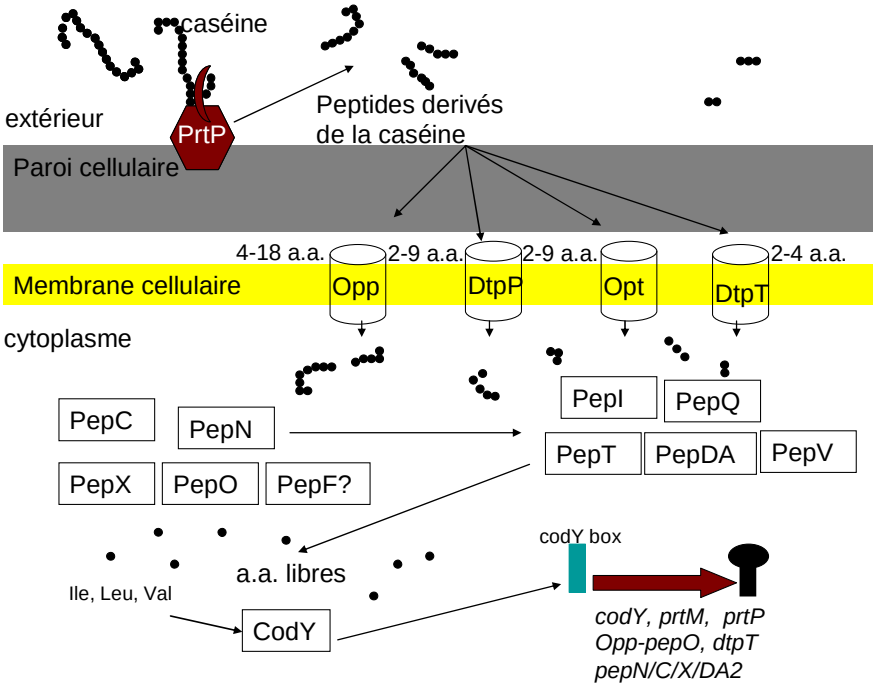


Figure 2-5 : Le système protéolytique chez *L. lactis* selon Guedon *et al* 2001.

D'autres protéines faisant partie de la machinerie protéolytique semblent ne pas être régulées par CodY, comme c'est le cas de PepF, ou régulées partiellement par CodY comme PepO1 et PepC, qui sont induites dans des conditions d'aérobiose même dans un mutant *codY*. Par ailleurs, la source de carbone affecte l'expression de *pepP* chez *L. lactis* (Guedon *et al.*, 2001a).

Des études de différents mutants des gènes faisant partie de la machinerie protéolytique ont permis de comprendre le rôle clé de la protéase de paroi et du transporteur d'oligopeptides dans la libération d'acides aminés à partir de la caséine. En effet, PrtP et Opp se sont avérés cruciaux pour la croissance dans un milieu contenant des caséines comme seule source d'azote. Au contraire, le manque de DtpT n'a aucune influence sur la croissance dans un tel milieu (Kunji *et al.*, 1998).

Le système protéolytique joue également un rôle très important dans la nutrition azotée chez d'autres bactéries, notamment des bactéries capables de se développer dans le lait. Ainsi, par exemple, *Staphylococcus aureus* présente une croissance diauxique dans le lait avec deux phases exponentielles de croissance. Pendant la phase de transition se produit la coagulation du lait. La coagulation semble être provoquée par l'activité protéolytique des enzymes sécrétées par *S. aureus*, qui produit 10 protéases extracellulaires : les protéases SspA, SplA, B, C, D, E, et F, ScpA et SspB, et la métalloprotéase Aur (Hiron *et al.*, 2007). Les auteurs ont également démontré qu'un des quatre transporteurs d'oligopeptides (Opp3) est indispensable pour le développement dans le lait.

2.1.2.2 La dégradation des protéines mal repliées

Une des réponses cellulaires au stress est l'induction rapide et temporaire de la machinerie protéolytique pour éliminer les protéines mal repliées suite à un changement de conditions

environnementales (Savijoki *et al.*, 2006). Les bactéries lactiques, par exemple, sont soumises à des stress durant les processus industriels. Chez *E. coli*, la protéase Lon est la première responsable de l'élimination des protéines non natives. Des homologues de Lon ne sont pas connus chez les bactéries lactiques. Néanmoins, trois protéases ont été décrites en relation avec le stress : Clp, HtrA et FtsH (Savijoki *et al.*, 2006). HtrA accomplit apparemment le rôle de la protéase Lon chez *L. lactis*, en dégradant les protéines anormales sécrétées (Poquet *et al.*, 2000). La sous-unité protéolytique de la protéase Clp, ClpP et les unités de régulation ClpC, ClpE et ClpB sont régulées par le régulateur de la réponse au stress CtsR (Varmanen *et al.*, 2000). Par ailleurs, l'inactivation du gène *trmA* stimule la dégradation de protéines mal repliées indépendamment de Clp. Ces résultats indiquent que TrmA est un régulateur négatif de la protéolyse chez *L. lactis* (Frees *et al.*, 2001).

2.1.2.3 La virulence

Plusieurs espèces bactériennes sécrètent des protéases avec des fonctions différentes. Certaines de ces protéases interagissent avec les systèmes de défense de l'hôte, d'autres attaquent et endommagent des tissus de l'hôte. Ces protéases constituent des facteurs de virulence importants (Miyoshi & Shinoda, 2000).

La plupart des protéases toxiques sont des métalloprotéases. Le motif HEXXH qui les caractérise a été retrouvé dans les neurotoxines de *Clostridium* sp., enterotoxines de *Bacteroides fragilis* et le facteur létal de *Bacillus anthracis*. Ces exotoxines montrent une activité protéolytique très spécifique contre les protéines cibles. Ainsi, par exemple, les neurotoxines de *Clostridium sp.* inhibent la libération de l'acétylcholine au niveau des jonctions musculaires et synapses. Il est connu maintenant que les neurotoxines ont une activité protéolytique dirigée vers la protéine membranaire du vésicule synaptique (Synaptobrevin-2), la protéine membranaire pré-synaptique (SNAP-25) et la syntaxine (Miyoshi & Shinoda, 2000).

Streptococcus pyogenes, se sert de ses protéases au cours de l'infection pour détruire des tissus et produit en même temps des inhibiteurs de protéases de l'hôte (Rasmussen & Bjorck, 2002). Les auteurs proposent que lors de la première partie de l'infection, *S. pyogenes* inhibe la protéolyse de l'hôte en exportant des inhibiteurs de protéase à sa surface. Dans la deuxième étape de l'infection, *S. pyogenes* produit massivement des enzymes protéolytiques et leur libération permet la dégradation des tissus humains et ainsi l'invasion par la bactérie. Ces événements sont régulés temporellement et spatialement et influencent fortement la virulence de *S. pyogenes* (Rasmussen & Bjorck, 2002).

La protéase la mieux connue, participant au processus d'infection chez *S. pyogenes* est la protéase C5a (Rasmussen & Bjorck, 2002). La peptidase C5a est une endopeptidase qui dégrade le facteur chemotactique C5a. Cette dégradation affecte la chemotaxie des neutrophiles (chargés de la première défense contre les bactéries) et permet à *S. pyogenes* de coloniser plus facilement les tissus. SpeB est une autre protéase sécrétée qui participe à la virulence de différentes manières. C'est une protéine capable de dégrader des protéoglycanes, libérant du dermatane sulphate qui inactive des peptides antibactériens. De plus, elle est capable de cliver des immunoglobulines humaines, ainsi que de libérer l'agent inflammatoire bradykinine à partir du clivage de kininogènes contribuant à l'infection par *S. pyogenes* (Rasmussen & Bjorck, 2002).

2.1.2.4 Le recyclage ou *turnover* de protéines

La synthèse des nouvelles protéines nécessite des acides aminés libres. Souvent ces acides aminés proviennent du recyclage (turnover) d'autres protéines. Chez *E. coli*, une des protéines chargées de la destruction de protéines en fin de vie est la protéase Lon mentionnée auparavant (Gottesman, 1996). Néanmoins, le plus probable est que plusieurs enzymes protéolytiques participent au turnover des protéines.

La dégradation des protéines doit être bien régulée et ciblée pour éviter l'hydrolyse des protéines utiles. Chez les eucaryotes, la dégradation de protéines est réalisée principalement par le protéasome 26S, qui est formé par l'associoation du régulateur 19S au protéasome 20S. La conformation cylindrique du complexe 20S fait que les sites actifs se trouvent à l'intérieur du cylindre, accessibles par des pores étroits. Ce compartimentage évite la dégradation non régulée de protéines. Le complexe 19S fait entrer les substrats à dégrader en utilisant de l'ATP (Zwickl *et al.*, 2001).

Chez les procaryotes, les « protéasomes » contiennent aussi un module ATP et un module protéolytique attaché de manière covalente (Lon) ou libre (ClpP). Ces protéasomes sont chargés de la dégradation de protéines mal repliées ou identifiés pour être dégradées. Ainsi, la peptidase ClpP de *Escherichia coli*, présente une architecture similaire au protéasome 20S. L'accès aux sites actifs est coordonné par plusieurs protéines de type Hsp100 (par exemple ClpA) qui fonctionnent avec de l'ATP (Mogk *et al.*, 2007).

Les protéines à dégrader doivent posséder un signal pour être reconnu par le système protéolytique. Chez les eucaryotes c'est l'ubiquitine qui est attaché aux protéines à dégrader. Chez les procaryotes, ce type de système existe aussi, mais est complètement indépendant de l'ubiquitine. La spécificité des substrats passe par les N-domaines des protéines Hsp100. Les N-domaines interagissent directement avec les substrats. Les sites de reconnaissance reconnus par les N-domaines se trouvent dans la partie N-terminale des protéines. Dans cette région, certains acides aminés favorisent ou défavorisent, selon leur nature, la liaison avec les substrats (Mogk *et al.*, 2007).

2.1.2.5 Protéolyse intermembranaire et régulatrice

Une protéolyse spécifique permet de libérer des peptides qui ont un rôle régulateur. Il s'agit de la protéolyse intermembranaire régulatrice ou RIP (Brown *et al.*, 2000). La RIP est un mécanisme dans lequel une peptidase membranaire réalise un rôle régulateur par la dégradation de lipoprotéines ou de protéines membranaires. La protéolyse intermembranaire agit sur un morceau de la protéine, de moins de 30 acides aminés. Le système le plus étudié est le SREBP (sterol regulatory element-binding proteins) humain. La protéine nécessaire dans ce système et chargée de la RIP a été identifiée comme une métalloprotéase membranaire S2P. Le peptide produit par S2P régule la transcription des gènes nécessaires pour la synthèse de lipides.

Des homologues de S2P existent chez les archaeas, les bactéries et les eucaryotes. La plupart des génomes d'archae contiennent plusieurs séquences S2P. Chez les bactéries, le nombre de copies est variable allant de un chez la plupart de bactéries à sept trouvés chez *B. subtilis*. Les bactéries qui

possèdent un petit génome n'ont souvent pas de S2P, comme par exemple *Bifidobacterium longum, Buchnera aphidicola, Mycoplasma pneumoniae, Ureaplasma urealyticum* (Kinch *et al.*, 2006).

Des exemples d'enzymes appartenant à la famille S2P seront présentées en section 2.2.4.2 au moment où YvjB, une des protéines cibles de ce travail, faisant partie de cette famille, sera présentée.

2.1.2.6 Peptidases participant à la maturation et l'export des protéines

Une fois les protéines synthétisées, une première peptidase, PepM, est chargée de libérer le premier acide aminé en position N-terminale, la méthionine. PepM a été étudiée principalement chez *E. coli* (Ben-Bassat *et al.*, 1987).

Les protéines exportées sont synthétisées comme précurseurs avec un peptide signal. Ce peptide signal est nécessaire pour leur translocation à travers la membrane. Durant le processus de translocation, ce peptide signal est enlevé par les signal peptidases. Il existent deux signal peptidases : une pour les lipoprotéines, la SPase II et une pour les non lipoprotéines, la SPase I. Les deux sont des protéines membranaires. La SPase I appartient à un groupe de sérine protéases qui utilise la dyade catalytique Ser-Lys ou Ser-His au lieu de la triade plus commune Ser-His-Asp (Paetzel *et al.*, 2002). Les SPases II appartiennent à une nouvelle famille de protéases (Tjalsma *et al.*, 1999b). Les mécanismes auxquels participent ces peptidases sont présentés en détails dans la section 2.4.1 qui concerne l'export de protéines.

2.2 Les peptidases et protéases du lactocoque

J'ai réalisé un bilan des peptidases et protéases décrites dans la littérature à ce jour, annotées lors du séquençage du génome de *L. lactis* (Bolotin *et al.*, 2001) et en interrogant les bases des données comme NCBI (http://www.ncbi.nlm.nih.gov/) pour la recherche de « cluster of orthologous groups » (COG), en tenant ainsi compte de domaines protéolytiques conservés. Il est possible d'établir une liste de 41 peptidases connues ou potentielles chez *L. lactis* (voir Tableau 2-1). Certains motifs, des motifs clés d'enzymes protéolytiques, peuvent ne pas avoir été caractérisés à l'heure actuelle, ce qui rend cette liste potentiellement non exhaustive.

Plusieurs protéases et peptidases de lactocoque sont bien caractérisées et leurs participations dans une ou différentes fonctions cellulaires bien établies. Les fonctions de certaines d'entre elles ont été déterminées grâce à leur similarité avec des peptidases et protéases connues chez d'autres organismes. D'autres sont complètement inconnues. Par la suite, je m'intéresse particulièrement aux peptidases du lactocoque dans le but de faire un bilan sur les connaissances des ces enzymes protéolytiques. Cela me permettra d'introduire les peptidases cibles de ce travail, notamment PepF, YvjB et la prényl-peptidase CAAX.

Tableau 2-1 : Liste de protéases et peptidases de *L. lactis* IL1403.

	Gène	**Protéine**
1	*clpP*	Sous-unité protéolytique de la protéase Clp
2	*comC*	Prépiline peptidase / N-méthyltransferase
3	*dacA*	D-alanyl-D-alanine carboxypeptidase
4	*dacB*	D-alanyl-D-alanine carboxypeptidase
5	*gcp*	O-sialoglycoprotéine endopeptidase
6	*htrA*	Protéase de ménage
7	*lspA*	Lipoprotéine signal peptidase
8	*pepA*	Glutamyl aminopeptidase
9	*pepC*	Aminopeptidase C
10	*pepDA*	Dipeptidase
11	*pepDB*	Dipeptidase
12	*pepF*	Oligoendopeptidase F
13	*pepM*	Méthionine aminopeptidase
14	*pepN*	Aminopeptidase N
15	*pepO*	Métalloendopeptidase
16	*pepP*	Xaa-Pro aminopeptidase
17	*pepQ*	Proline dipeptidase
18	*pepT*	Di- et tripeptidase
19	*pepV*	Dipeptidase
20	*pepX*	X-prolyl dipeptidyl aminopeptidase
21	*pi136*	Prohead protéase
22	*pi323*	Clp protéase putative
23	*sipL*	Signal peptidase I
24	*yaiF*	Protéine hypothétique avec un domaine prényl-peptidase CAAX
25	*yaiH ou caax*	Protéine hypothétique avec un domaine prényl-peptidase CAAX
26	*yajF*	Protéine hypothétique avec un domaine prényl-peptidase CAAX
27	*ybdI*	Protéine hypothétique avec un domaine prényl-peptidase CAAX
28	*ybdJ*	Protéine hypothétique avec un domaine prényl-peptidase CAAX
29	*ybcH*	Peptidase X-Pro dipeptidyl putative
30	*yciA*	Putative N-acetyldiaminopimelate déacetylase
31	*yddA*	Prépiline signal peptidase PulO and related peptidases
32	*yjiB*	Metal-dependent amidase/aminoacylase/carboxypeptidase
33	*yjgB*	Gamma-glutamyl-diamino acid-endopeptidase
34	*yueE*	Predicted Zn-dependent peptidases
35	*yueF*	Protease
36	*yugD*	Protease
37	*yuhB*	Protease
38	*yvdC*	Protéine hypothétique avec un domaine prényl-peptidase CAAX
39	*yvdE*	Glutamine amidotransferase putative avec domaine peptidase
40	*yvjB ou eep*	Métalloprotéase à zinc hypothétique
41	*yxdF*	Protéine hypothétique avec un domaine prényl-peptidase CAAX

2.2.1 Les peptidases et protéases participant principalement à la nutrition azotée

Comme détaillé dans la section 2.1.2.1 , la machinerie protéolytique de *L. lactis* dédiée à la nutrition azotée est bien connue. De ce fait, les peptidases participant à ce processus ont été bien caractérisées. Il s'agit de PepC, PepN, PepO, PepQ, PepT, PepV et PepX (Kunji *et al.*, 1998) ; (Guedon *et al.*, 2001b) ;(den Hengst *et al.*, 2005b) ; (Mierau *et al.*, 1996). La plupart ont des spécificités de substrat générales et mêmes chevauchantes comme PepC, PepN et PepT. Quatre de ces peptidases sont spécialisées dans le clivage des peptides contenant des prolines. Etant dérivés des caséines, les peptides constituant les substrats naturels de *Lactococcus lactis* sont riches en proline (Kunji *et al.*, 1998). Dans la suite, des détails sont donnés sur la spécificité des peptidases impliquées dans la nutrition azotée.

PepC a été isolée à l'origine de la souche *L. lactis* AM2 et décrite comme étant une aminopeptidase générale (Neviani *et al.*, 1989). Elle a été classée parmi les cystéine aminopeptidases par (Chapot-Chartier *et al.*, 1993). La spécificité pour l'acide aminé N-terminal est assez générale, bien que des peptides contenant de la proline en cette position ne soient pas hydrolysés. Il s'agit d'une enzyme cytoplasmique (Tan *et al.*, 1992).
PepN ou la lysyl aminopeptidase est une peptidase de spécificité large capable de libérer l'acide aminé N-terminal des peptides. L'enzyme a été appelée PepN chez le lactocoque car elle complémente une mutation *pepN* chez *E. coli.* PepN de *L. lactis* hydrolyse des di-, tri- et oligopeptides, mais la longueur optimale du peptide substrat est un tétramère (Niven *et al.*, 1995). L'enzyme a une préférence pour les peptides contenant arginine comme résidu N-terminal, clivant aussi des peptides avec Leu ou Lys à cette position.
PepO a été appelé ainsi pour la première fois par Mierau et collaborateurs (Mierau *et al.*, 1993). Il s'agit d'une peptidase capable de cliver des peptides de séquences très différentes composées de 5 à 30 acides aminés (Tan *et al.*, 1991). Le gène *pepO* fait partie de l'opéron codant pour le transporteur d'oligopeptides Opp. L'inactivation du gène (simple mutant *pepO*) n'affecte pas la croissance en lait (Mierau *et al.*, 1993).
PepP ou l'aminopeptidase P a été caractérisée chez *L. lactis* (Mars & Monnet, 1995). La taille optimale de son substrat est de 5 acides aminés et la peptidase présente la spécificité X-Pro-Pro pour les peptides contenant de la proline.
PepT a été caractérisée comme une tripeptidase capable de cliver des acides aminés de plusieurs peptides libérant un acide aminé libre et un dipeptide (Mierau *et al.*, 1994).
PepV a été trouvé pour la première fois chez *Lactobacillus delbruecki* (Vongerichten *et al.*, 1994). L'enzyme clive plusieurs peptides (aussi des β-ala-dipeptides) et certains tripeptides.
PepX est une X-prolyl dipeptidyl aminopeptidase libérant des dipeptides X-Pro à partir des extrémités N-terminales des peptides. Elle a été caractérisée chez *L. lactis* (Nardi *et al.*, 1991). Elle est capable de dégrader des peptides provenant des caséines et particulièrement de la caséine β riche en proline. Elle n'est pas essentielle pour la croissance mais la composition de peptides du lait fermenté change en son absence (Mierau *et al.*, 1996).
Des mutations dans chacun de ces gènes individuellement affecte peu la croissance de la bactérie dans le lait mais l'accumulation de ces mutations conduit à une réduction très importante de cette croissance (Mierau et al., 1994). Les peptidases participant à la nutrition azotée de cette manière ont été identifiés comme faisant partie du régulon CodY qui régule leur transcription répondant au *pool* d'acides aminées branchés (Guedon *et al.*, 2001).

2.2.2 Les peptidases dédiées aux fonctions autres que la nutrition azotée

Compétence naturelle : ComC, une prépiline-like peptidase

Wydau et collaborateurs ont étudié les éléments participant à la compétence chez *L. lactis*, identifiés sur la base d'homologies de séquences (Wydau *et al.*, 2006). Ces éléments sont présents bien que la compétence naturelle de *L. lactis* n'ait pas été montrée. La peptidase ComAB fait partie de ces éléments et participe aux phases tardives de la compétence chez *Streptococcus pneumoniae* (Pestova *et al.*, 1996). Les résultats obtenus par Wydau et collaborateurs montrent que les protéines codées par les gènes de compétence homologues chez *S. pneumoniae* doivent avoir la même fonction que celle décrite pour *S. pneumoniae* (Pestova *et al.*, 1996). Wydau et collaborateurs ont observé que l'expression des gènes qui interviennent dans la phase tardive de la compétence est augmentée lors d'une surexpression de ComX, qui est le régulateur de la compétence chez *S. pneumoniae.* En réponse à un signal externe, des facteurs sigma provoquent le passage à la phase tardive de la compétence en stimulant l'expression des gènes associés à cette phase, dont *comC* fait partie (Pestova *et al.*, 1996).

Par ailleurs, ComC possède des motifs communs avec les protéases qui clivent les prépilines (précurseur des protéines fibreuses trouvées dans les « pili » bactériens) de type IV, il s'agirait alors d'un type de signal peptidase qui cliverait le peptide signal de la prépiline. (Chung & Dubnau, 1995).

Synthèse du peptidoglycane : DacA* et *DacB, des D-alanyl-D-alanine-carboxypeptidases/ transpeptidase

Les peptidases DacA et DacB ont des homologues chez d'autres organismes (DacA par exemple présente 42% de similarité avec DacA de *E. coli*). Il s'agit de peptidases nécessaires à l'étape finale de synthèse du peptidoglycane, c'est à dire à la liaison de chaînes de N-acétyleglucosamine et d'acide N-acétylmuramique par des peptides courts. La réaction consiste à libérer de la D-alanine à partir du peptide donneur L-Ala-γ-D-Glu-L-R3-D-Ala et à son transfert au groupe L-R3 d'un autre peptide similaire récepteur. Ces enzymes sont en fait des transpeptidases (Rhazi *et al.*, 2003).

Plus en détail, la réaction consiste en l'activation de la sérine du site actif de la peptidase qui attaque le groupement carbonyle de l'avant-dernier résidu D-Ala du précurseur du peptidoglycane. Cela libère le dernier D-Ala, donnant un intermédiaire acyl-enzyme. Une amine provenant d'une autre chaîne de peptidoglycane réagit alors avec l'ester de l'intermédiaire acyl-enzyme pour donner deux chaînes de peptidoglycane unies par le peptide (Lee *et al.*, 2001).

Ces enzymes sont inhibées par les antibiotiques de type beta-lactames qui acylent la sérine dans le site actif de l'enzyme. Chez *L. lactis*, DacB a été caractérisée comme étant impliquée dans la maturation du peptidoglycane, confirmant les observations faites chez d'autres organismes (Courtin *et al.*, 2006).

Hydrolyse du peptidoglycane : Yjgb, une gamma-D-glutaminyl-L-lysyl-endopeptidase
Il s'agit d'une des 5 hydrolases du peptidoglycane chez *L. lactis.* Redko et collaborateurs ont montré que YjgB est une endopeptidase qui hydrolyse les peptides du peptidoglycane contenant la séquence gamma-D-Gln-L-Lys (Redko *et al.*, 2007).

Maturation des polypeptides nouvellement synthétisés : PepM, la méthionine aminopeptidase
Il s'agit de la peptidase chargée d'enlever la méthionine en position N-terminale des protéines nouvellement synthétisées. Elle a été étudiée principalement chez *E. coli* (Ben-Bassat *et al.*, 1987), mais annotée dans la plupart de génomes connus comme méthionine aminopeptidase grâce à son homologie avec la protéine de *E.coli.*

PepA, une glutamyl aminopeptidase
Parmi les bactéries, PepA est bien caractérisée chez *E. coli* et *S. thyphimurium.* Il s'agit d'une enzyme capable de libérer préférentiellement un acide glutamique (Vogt, 1970). PepA a été étudiée chez *L. lactis* MG1363 (l'Anson *et al.*, 1995). Les auteurs ont conclu que PepA n'est pas une peptidase essentielle mais qu'elle est nécessaire pour une croissance optimale dans le lait, néanmoins elle ne fait pas partie du régulon CodY (Guedon *et al.*, 2001b).

Clivage des peptides signaux : Signal peptidases I et II
Strictement parlant il s'agit de protéases. Elles sont chargées d'enlever les peptides signaux de protéines pendant le processus de translocation (Novak *et al.*, 1986). Les peptidases signal de type II sont spécifiquement chargées de cliver les peptides signaux des lipoprotéines (Tjalsma *et al.*, 2000).

2.2.3 Les peptidases et protéases de *Lactococcus lactis* de rôle inconnu ou mal connu

Nous pouvons diviser les peptidases les moins connues chez le lactocoque en trois groupes :

1. Etudiées, mais dont le rôle n'a pas été complètement élucidé
C'est le cas de l'oligoendopeptidase PepF. C'est une endopeptidase qui a été décrite chez *L. lactis* (Monnet *et al.*, 1994) dont la spécificité sur substrats synthétiques et l'activité ont été bien caractérisées, mais dont le rôle précis est resté inconnu après ces premières études. Plusieurs données, assez diverses, sont disponibles sur cette peptidase chez diverses bactéries. Comme il s'agit d'une de nos peptidases d'intérêt, les informations disponibles dans la littérature seront présentées plus en détail dans la section 2.2.4.1

2. Annotées grâce à la présence d'un motif ou une similarité avec des peptidases connues chez d'autres organismes
Dans ce groupe, nous retrouvons YciA, YddA, YjiB, YvdE, YueE (pour les motifs voir Tableau 2-1) et YvjB. La dernière peptidase est un homologue de Eep de *Enterococcus faecalis,* qui sera traitée en détail dans la section 2.2.4.2 .

Gcp, une O-sialoglycoprotéine endopeptidase
La fonction de cette peptidase est également proposée à la vue de son homologie avec les O-siaglycoprotéine endopeptidases chez d'autres organismes. Gcp a été étudiée principalement chez *Pasteurella haemolytica* (Cladman *et al.*, 1996). Il s'agit d'une peptidase qui clive principalement des peptides contenant de l'acide sialique, plus exactement ceux dont les résidus sérine ou thréonine contiennent cet acide (Cladman *et al.*, 1996).

PepDA et PepDB, des dipeptidases
PepDA a été la première dipeptidase identifiée chez *Lactobacillus helveticus* (Dudley *et al.*, 1996). Sa fonction dipeptidase chez *L. lactis* a été déduite par homologie (Bolotin *et al.*, 2001).

3. Les peptidases dont l'annotation automatique ne permet pas d'assigner une fonction :
Dans cette situation nous trouvons plusieurs copies de la prényl-endopeptidase CAAX : YaiH, YaiF, YajF, YbdI, YbdJ, YvdC, YxdF. Tous ces gènes ne sont pas annotés. Néanmoins, le premier est à 83% similaire à la prényl-peptidase de *L. lactis* subsp. *cremoris* utilisée dans la description de cette famille (Pei & Grishin, 2001). Les connaissances sur ce type de peptidases seront présentées dans la section 2.2.4.3 .

2.2.4 Les peptidases d'intérêt

Au cours de ce travail, nous nous sommes intéressés à trois peptidases qui, selon les connaissances disponibles, pourraient être impliquées dans des processus cellulaires globaux et régulateurs. De plus, nous avons ciblé les peptidases présentes chez différents organismes et pour lesquelles le rôle, déterminé chez un organisme modèle comme *L. lactis*, pourrait être généralisé. Dans la suite, je présente les connaissances sur les peptidases d'intérêt PepF, YvjB et la prényl-peptidase CAAX. Pour cette dernière, compte tenu du peu de connaissances disponibles dans la littérature, je presenterai quelques résultats préliminaires issus de la recherche dans les bases des données.

2.2.4.1 PepF ou l'oligoendopeptidase F

PepF est une protéine présente chez de nombreuses espèces de bactéries à faible contenu en GC: (*Staphylococcus* sp., *Bacillus* sp., *Streptococcus* sp., *Lactobacillus* sp., *Mycoplasma* sp., *Clostridium* sp., etc), chez des Spyrochetes, des Proteobacteria (*Agrobacterium* sp., *Escherichia coli, Salmonella* sp., *Yersinia* sp., etc.), des Archaea (*Halobaterium* sp., *Methanosarcina* sp., etc), Protozoa (*Plasmodium* sp.), et d'autres (*Thermus* sp., *Deinococcus* sp., *Rhodopirellula* sp., etc). Les études réalisées sur PepF chez différents organismes ont montré qu'elle participe à divers processus cellulaires et que son inactivation a des effets pleïotropiques. Nous nous sommes intéressés en priorité à cette peptidase parce que c'est la mieux caractérisée des trois peptidases que nous avons sélectionnées. De plus, son probable rôle pleïotrope et sa présence chez différentes bactéries permettent d'augmenter les connaissances sur les différentes fonctions cellulaires et de les étendre à d'autres bactéries.

La première copie de *pepF*, caractérisée chez *L. lactis*, se trouve sur un plasmide de 55 kb de la

souche *L. lactis* NCDO 763. Ce plasmide contient plusieurs gènes codant pour des protéines qui permettent l'utilisation du lactose, et la protéase de paroi. Il s'agit de gènes essentiels pour la croissance dans le lait (Monnet *et al.*, 1994). Le même cas de figure existe chez la souche *L. lactis* SK11, où le gène codant PepF est situé sur un plasmide de 47 kb qui contient également les gènes pour l'utilisation du lactose, ainsi que d'autres codant des protéines participant dans la protéolyse et le transport de peptides (Siezen *et al.*, 2005). Chez ces deux souches, le gène *pepF* est dupliqué puisqu'il s'en trouve une copie sur le chromosome en plus de celle du plasmide.

PepF, avec une masse de 70 kDa et 602 acides aminés, a été caractérisé comme une enzyme capable d'hydrolyser des oligopeptides entre 7 et 17 acides aminés, ayant une activité cytoplasmique (Monnet *et al.*, 1994). Les conditions pour une activité optimale sont un pH de 8 et une température de 40°C. Il s'agit d'une métallopeptidase inactivée par le chelatant EDTA. Elle présente le motif His-Glu-X-X-His caractéristique des métallopeptidases dépendantes du zinc.

Chez *L. lactis* NCDO763, les mutants de deux copies de *pepF* présentent un phénotype perturbé. Leur croissance est affectée en milieu minimum contenant uniquement les acides aminés essentiels. Ces résultats montrent que PepF est bénéfique dans ce type de milieu, dans lequel un stress causé par une source d'azote minimale provoque certainement la dégradation de protéines. Comme le taux de croissance est plus faible en absence de PepF, son possible rôle dans le recyclage (*turnover*) de protéines a été proposé (Nardi *et al.*, 1997).

Chez *Bacillus subtilis*, la sporulation est inhibée, lorsque le gène *pepF* est surexprimé, (Kanamaru *et al.*, 2002). Les analyses réalisées par les auteurs, indiquent que l'effet inhibiteur provient de la régulation négative de la phosphatase RapA. L'effet inhibiteur est dû à l'hydrolyse du peptide PhrA. Cette hydrolyse provoque la dérégulation de la phosphatase RapA et entraîne la déphosphorylation de Sp0A_P qui régule directement l'initiation de la sporulation. Quand elle est surexprimée, PepF hydrolyse d'autres peptides de la famille Phr, comme par exemple PhrC qui participe à la compétence (Kanamaru *et al.*, 2002), voir Figure 2-6. Il s'agit de mécanismes de type quorum sensing dans lesquels Phr est le peptide de signalement. L'hypothèse des auteurs est que PepF pourrait participer à la maturation de ce peptide. Néanmoins, ceci n'a pas été complètement prouvé.

En contradiction avec les connaissances chez *L. lactis* et *B. subtilis*, où PepF est une protéine cytoplasmique, chez *B. amyloliquefaciens*, une protéine sécrétée et semblable à PepF a été décrite. Il s'agit de la peptidase $PepF_{BA}$, qui apparemment participe au processus de sporulation (Chao *et al.*, 2006). Chez *Geobacillus collagenovorans*, Miyake et collaborateurs ont trouvé que l'homologue de PepF est nécessaire pour la dégradation du collagène, étant capable de dégrader de dimères et trimères des peptides de collagène Gly-Pro-Leu (Miyake *et al.*, 2005). Ils lui attribuent alors une fonction nutritionnelle. PepF a été caractérisée également chez *Caulobacter crescentus* (Braz *et al.*, 2002). Les auteurs ont confirmé que *pepF* n'est pas un gène essentiel comme cela avait été observé chez le lactocoque (Nardi *et al.*, 1997). Par ailleurs, ils ont testé la croissance de *C. crescentus* dans un milieu contenant des oligopeptides comme seule source d'azote et ont pu constater que d'autres peptidases peuvent prendre le relais en l'absence de PepF.

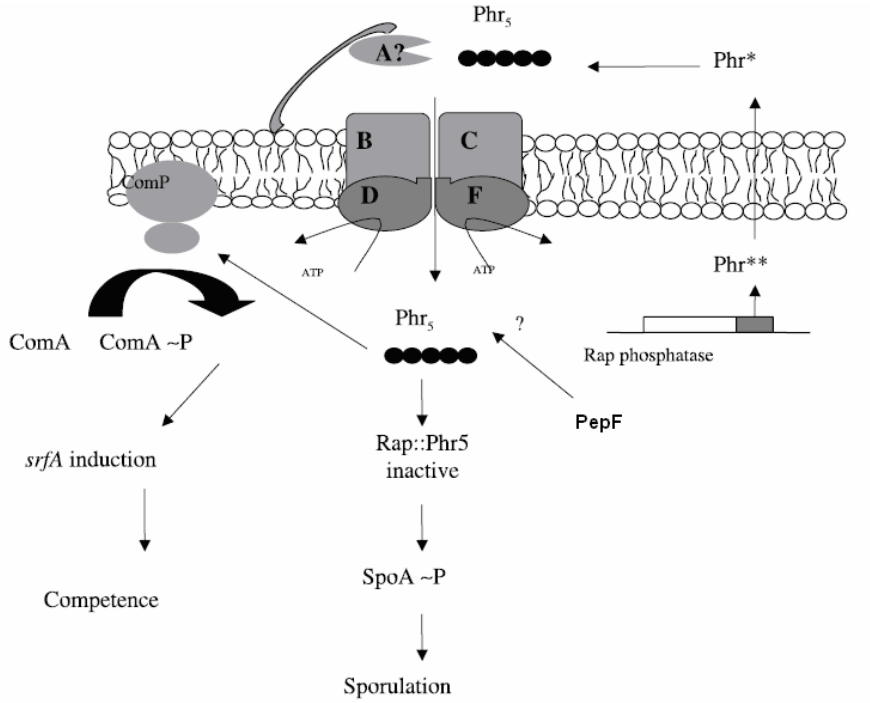


Figure 2-6 : Hypothèse concernant la participation de PepF dans les mécanismes de sporulation et compétence chez *B. subtilis* par hydrolyse des peptides de type Phr.

L'homologue de PepF chez *E. coli*, OpdA, a été étudié pour différents aspects. C'est une protéine qui présente 53% de similarité avec PepF de *L. lactis* au niveau du site actif sur une longueur de 68 acides aminées, entre les positions 378 et 439 pour PepF et 457 et 525 pour OpdA et une similarité globale de 10%. Comme pour les autres homologues de PepF, plusieurs rôles lui ont été attribués par son implication dans des processus très divers. Le gène codant OpdA a été identifié au cours d'une étude dont le but était de trouver des gènes participant dans l'export des protéines (Emr, 1982). Il avait été observé que LamB, localisé dans la partie extérieure de la membrane et nécessaire au transport du maltose chez *Escherichia coli*, était pourtant mal localisée chez certains mutants (Emr & Silhavy, 1980). Si cette protéine n'est pas bien localisée, la bactérie n'est pas capable de se développer dans un milieu contenant le maltose comme seule source de carbone. De plus, le phage λ utilise LamB comme récepteur. Si elle n'est pas présente à l'extérieur de la membrane, la bactérie devient résistante au phage λ. Cela fait deux phénotypes facilement observables qui ont permis à Emr & Silhavy (1980) de détecter des souches avec un défaut de localisation de LamB (un défaut de sécrétion). Le précurseur de LamB s'accumule dans le cytoplasme et n'arrive pas à être secrété pour prendre sa place à l'extérieur de la membrane. Dans une deuxième étude, Emr a cherché des « révertants » de ce phénotype dans le but de découvrir des gènes impliqués dans la sécrétion des protéines chez *E. coli*. Les trois loci identifiés ont été nommés *prlA*, *prlB* et *prlC* pour « protein localization ». C'est PrlC qui a été renommé ensuite OpdA. Ces loci affectent la sécrétion de la protéine LamB de manière différente et l'effet des mutations n'est observable que dans le cas d'une induction de la transcription de LamB (induction de la sécrétion) par croissance sur un milieu contenant du maltose (operon : *malK lamB*) (Emr, 1982). La participation de OpdA a été montrée également dans la sécrétion du gène P22 d'un bactériophage. La séquence du peptide signal n'est pas homologue à celle du LamB, mais les mutants de *opdA* ne sont pas capables d'exporter cette protéine non plus (Conlin *et al.*, 1992). Il s'agit de la même enzyme que celle identifiée par Novak comme une enzyme capable d'hydrolyser des peptides signaux après leur clivage chez *E. coli* (Novak *et al.*, 1986). Ces études montrent qu'OpdA participe

à la sécrétion car elle est capable d'hydrolyser des peptides signaux, mais il n'a pas été prouvé qu'il s'agisse d'une signal peptide peptidase (SPPase).

Vimr et al. (1983) ont, de leur côté, décrit OpdA comme une enzyme nécessaire chez *Salmonella typhimurium* pour utiliser N-acetyl-L-Ala_4 comme seule source d'azote. Chez *S. typhimurium* encore, Conlin & Miller ont montré que *opdA* est transcrit avec une ORF se trouvant en aval : *yhiQ* (Conlin & Miller, 2000). Ils ont montré que *opdA* et *yhiQ* forment un opéron qui est induit par un changement de température et que son induction dépend de RpoH (un facteur sigma). Ils ont trouvé une séquence promotrice proche de celle du facteur sigma32 pour *opdA*. Le rôle de ces deux protéines dans la réponse au changement thermique reste douteux, puisque les tailles prédites pour *opdA* et *yhiQ* ne correspondent pas aux protéines identifiées dans la réponse au stress.

En conclusion, PepF et ses homologues sont des métallopeptidases cytoplasmiques avec des spécificités de clivage très larges qui ont été décrites comme intervenant dans plusieurs processus pleïotropes.

2.2.4.2 YvjB ou Eep (enhanced expression of pheromone)

YvjB, composée de 428 acides aminés chez *L. lactis* et d'une masse théorique de 46 kDa, est une protéine contenant un motif métallopeptidase. Sa séquence permet la prédiction de quatre segments transmembranaires ce qui suggère une localisation dans la membrane. Aucune donnée n'existe sur YvjB chez *L. lactis*. Certaines informations peuvent être obtenues à partir de l'annotation du génome de *L. lactis* et la comparaison de la séquence de YvjB avec d'autres bactéries. Chez *L. lactis* les gènes *uppS, cdsA, yvjB* et *proS* sont probablement en opéron. Ces gènes codent respectivement l'undecaprényl diphosphate synthétase, la phosphatidate cytidil transférase, YvjB et la prolyl-ARNt-synthétase. L'undecaprényl diphosphate synthétase participe à la biosynthèse des polysaccharides de la paroi cellulaire. La phosphatidate cytidil transférase participe à la synthèse de phospholipides. La prolyl-ARNt-synthétase est chargée d'attacher le résidu prolyl (radical acide de l'acide aminé hydrophobe proline) à l'ARN (acide ribonucléique) de transfert.

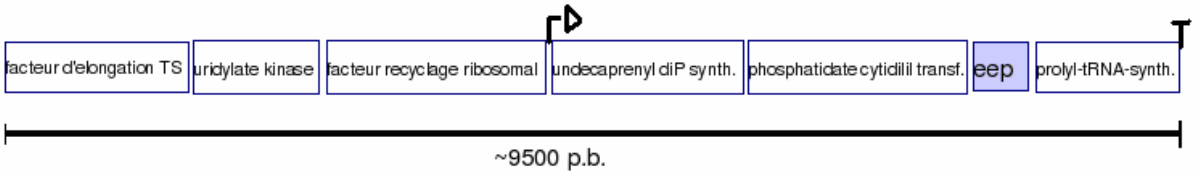


Figure 2-7 : Voisinage génétique de *yvjB=eep* chez *L. lactis.*

Comme YvjB n'a pas été étudiée, nous présenterons ci-dessous la famille à laquelle elle appartient et la fonction de certains de ses homologues. Des gènes similaires à environ 50% sont présents chez plusieurs bactéries Gram positives: *Streptococcus* sp., *Lactobacillus* sp., *Lactococcus* sp. *et Enterococcus* sp. Chez ces bactéries, le voisinage de *yvjB(eep)* est très conservé sur une région d'approximativement 9500 pb (Figure 2-7). Des homologues plus lointains de YvjB (environ 30% de similarité) sont présents chez les bactéries Gram négatives. Les protéines similaires à YvjB ont été regroupés dans la famille S2P (Rudner *et al.*, 1999). Rudner et collaborateurs décrivent cette famille comme une famille de métalloprotéases possédant les motifs HEXXH et NPDG, qui se trouvent à l'intérieur d'un segment transmembranaire (Rudner *et al.*, 1999). Les auteurs ont trouvé plusieurs protéines appartenant à cette famille (Figure 2-8), dont la protéine S2P humaine (Rudner

et al., 1999). Une troisième région conservée est présente chez plusieurs protéines de cette famille. Il s'agit d'un motif PDZ trouvé dans les protéines Yael et MmpA (*E. coli* et *Caulobacter* sp. respectivement) mais qui n'est pas présente dans SpoIVFB (Figure 2-9) (Chen *et al.*, 2005). Ce motif est impliqué dans la reconnaissance de polypeptides (Fuh *et al.*, 2000). Des protéases comme HtrA (*L. lactis*) et DegS (*E. coli*) possèdent des motifs PDZ. Dans le cas de cette dernière, le motif semble être impliqué dans la régulation négative de sa propre activité (Walsh *et al.*, 2003).

Les protéines de cette famille participent souvent à la protéolyse intermembranaire régulatrice (RIP) (voir 2.1.2.5) (Brown *et al.*, 2000). La RIP est un mécanisme où une protéase membranaire produit des peptides de signalement au cours de la dégradation de lipoprotéines et protéines membranaires. La protéolyse intermembranaire suit un premier clivage fait sur la protéine entière et agit donc sur un peptide de moins de 30 acides aminés. Dans la plupart des cas, la fonction du peptide enlevé est inconnue. Dans le système SREBP (sterol regulatory element-binding proteins) humain, par exemple, la protéine réalisant la RIP a été identifiée comme une métalloprotéase membranaire. Elle s'appelle S2P et fait partie des protéines similaires à YvjB. La théorie proposée par (Brown *et al.*, 2000) est que, à l'origine, les peptidases membranaires étaient chargées de dégrader des morceaux restants des lipoprotéines et que ce mécanisme a été développé, dans certains cas, pour devenir un mécanisme de régulation. Tel est le cas du RIP SREBP chez l'homme, où le peptide produit par S2P régule la transcription des gènes nécessaires pour la synthèse de lipides (Brown *et al.*, 2000).

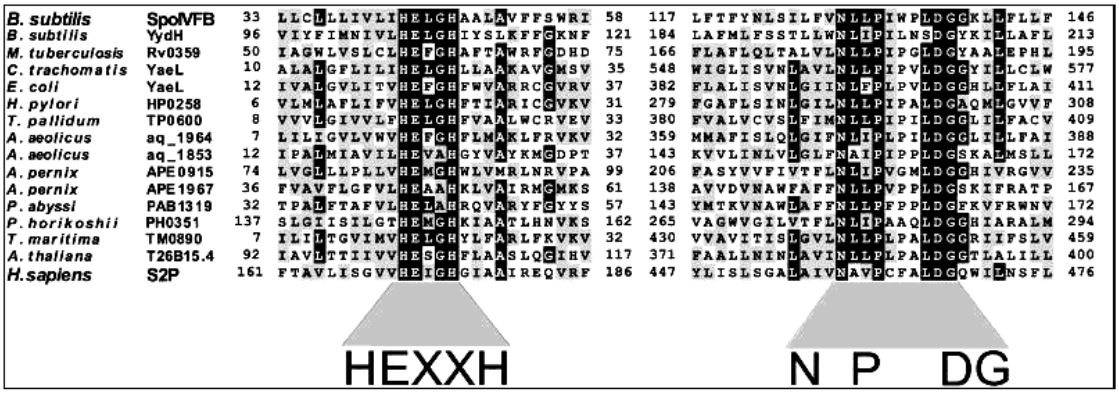


Figure 2-8: Alignement des protéines appartenant à la famille S2P montrant les motif conservés chez plus de 50% de protéines membres en noir (Rudner *et al.*, 1999).

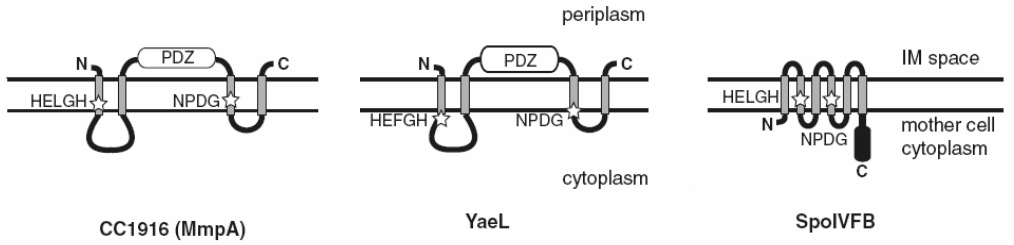


Figure 2-9 : Représentation schématique de la structure des protéines de la famille S2P chez *Caulobacter* sp. (MmpA) , *E. coli* (YaeL), *B. subtilis* (SpoIVFB) (Chen *et al.*, 2005).

Récemment, la phylogénie de la famille S2P a été étudiée par Kinch et collaborateurs. Ils confirment que les caractéristiques les plus conservées sont trois segments transmembranaires, dont

le premier et le troisième contiennent les sites catalytiques HEXXH et NPDG. Le domaine PDZ n'est pas présent chez tous les organismes. Le deuxième segment transmembranaire contient également un motif conservé : GXXXN/S/G. Sur la base de leurs résultats et la prédiction de la structure 3D prédite par Molscript (Esnouf, 1999), les auteurs proposent un modèle tridimensionel pour S2P dans lequel les deux sites catalytiques se trouvent l'un en face de l'autre et le substrat se trouve entre les deux (Kinch *et al.*, 2006), voir Figure 2-10.

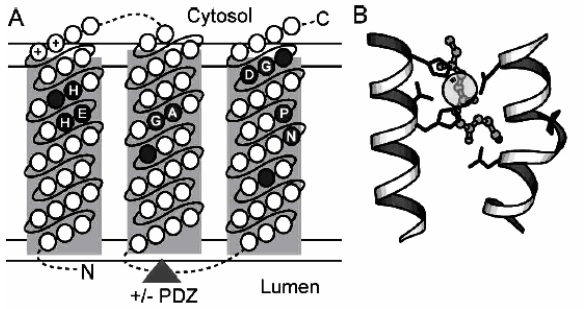


Figure 2-10 : Modèle de la topologie de S2P humaine. A) Représentation des segments transmembranaires de S2P indiquant les acides aminés conservés et le motif PDZ par un triangle. B) Modèle des deux sites catalytiques protéolytiques agissant sur le substrat représenté entre eux.

Les protéines les mieux caractérisées, appartenant à la famille S2P chez les bactéries sont : YaeL chez *E. coli,* YluC et SpoIVFB chez *B. subtilis* et Eep chez *Enterococcus faecalis* (Antiporta & Dunny, 2002; Chandler & Dunny, 2008) .

YaeL, aussi connu comme RseP (pour « regulator of sigma E , protease »), fait partie de la cascade de réponse au stress σ^E ; elle participe à la dégradation d'un fragment d'une protéine anti σ^E (Akiyama *et al.*, 2004; Bohn *et al.*, 2004). Dans des conditions normales, σ^E est séquestré par la protéine membranaire RseA. Dans le cas d'une accumulation des protéines mal repliées dans le périplasme, une réponse au stress est activée par le clivage de RseA par DegS. Ce premier clivage rend RseA sensible à la dégradation par YaeL, qui possède des sites actifs dans la membrane. La libération de σ^E permet la transcription de son régulon qui contient plusieurs protéines chaperonnes et des protéines participant dans la synthèse de lipides, lipopolysaccharides et la synthèse du peptidoglycane. De plus, les deux protéines YaeL et DegS sont nécessaires pour la dégradation de RseA, mais leur activité sur le substrat est indépendante. Par ailleurs, Bohn et collaborateurs ont pu déterminer que les sites protéolytiques de Yael sont essentiels pour la viabilité de la cellule, mais que le motif PDZ présent ne l'est pas. Même avec une délétion de 40 acides aminés dans le site PDZ, la reconnaissance et dégradation de RseA ne sont pas affectées (Bohn *et al.*, 2004).

YluC, l'analogue de YvjB chez *B.subtilis,* participe à la régulation de la division cellulaire en degradant FtsL par protéolyse (Bramkamp *et al.*, 2006), ainsi que dans la dégradation d'un facteur anti-σ^W (Schöbel *et al.*, 2004). La réponse σ^W est activé lors de stress alcalins et osmotiques et contrôle la transcription de gènes de détoxification et transport. Schöblel et collaborateurs (2004) ont montré que chez *B. subtilis* YluC n'est pas une protéine essentielle, mais les cellules du mutant négatif ne sont plus capables de sporuler.

SpoIVFB est l'un des membres de cette famille le mieux caractérisé et souvent utilisé comme exemple de la RIP (voir section 2.1.2.5). Cette protéine participe à la sporulation chez *B. subtilis* (Rudner *et al.*, 1999). Les auteurs ont analysé les deux motifs HEXXH et NPDG en détail par

mutagenèse dirigée. La substitution d'acides aminés dans le motif HEXXH arrête la maturation de précurseur du facteur s^{Ki}. Dans le motif NPDG, c'est uniquement la substitution de l'acide aspartique (D) qui affecte la maturation du précurseur ce qui confirme que l'asparagine fait partie du site actif. La structure de SpoIVFB proposée est en accord avec l'hypothèse d'un clivage des protéines substrats à l'intérieur de la membrane et de la proximité de l'acide aspartique de NPDG et du motif HEXXH pour accomplir un rôle de ligation de l'ion métallique. Néanmoins, la confirmation de ce mécanisme nécessite la purification des protéines et une étude biochimique détaillée (Rudner *et al.*, 1999).

Les travaux de l'équipe de Clewell ont permis de caractériser la fonction de la protéine Eep chez *E. faecalis.* Il s'agit d'une enzyme qui présente une similarité de 69% avec YvjB de *L. lactis*. Elle a été décrite comme participant à la production de la phéromone cAD1, ainsi que d'autres phéromones qui déclenchent la conjugaison. La phéromone cAD1 est produite par les bactéries ne possédant pas le plasmide pAD1 dans le but de trouver des donneurs de ce plasmide (An *et al.*, 1999). Une fois le plasmide acquis, la phéromone n'est plus détectée car la bactérie produit un peptide inhibiteur codé par un gène plasmidique. Pour pAD1, l'inhibiteur est appelé iAD1 et provient des huit derniers acides aminés d'un précurseur de 21 acides aminés. Les auteurs ont identifié la participation de Eep à la production de la phéromone grâce à la comparaison de la production de celle-ci entre des souches possédant Eep et d'autres souches ne possédant pas le gène codant pour cette protéine. Par ailleurs, elle participe aussi à la production d'autres phéromones. Dans une étude postérieure, An et Clewell décrivent le précurseur de la phéromone cAD1 (An & Clewell, 2002). Il s'agit d'une lipoprotéine de 309 acides aminés (Cad). L'octapeptide correspondant à la phéromone fait partie de la séquence signal de 22 acides aminés de cette lipoprotéine. Dans cette même étude, les auteurs ont montré que Eep participe aussi à la maturation de l'inhibiteur de la conjugaison, iAD1. Le précurseur est une petite protéine possédant également un peptide signal avec une séquence similaire a celle du précurseur de la phéromone (An & Clewell, 2002). Les auteurs concluent que la maturation des précurseurs de la phéromone et de l'inhibiteur doivent être le produit d'un travail d'équipe entre la signal peptidase II et Eep.

L'équipe de Dunny s'est également intéressée à la participation de Eep dans la production de phéromones (Antiporta & Dunny, 2002; Chandler & Dunny, 2008). Antiporta et Dunny (2002) ont décrit la participation de Eep dans la conversion du précurseur CcfA en peptide cCF10. Ce peptide déclenche également la conjugaison chez *E. faecalis* pour le transfert du plasmide pCF10. Là encore le précurseur est une lipoprotéine. La phéromone de sept acides aminés se trouve dans le peptide signal. Compte tenu des résultats obtenus par l'équipe de Clewell et de Dunny, un modèle pour la production de phéromones a été proposé (Antiporta & Dunny, 2002). Il implique l'utilisation de précurseurs de lipoprotéines et l'hydrolyse par Eep après un clivage par la signal peptidase II et finalement l'action d'une exopeptidase non identifié pour obtenir la phéromone mature (Figure 2-11). Par ailleurs, il a été montré que Eep reconnaît la séquence N-terminale du peptide signal qu'elle mature puisque la production de la phéromone cCF10 est affectée si cette extrémité est modifiée (Chandler & Dunny, 2008). Les auteurs ont étudié l'action de Eep sur des séquences semblables aux peptides signaux sans que la peptidase signal II les ait clivés. Ces résultats indiquent que l'action de Eep est indépendante de celle de la signal peptidase II (Chandler & Dunny, 2008).

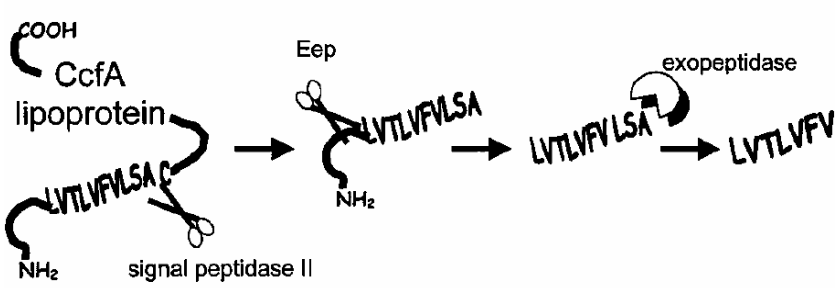


Figure 2-11 : Modèle pour la production de la phéromone cCF10 (Antiporta & Dunny, 2002).

Récemment, le clivage alternatif du peptide signal d'une lipoprotéine, réalisé par l'homologue de YvjB chez *Streptococcus uberis*, en absence de la signal peptidase II a été décrit (Denham *et al.*, 2008). Les auteurs proposent que l'homologue de d'YvjB pourrait participer au maintien de la sécrétion en cas d'absence de la signal peptidase II. En accord avec les clivages du peptide signal connus chez *E. faecalis* (Antiporta & Dunny, 2002), la partie clivée est plus courte que le peptide signal. Ces résultats indiquent que l'homologue de YvjB clive le peptide signal de lipoprotéines à un endroit différent que la signal peptidase II, mais il n'est pas très clair dans quel ordre se succèdent ces clivages et si un est nécessaire pour que l'autre puisse être réalisé.

En conclusion, les peptidases similaires à YvjB appartiennent à la famille S2P et participent à la maturation de protéines, principalement les lipoprotéines, libérant des peptides de signalement qui possèdent des propriétés de régulation à l'intérieur ou à l'extérieur de la cellule.

2.2.4.3 YaiH ou la prényl-peptidase CAAX

YaiH, composée de 236 acides aminés et d'une masse prédite de 27kDa, est une peptidase faisant partie de la famille de CAAX prényl-endopeptidases de type II (Pei & Grishin, 2001). La prénylation (addition de molécules hydrophobes à une protéine) est cruciale pour le fonctionnement et l'identification de plusieurs protéines chez les eucaryotes (Clarke, 1992). Le groupement prényl est attaché à la cystéine du motif CAAX. Ensuite, le tripeptide AAX est enlevé par une prényl-endopeptidase. Deux types de prényl-endopeptidases sont connus. Ceux de type I sont des métalloprotéases dépendantes du zinc avec un motif HEXXH. Les prényl-endopeptidases de type II possèdent trois autres résidus conservés, potentiellement impliqués aussi dans la liaison d'un ion métallique (Pei & Grishin, 2001). Il s'agit d'une superfamille de prényl-endopeptidases CAAX retrouvée chez les archaea et les plantes, en particulier *Arabidopsis thaliana*, où elle a été étudiée en détail (Cadiñanos *et al.*, 2003b).

Les prényl-endopeptidases ont surtout été étudiées chez les eucaryotes, leur rôle chez les procaryotes reste mal connu. Chez les eucaryotes, elles participent à la maturation post-traductionnelle des protéines. Ainsi chez *Caenorhabditis elegans*, par exemple, deux prényl-endopeptidases CAAX ont été identifiées. Elles sont essentielles pour la maturation des Ras-GTPases qui contiennent le motif CAAX (Cadiñanos *et al.*, 2003a). De plus, des protéines appartenant à cette famille ont été étudiées chez des parasites comme *Tripanosoma* sp. (Gillespie *et al.*, 2007).

Ce qui semble intéressant chez cette famille de peptidases est son universalité dans le monde vivant. Cependant, le voisinage des prényl-endopeptidases n'est pas conservé. De plus, chez *L. lactis*,

plusieurs copies avec différents degrès d'homologie à la prényl-peptidase de référence (CAAX de *L. lactis* subsp. *cremoris*, utilisée dans la description de cette famille (Pei & Grishin, 2001)) existent (Tableau 2-2). Ces copies sont entourées d'espaces intergéniques grands et/ou de transposases, indiquant des possibles transfert horizontaux. Un indice de plus de ces acquisitions est donné par la déviation de ces gènes en ce qui concerne l'indice d'usage de codons moyen de *L. lactis* IL1403. La copie la plus homologue à la prényl-endopeptidase de référence présente un biais par rapport à cet indice, renforçant l'hypothèse d'un transfert horizontal (Figure 2-12).

Tableau 2-2 : Similarité des copies des prényl-endopeptidases trouvés chez *L. lactis* IL1403 avec la prényl-peptidase de *L. lactis* subsp. *cremoris* utilisée dans la description de la famille des prényl-endopeptidases par Pei et Grishin (2001).

	Gène	**E-value**	**Similarité (%)**	**Identités**
1	*yaiH*	2^{-110}	83	191/230
2	*yaiF*	2^{-20}	32	72/222
3	*yajF*	2^{-18}	31	63/200
4	*ybdJ*	8^{-16}	29	66/221
5	*ybdI*	2^{-15}	31	51/163
6	*yxdF*	2^{-13}	31	56/178
7	*yvdC*	2^{-5}	27	60/222

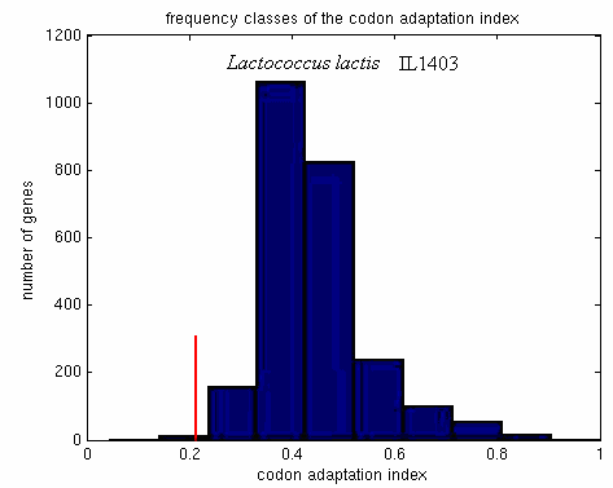


Figure 2-12 : Histogramme du nombre de gènes appartenant à une classe d'indice d'usage de codons. Le trait rouge indique la valeur de cet indice pour *yaiH*.

En conclusion, YaiH est la peptidase d'intérêt la moins connue chez les bactéries et dont plusieurs copies sont présentes chez *L. lactis* IL1403. Pour cette raison, nous avons étudié en priorité les peptidases PepF et YvjB au cours de ce travail.

2.3 Prédiction du rôle des protéines

Il est difficile de savoir quel est le nombre de gènes codant des protéines de rôle inconnu dans un organisme au génome séquencé. Rares sont les travaux rapportés dans la littérature concernant ce sujet. Fukuchi et Nishikawa se sont par exemple intéressés à la prédiction du nombre de gènes potentiels dans les génomes bactériens (Fukuchi & Nishikawa, 2004). Selon les auteurs le nombre de gènes potentiels varie de 25 à 50%. Ils démontrent que la proportion de gènes potentiels est corrélée à « l'index d'isolement de l'organisme », mesuré par la similarité de séquences. Cela permet de connaître le nombre attendu de vrais gènes potentiels dans un génome. Le nombre de protéines de rôle inconnu est en tout cas très grand et cela a motivé le recours à des approches statistiques et mathématiques afin de prédire de rôles de protéines. Certaines de ces approches seront présentées dans cette section.

2.3.1 Approches existantes

Le séquençage qui a permis d'identifier des gènes codant pour des protéines inconnues a, dans un même temps, permis l'obtention de données « génomiques » à haut débit telles les données transcriptomiques. Pour aller plus loin dans la prédiction des rôles des protéines identifiées par le séquençage que l'annotation des génomes, plusieurs méthodes ont été proposées. Elles utilisent les données génomiques disponibles pour tous les gènes d'un organisme pour estimer des réseaux métaboliques, de régulation de gènes ou des réseaux d'interaction de protéines, ou encore pour classer les protéines dans un groupe fonctionnel. Compte tenu de l'avènement des données génomiques à haut débit, des méthodes d'analyse et prédiction sont publiées constamment depuis la fin des années 1990. Une des premières revues de méthodes utilisées pour la modélisation et la simulation de réseaux de régulation de gènes a été réalisée en 2002 (de Jong, 2002) et déjà une panoplie de méthodes basées sur différents modèles sont présentées : méthodes bayésiennes, réseaux booléens, équations différentielles ordinaires et aux dérivées partielles, équations stochastiques, formalismes à base de règles. Depuis, les mêmes modèles, utilisés de manière différente ou de nouveaux modèles ont été mis à l'épreuve pour la reconstruction et la modélisation de réseaux de gènes et il est presque impossible de faire une revue exhaustive de toutes ces méthodes. On distingue les méthodes selon deux critères essentiels qui sont leur caractère supervisé ou non supervisé et leur acceptation ou non acceptation de données hétérogènes. Dans cette section, je présenterai donc quelques méthodes relevant de ces critères (Tableau 2-3) et m'étendrai plus particulièrement sur l'analyse de corrélation canonique bâtie sur des noyaux ou *kernel canonical correlation analysis* (KCCA), qui nous a servi pour la prédiction des rôles des protéines.

Tableau 2-3 : Synthèse des méthodes de prédiction des rôles des protéines. NS : Non-supervisée. S : Supervisée. Int : intégration de données ;

	Auteurs	NS	S	Int	Modèle
Approches basées sur des modèles statistiques	Friedmann *et al.* 2000	oui	-	-	Réseaux bayésiens
	Kauffman 1993	oui	-	-	Booléen
	de Hoon *et al.* 2002	oui	-	-	Equations différentielles
	Schäfer & Strimmer 2005	oui	-	-	Modèle Gaussien
	Meinshausen & Bühlmann 2006	oui	-	-	Modèle Gaussien
	Westra *et al.* 2007	oui	-	-	Equations différentielles
	Werhli & Husmeier 2007	-	oui	oui	Réseaux bayésiens
Approches basées sur la similarité	Yamanishi *et al.* 2003	-	oui	oui	Méthode à noyaux (KCCA)
	Covert *et al.* 2004	-	oui	oui	Booléen
	Qi *et al.* 2005	-	oui	oui	Forêts aléatoires
	Kato *et al.* 2005	-	oui	oui	Méthode à noyaux (em)
	Aerts *et al.* 2006	-	non	oui	Corrélation
	De Bie *et al.* 2007		oui	oui	Méthode à noyaux (SVM)
	Bleakley *et al.* 2007	-	oui	oui	Méthode à noyaux (SVM local)
	Von Mering *et al.* 2007	-	oui	oui	Transfert d'interactions connues
Text mining	Kim *et al.* 2007	-	oui	-	Méthode à noyaux

Le défi est d'utiliser les données disponibles d'une manière intelligente pour prédire le rôle des protéines inconnues alors que le nombre d'échantillons (hybridations sur microarray, génomes séquencés, etc.) est très inférieur au nombre d'objets (gènes). Les données de prédiction ou d'inférence de rôles vont utiliser les données génomiques pour produire des réseaux de protéines et renseigner comment les protéines sont reliées entre elles. Les liaisons peuvent être de différentes sortes comme la catalyse de deux réactions successives dans une voie métabolique, ou aussi des relations de régulation génétique.

A ce stade, il est important de clarifier que l'objet d'étude de ces méthodes est le gène ou la protéine codée par un gène, selon le type de données utilisé. Même si la différence au niveau biologique est claire, les deux objets sont confondus dans les analyses qui visent la prédiction de rôles de protéines. Le résultat de ces prédictions est toujours un réseau d'interaction entre protéines, puisque la participation à une même fonction ou une fonction proche est prédite. Par contre, les données qui sont utilisées pour l'analyse peuvent concerner les gènes codant cette protéine (les données transcriptomiques, profils phylogénétiques, distance sur le chromosome) ou les protéines (données double-hybride, localisation dans la cellule, graphe des voies métaboliques) ou toute autre information y compris la littérature scientifique.

2.3.1.1 Méthodes non supervisées et sans intégration des données

Modèles dynamiques

Ces modèles concernent des données temporelles. Une des premières méthodes pour inférer un réseau de régulation de gènes a été présenté par Kauffman (Kauffman, 1993). Elle considère des données quantitatives obtenues pour une collection de protéines. C'est une méthode basée sur des réseaux booléens. Dans cette approche, une des premières approximations qui est faite est de décrire l'état d'un gène comme actif (1) ou comme non actif (0) et donc si les produits de ces gènes sont absents ou présents. Les interactions entre gènes peuvent être représentées par des fonctions booléennes qui calculent l'état d'un gène à partir de l'activation d'autres gènes. Le résultat est un réseau booléen. Un vecteur $\hat{x}$ de taille *n* représente l'état d'un système de régulation de *n* éléments. Chaque $\hat{x}_i$ a la valeur 1 ou 0, et l'espace du système pourrait comporter 2^n états. L'état $\hat{x}_i$ d'un élément au temps $t+1$ est calculé par une fonction booléenne, ou règle $\hat{b}_i$ à partir de l'état de *k* des *n* éléments au moment *t* (*k* peut différer d'un élément à l'autre). En résumé, les dynamiques d'un réseau booléen décrivant un système de régulation sont données par : $\hat{x}_i(t+1) = \hat{b}_i(\hat{x}(t))$, $1 \le i \le n$, où $\hat{b}_i$ projette *k* entrées sur une valeur de sortie. Un exemple est montré en Figure 2-13 (de Jong, 2002) . Si les trois gènes sont inactivés au temps t (000), en utilisant les fonctions (les règles) de l'exemple, au temps $t+1$, le système va changer pour l'état (011), où deux gènes seront activés (voir le diagramme (*wiring diagram*) Figure 2-13 c).

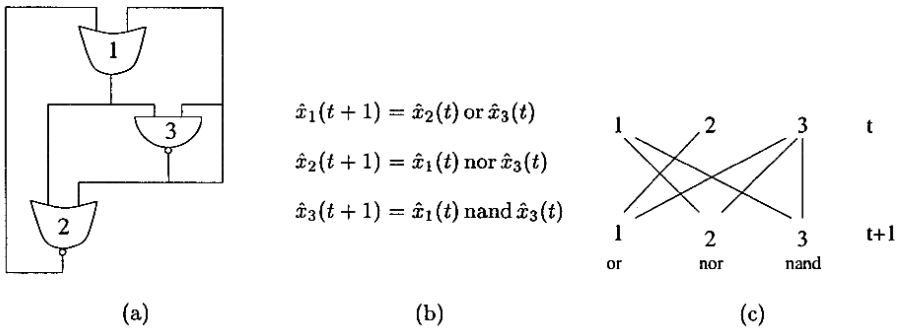


Figure 2-13 : a) exemple de réseau booléen, b) les fonction correspondantes, c) le *wiring* diagramme (de Jong, 2002).

Les réseaux booléens sont faciles d'application et peuvent être analysées efficacement, néanmoins, ils ne permettent pas de tenir compte de désynchronisations du changement d'état des gènes (de Jong, 2002).

Les équations différentielles sont aussi très utilisées pour modéliser des systèmes dynamiques en science et sont aussi souvent employées pour l'analyse et l'inférence de réseaux de gènes. De Jong

présente plusieurs cas d'étude où des équations de ce type sont utilisées pour modéliser le taux de production d'un composé au cours du temps. Ces modèles sont basés sur des équations différentielles ordinaires, linéaires par morceaux, qualitatives, aux dérivées partielles (de Jong, 2002). Je présenterai uniquement un exemple de l'utilisation d'équations différentielles pour la construction de réseaux de gènes proposé par Westra et collaborateurs (Westra *et al.*, 2007). Cet exemple est basé sur un modèle dynamique linéaire par morceaux (*piecewise linear state space model*) pour la reconstruction des réseaux de gènes à partir de données transcriptomiques. Dans un modèle dynamique les relations entre gènes peuvent être décrites en temps continu ou discrets. Pour modéliser les dynamiques du réseau d'interaction les auteurs proposent d'utiliser une équation différentielle stochastique : $\dot{x} = f(x,u|\theta) + \xi(t)$, où *x(t)* est le vecteur d'état (state vector) qui donne l'expression des *N* gènes (niveau d'ARN) au moment *t* , *u(t)* représente les *P* variables de contrôle du système (comme par exemple les agents toxiques) et $\xi(t)$ est un terme de bruit blanc gaussien. Ce formalisme permet de représenter un système qui passe par différents états transitoires et stationnaires (attracteurs). Si on associe à chaque attracteur une cellule *l* dans l'espace des configurations, alors, sous l'influence d'effets externes *u(t)* ou aléatoires $\xi(t)$ il peut s'éloigner d'un attracteur et continuer son parcours. Ainsi, il est possible d'approximer la dynamique du système, dans une cellule *l*, autour du point d'équilibre $x_{eq}^{(l)}$ par:

$$\dot{x}(t) \approx \frac{\partial f(x_{eq}^{(l)},u)}{\partial x}(x - x_{eq}^{(l)}) + \frac{\partial f(x_{eq}^{(l)},u)}{\partial u}u \equiv A_l x(t) + B_l u(t) + c_l$$

Autour du point d'équilibre, le système dynamique est representé par un modèle linéaire. L'approche de Westra consiste alors à identifier des systèmes de régulation de gènes à partir de données transcriptomiques temporelles. Pour cela les auteurs estiment les matrices d'interactions de gènes locales (par morceaux) A_t, directement reliées au graphe du réseau de régulation de gènes. De plus les matrices *B* vont donner l'information qui relie les gènes aux entrées qui perturbent le système.

Les approches présentées ci-dessus sont basées sur la modélisation des interactions selon un modèle précis et il est nécessaire que la dynamique des interactions s'adapte au modèle pour que la reconstruction des réseaux soit correcte.

Modèles statistiques :
Au lieu d'essayer de caler un modèle aux interactions entre gènes, il est possible d'utiliser des modèles statistiques basés sur la probabilité d'une relation entre gènes. On peut noter parmi ces

modèles les réseaux bayésiens (Friedman *et al.*, 2000) et les modèles graphiques gaussiens (Meinshausen & Bühlmann, 2006; Schäfer & Strimmer, 2005).

Modèles statistiques : Réseaux bayésiens

Friedman et collaborateurs proposent une procédure, basée sur des réseaux bayésiens (Friedman *et al.*, 2000) et utilisant des mesures d'expression de gènes, pour déceler des interactions de gènes chez la levure. Un réseau bayésien est un modèle qui représente la dépendance probabilistique entre objets interagissants. Il est défini par un graphe $G = \langle V, E \rangle$. Les nœuds $i \in V$, $1 \leq i \leq n$ représentent les gènes (ou d'autres objets) et correspondent aux variables X_i. Si *i* est un gène, alors X_i décrit le niveau d'expression de *i.* Pour chaque X_i une distribution conditionnelle $p(X_i | parents(X_i))$ est définie où *parents*(X_i) désigne le régulateur direct de *i* dans *G.* Le graphe *G* et la distribution conditionnelle $p(X_i | parents(X_i))$ définissent un réseau bayésien et une distribution de probabilité associé *p(X).* On a donc indépendance conditionnelle de X_i,*nondesc*(X_i)|*parents*(X_i), soit : $p(X) = \prod_{i=1}^{n} p(X_i | parents(X_i))$. Friedman et collaborateurs ont proposé un algorithme heuristique pour la reconstruction de réseaux bayésiens à partir de données d'expression de gènes et prennent en compte le fait que le nombre de gènes dépasse largement le nombre d'échantillons. Au lieu d'essayer de retrouver un seul réseau qui représente au mieux un jeu de données d'expression de gènes, ils explorent des réseaux avec des caractéristiques communes (appartenant à une même classe de réseaux). Ils cherchent des relations de Markov et causales entre paires de variables X_i et X_j (Friedman *et al.*, 2000).

Les réseaux bayésiens peuvent être utilisés même quand le système n'est pas complètement spécifié. De plus, ils permettent de réduire la complexité du problème de reconstruction sans avoir recours à des modèles très précis. Les réseaux bayésiens statiques comportent certains limites et notamment l'impossibilité de prendre en compte des feedbacks. Certains de ces limites sont levés par les réseaux bayésiens dynamiques (Werhli & Husmeier, 2007), voir section 2.3.1.2 .

Modèles statistiques : Modèles graphiques gaussiens

Plusieurs méthodes d'interaction entre entités décrites par un ensemble de « mesures dans R » ont été proposées ces dernières années pour retrouver des relations entre gènes à partir de données d'expression. Chacun des *p* gènes étudié est alors décrit par un vecteur $X_i \in R^n$ correspondant aux *n* conditions expérimentales. Ces *n* données sont considérées comme des répétitions et on modélise

$X = (X_i,..., X_p)$ par un modèle graphique gaussien, soit $X \sim N(\mu, \Sigma)$. De cette modélisation résulte la propriété suivante :

$X_i | X_{-i} \sim \mathrm{N}(\mu_i, \sigma_i)$, *i=1,...,p* où $\mu_i = \sum_{j \neq i} \theta_j^i x_j$ et σ_i^2 ne dépend pas de X_{-i}, soit encore

$X_i = \sum_{j \neq i} \theta_j^i X_j + \varepsilon_i$ avec $\mathrm{E}(\varepsilon_i) = 0$ et ε_i indépendant de $X_j \ \forall_{j \neq i}$, $X_{-i} = (X_i,..., X_{i-1}, X_{i+1},..., X_p)$.

La loi de X_i conditionnelle à X_{-i} est donc gaussienne. Il y a alors équivalence entre la nullité de certains coefficients θ_j et l'indépendance conditionnelle.

A un modèle gaussien on peut faire correspondre un graphe non dirigé $G = (V, E)$ ayant pour sommets les *p* variables et pour arête l'ensemble *E*. Il y a une arête entre les sommets *i* et *j* si et seulement si, les variables X_i et X_j sont conditionnellement dépendantes toutes les autres variables $X_{-\{i,j\}}$ données. De cette correspondance entre modèle gaussien et graphe vient le terme Modèle Graphique Gaussien (Lauritzen, 1996).

On définit alors le voisinage de *i* comme les sommets *j* pour lesquels il existe une arête entre *i* et *j* et qui ont la propriété de Markov locale. *X* vérifie la propriété de Markov locale au sommet *i* par rapport au graphe *G* si X_i est indépendant de $\{ X_j, j \in V \setminus (V(i) \cup \{i\}) \}$ conditionnellement à $\{X_j, j \in V(i)\}$. D'un point de vue pratique, inférer des relations fonctionnelles entre gènes revient alors, dans le cadre d'un modèle graphique gaussien, à reconstruire le graphe *G* et une arête entre sommets signifie une dépendance entre les données d'expression entre chacun de ces gènes, même lorsque l'on considère les données d'expression liées aux autres gènes. Des indépendances sont trouvées en inversant la matrice Σ, $\theta_j^i = -\frac{\Sigma_{i,j}^{-1}}{\Sigma_{i,i}}$.

Dans le cadre de l'inférence à partir de données d'expériences transcriptomiques, le nombre de variables *p* est souvent très supérieur au nombre d'expériences et la matrice Σ n'est pas inversible. Partant de là plusieurs auteurs (Meinshausen & Bühlmann, 2006; Schäfer & Strimmer, 2005; Verzelen & Villers, 2007) ont proposé des méthodes d'estimation du graphe (c'est à dire de coefficients de corrélations partielles θ_j^i).

Une des premières méthodes basée sur un modèle graphique gaussien dans un cadre bayésien a été proposée par (Schäfer & Strimmer, 2005). Pour la construction d'un graphe, les corrélations élevées entre deux gènes, qui indiquent une interaction directe, sont utilisées. Les corrélations par paires se trouvent dans la matrice de corrélations partielles $\prod = (\pi_{ij})$. Ces coefficients représentent la

corrélation entre gènes *i* et *j* conditionnellement à tous les autres gènes. La matrice $\prod$ peut être obtenue par sa relation à l'inverse de *P*, la matrice de corrélations. Pour reconstruire un graphe d'interaction de gènes à partir de la matrice de corrélations partielles estimée $\hat{\prod}$, des tests statistiques sont réalisés pour déterminer quelles sont les corrélations significativement différentes de zéro et déduire un graphe dont les arêtes correspondent à des corrélations non nulles. Comme la matrice de covariance n'est pas inversible, les auteurs proposent alors d'utiliser une matrice pseudo-inverse de Moore-Penrose. La matrice *P* peut être décomposée en $P = UDV^T$ où *D* est une matrice diagonale carrée de rang $m \leq \min(n, p)$. La matrice pseudo-inverse P^+ est alors définie comme $P^+ = VD^{-1}U^T$ où uniquement l'inversion de *D* est nécessaire. De plus, une approche bootstrap est utilisée pour stabiliser l'estimateur de $\prod$. Enfin, les auteurs utilisent des tests multiples basés sur le taux de faux positifs (FDR : false discovery rate, $FDR = \mathrm{E}\left(\frac{FP}{TP+FP}\right)$, où FP=faux positifs et TP=vrais positifs) pour choisir les corrélations significativement différentes de zéro. D'abord une liste de *p-values* ordonnée est calculée, une valeur pour chaque arête potentielle. Ensuite, l'hypothèse nulle d'une corrélation nulle est rejetée pour un nombre d'arête contrôlant le FDR. Cette méthode a été implémentée sous R.

Meinshausen et Bühlmann proposent une méthode alternative s'appuyant sur le *Lasso* (Meinshausen & Bühlmann, 2006). Le *Lasso* est une méthode de sélection en régression linéaire. Elle permet de minimiser la somme de carré des écarts habituelle avec une contrainte sur la somme $s = \sum|\theta_j|$ de valeurs absolues des coefficients. De manière équivalente cette approche consiste à optimiser une fonction objective pénalisée dans un espace de grande dimension : $\min_{\theta} \sum_{l=1}^{n} (x_i^l - \sum_{j \neq i} \theta_j^i x_j^l)^2 + \lambda \sum_{j \neq i} |\theta_j^i|$, où λ est un paramètre de pénalisation. Cette minimisation peut conduire à deux graphes distincts selon la définition de voisinage choisie : $i \sim$ j si $\theta_j^i \neq 0 \vee \theta_i^j \neq 0$ ou encore $i \sim$ j si $\theta_j^i \neq 0 \wedge \theta_i^j \neq 0$. Si *s* est très large (ou λ faible), la contrainte n'a pas d'effet et la solution est celle des moindres carrés habituelle. Néanmoins, pour des *s* plus faibles, les solutions sont différentes. Souvent *s* est choisi par validation croisée. Les auteurs proposent aussi de contrôler le taux d'erreur (Meinshausen & Bühlmann, 2006).

Une méthode récente, pour tester la complétude d'un graphe d'un modèle gaussien, a été proposée (Verzelen & Villers, 2007). Au contraire des méthodes présentées plus haut il s'agit de tester un

graphe existant et non pas de déduire un nouveau. Si nous supposons l'existence d'un vecteur *X* avec *n* échantillons et un graphe *G,* la méthode proposée consiste à tester l'hypothèse : « *X* suit la propriété de Markov locale au sommet *i* par rapport au graphe *G* » contre l'hypothèse contraire. C'est un test de voisinage dont l'objet est de tester s'il y a des voisins manquants. Ces tests sont réalisés dans un cadre où la puissance : $E\left(\frac{TP}{TP+FN}\right)$, où TP=vrais positifs, FN=faux négatifs, est contrôlée.

Les modèles statistiques présentés dans cette section reposent sur un formalisme solide et permettent de tenir compte des aspects stochastiques de l'expression de gènes et du bruit inhérent à ces données.

2.3.1.2 Méthodes supervisées et permettant l'intégration des données

Il est clair que les méthodes présentées plus haut peuvent être utilisées dans certains cas, par exemple quand le modèle des interactions entre gènes est bien caractérisé et peut être modélisé facilement ou quand les données disponibles reflètent très bien les interactions recherchées. Cependant, dans le cas de questions plus générales, comme la reconstruction de réseaux de gènes globaux, des méthodes plus générales sont nécessaires. De plus, souvent des informations autres que le niveau d'expression des gènes sont disponibles sur les gènes eux mêmes et les protéines codées par eux : interactions physiques obtenues par double-hybride, voies métaboliques, profils phylogénétiques, localisation dans la cellule (chez les eucaryotes), structure de protéines, relations obtenues par l'analyse de textes, entre autres. Dans cette section nous présentons des méthodes de reconstruction de réseaux de gènes qui incluent un apprentissage à l'aide de parties connues du réseau (méthodes supervisées), ainsi que des méthodes qui permettent l'utilisation de différents types de données. Les formalisme pour pouvoir réaliser un apprentissage et pour pouvoir inclure des données provenant de sources diverses sont très différentes d'une méthodes à l'autre. Encore une fois le nombre d'approches existantes ne permet pas de faire une revue exhaustive. Comme la méthode utilisée pour faire les prédictions de rôles de protéine dans ce travail est supervisée et permet l'intégration de données, d'autres méthodes semblables seront présentées. Les méthodes choisies montrent une évolution dans la manière d'approcher la reconstruction de réseaux de gènes. D'abord, les méthodes ont été améliorées par l'intégration de différentes sources de données, ce qui a, dans un premier temps, conduit à une simplification des données (Aerts *et al.*, 2006; Covert *et al.*, 2004; Qi *et al.*, 2005) pour passer ensuite à des représentations des données plus fidèles et

permettant de garder le plus d'information possible comme les noyaux (De Bie *et al.*, 2007; Kim *et al.*, 2008; Yamanishi *et al.*, 2003). Une autre évolution a été faite passant d'une simple classification des gènes en deux classes (interagissant, non-interagissant) (Qi *et al.*, 2005) à une modélisation plus complète des interactions (Werhli & Husmeier, 2007). De nouvelles approches statistiques plus sophistiquées basées sur la similarité et permettant le contrôle de faux positifs ont été aussi développées (Aerts *et al.*, 2006; De Bie *et al.*, 2007).

Méthode itérative basée sur un modèle booléen

Dans cet esprit d'intégration d'informations déjà existantes, Covert et collaborateurs ont construit un réseau métabolique et de régulation de gènes booléen (basé sur des règles simples) intégrant des données disponibles dans la littérature et des bases de données (Covert *et al.*, 2004). L'approche proposée par les auteurs est une approche itérative consistant d'abord en la construction du modèle avec les données disponibles (littérature et bases de données) et ensuite en la réalisation des expériences de laboratoire. Les résultats des expériences sont ensuite utilisés pour corriger le modèle qui avait prédit un phénotype de croissance contraire au phénotype observé. De cette manière les expériences ont permis de compléter et corriger le modèle. La même approche a été utilisée pour prédire et ensuite vérifier les gènes dont la transcription serait affectée par un manque d'oxygène. L'approche est résumée en Figure 2-14.

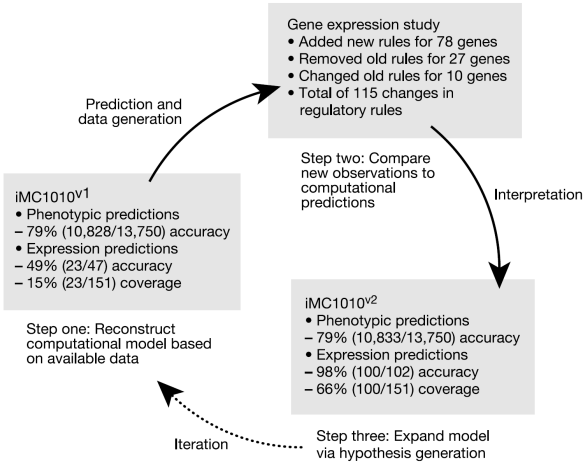


Figure 2-14 : Représentation de l'approche utilisé par (Covert *et al.*, 2004) où d'abord un modèle booléen est construit, puis vérifier et compléter avec des données expérimentales, ce qui conduit à un modèle plus complet.

Transfert de connaissances sur les relations entre gènes

Il est important de citer la base de données relationnelle STRING (http://string.embl.de/) (Von Mering *et al.*, 2007) ici, parce qu'il s'agit d'un outil qui cherche à prédire des relations entre protéines. Il ne s'agit pas vraiment d'une méthode mais d'un programme qui transfère les relations connues entre protéines chez un organisme modèle aux organismes moins connus. D'abord,

l'information sur les relations connues entre protéines est recueilli dans plusieurs bases de données et complétée par des recherches de text-mining. Ensuite, des interactions sont prédites compte tenu de la similarité entre protéines, les motifs communs et le voisinage de gènes (Von Mering *et al.*, 2007). Dans STRING les prédictions sont faits sur la base des données qui sont disponibles dans la base. Si par exemple uniquement des données de voisinage sont disponibles, les prédictions seront faites uniquement sur la base du voisinage et seront alors assez incomplètes.

Méthodes statistiques : forêts aléatoires

Qi et collaborateurs proposent une approche fondée sur la similarité de classement à l'aide de forêts aléatoires (*random forests)* pour prédire l'interaction de paires de protéines. Ils font l'apprentissage d'arbres de décision (qui composent la forêt aléatoire) en s'appuyant sur des sources de données différentes (Qi *et al.*, 2005). Les données génomiques sont utilisées pour la construction d'un vecteur X_i à d dimensions pour chaque paire de protéines. Chaque entrée de ce vecteur résume un jeu de données donnant un attribut à la paire. Par exemple, une valeur qui exprime si les deux gènes sont régulés par le même facteur de transcription ou le coefficient de corrélation entre leurs profils d'expression. Une fois ce vecteur construit la prédiction d'interaction entre gènes est abordée par les auteurs comme un problème de classification binaire : la paire de protéines *i* interagit-elle ($Y_i=1$) ou pas ($Y_i=-1$)? Néanmoins, il est nécessaire de tenir compte de certaines caractéristiques propres à ce problème de classification. Le nombre de paires non-interagissantes est beaucoup plus élevé que le nombre de paires interagissantes. Chaque paire de protéines est résumée par un vecteur X_i, aux composants scalaires ou catégoriels. Pour gérer ces difficultés, la classification est divisée en deux parties : I) Utilisation d'une forêt aléatoire (collection d'arbres de décision) pour déterminer la similarité entre paires de protéines. Plusieurs arbres de décision sont construits par un échantillonnage bootstrap. Un nombre m<<M (M étant le nombre total d'attributs) est choisi pour la séparation d'un noeud, m est choisi aléatoirement à partir de M. Pour une forêt aléatoire ainsi construite, la similarité entre deux paires de protéines X_1 et X_2 est calculée en propageant leur valeurs le long de la forêt. La position finale de chaque paire est retenue. Si $Z_1 = (Z_{11},...,Z_{1K})$ (resp. Z_2) est la position de X_1 (respectivement X_2) dans les arbres de la forêt, la similarité entre X_1 et X_2 est : $S(X_1, X_2) = \frac{1}{K}\sum_{i=1}^{K} I(Z_{1i} == Z_{2i})$. II) Utilisation de la similarité pour classifier les paires de protéines en interagissants ou non-interagissants basée sur un algorithme de k plus proches voisins. Les auteurs ont appliqué l'algorithme pour reconstruire une réponse à une phéromone connue chez la levure retrouvant parmi les interactions détectées 70% d'interactions connues, 5% d'interactions

fausses et 25% d'interactions putatives.

Méthodes statistiques : réseaux bayésiens (statiques ou dynamiques)

Werhli et Husmeier proposent une approche pour la reconstruction de réseaux de régulation de gènes basée sur des réseaux bayésiens. Leur méthode permet l'utilisation de différents types des données (en plus des données transcriptomiques) provenant de connaissances *a priori* (Werhli & Husmeier, 2007). Chaque source, *s*, d'*a priori* est représentée par une matrice $B^s = (B_{ij}^s)_{i,j=1...p}$ où B^s est une matrice traduisant l'*a priori* ; B_{ij}^s vaut 0.5 lorsqu'on a pas d'*a priori* par rapport à une arête entre *i* et *j*, >0.5 (resp. <0.5) s'il y a une forte (faible) présomption pour l'existence d'une arête. Ils définissent alors une fonction qui mesure la correspondance entre un graphe *G* et les connaissances biologiques *a priori* qu'ils appellent fonction d'énergie : $E_s(G) = \sum_{i,j=1}^{p} \left| B_{ij}^s - G_{ij} \right|$. Les différentes sources, *s=1,...,S,* sont enfin combinées dans une distribution de Gibbs de la forme $P(G|\beta_1,...,\beta_S) = \frac{e^{-\sum_{s=1}^{S} \beta_s E_s(G)}}{Z(\beta_1,...,\beta_S)}$ où $\beta_1,...,\beta_S$ sont des hyperparamètres. Ils utilisent une procédure d'échantillonnage MCMC (Monte Carlo Markov Chain) pour simuler la distribution des réseaux *a posteriori* et inférer simultanément le réseau et des hyperparamètres. Ceci leur permet aussi d'évaluer l'importance relative des différentes sources de connaissances *a priori*.

Méthodes statistiques : classification basée sur la similarité

Une méthode permettant également l'intégration de données a été proposée par l'équipe d'Yves Moreau (Aerts *et al.*, 2006) et ensuite améliorée grâce à l'utilisation de noyaux par la même équipe (De Bie *et al.*, 2007). Les auteurs s'intéressent à l'identification de gènes impliqués dans des maladies humaines. Leur approche est disponible dans le logiciel Endeavor. La prédiction de liens est basée sur la similarité des gènes candidats à des gènes d'un ensemble d'apprentissage constitué de gènes connus pour participer à la même maladie ou des maladies similaires. Pour chaque gène candidat et chaque source de données un rang est calculé qui reflète la similarité du gène candidat aux gènes de l'ensemble d'apprentissage (voir step 1 Figure 2-15). Les gènes de chaque source de données sont alors agrégés pour obtenir un rang unique par rapport à toute les sources de données utilisées (voir step 3 Figure 2-15), (Aerts *et al.*, 2006).

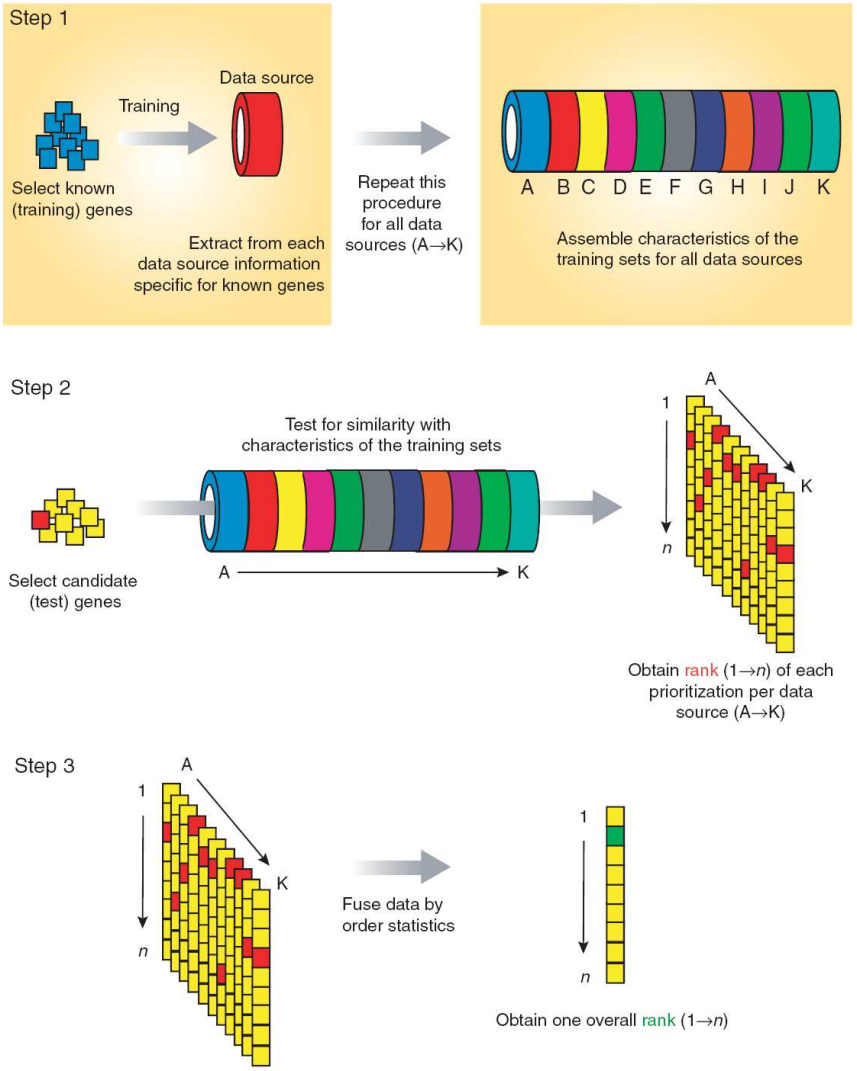


Figure 2-15 : Schéma de l'algorithme Endeavour (Aerts *et al.*, 2006). Sources de données A-K : A) Résumés d'articles scientifiques, B) Annotation fonctionnelle, C) données d'expression de gènes, D) données provenant de la base de données Ensembl, E) dynamique de protéines (Interpro), F) interaction entre protéines (base de données BIND), G) participation à une voie métabolique, H) modèles de régulation en *cis*, I) motifs transcriptionels (TRANSFAC), J) similarité de séquences (BLAST), K) autres sources de données qui peuvent être rajoutées.

2.3.1.3 Méthodes à noyaux

Après avoir introduit les méthodes qui permettent la réalisation d'un apprentissage et l'intégration de données nous allons nous intéresser maintenant aux méthodes à noyaux qui, en plus de ces deux caractéristiques, ont en commun le fait que les données sont représentées sous forme de comparaison entre objets, par des noyaux. Comme les méthodes que nous présenterons à partir de maintenant utilisent toutes ce type de représentation de données, c'est le moment d'introduire les noyaux et l'astuce du noyau (*kernel trick*), éléments principaux des méthodes à noyaux.

Les noyaux

Les noyaux semi-définis positifs ont été utilisés pour la première fois par Aronszajn (Aronszajn, 1950). Mais ce n'est que plus tard que l'astuce du noyau, essentielle dans les méthodes à noyaux utilisées des nos jours, a été introduite. Plus tard, nous verrons qu'il s'agit de dire qu'un noyau semi-défini positif est identique au produit scalaire dans un autre espace (appelé *feature space*), dans lequel l'algorithme peut être réalisé (Vert *et al.*, 2004).

Soit $S = \{x_1,...,x_n\}$, un ensemble de *n* objets à analyser. On suppose que chaque objet x_i est un élément de X (par exemple une collection de gènes). A chaque objet correspondent des descripteurs scalaires, catégoriels ou plus complexes (e.g graphe); soit E l'espace de ces descripteurs. Habituellement, les algorithmes d'analyse de ces ensembles opèrent directement sur ces descripteurs. Dans le cas des méthodes à noyaux, à ces descripteurs vient s'ajouter la construction d'un (ou plusieurs) noyau $k : X \times X \to R$, associant à chaque paire d'objets un scalaire traduisant généralement la similarité entre ces objets. Les méthodes à noyaux pour la classification de ces objets ou d'autres objectifs s'appuient alors sur ces comparaisons de paires d'objets. Les données S sont représentées par une matrice $n \times n$ de comparaison par paires $k_{i,j} = k(x_i, x_j)$ (Vert *et al.*, 2004).

Définition d'un noyau:

Une fonction $k : X \times X \to R$ est appelée un noyau semi-défini positif si elle est symétrique : $k(x,x') = k(x,x')$ pour tous les objets $x, x' \in X$, et définie positive : $\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i x_j) \geq 0$ pour tous $n > 0$, tous les objets $x_1,...,x_n \in X$ et toutes les valeurs réelles $c_1,...,c_n \in R$ (Vert *et al.*, 2004).

Plusieurs types de données, assez différents les uns des autres, peuvent être utilisés pour construire une matrice de ce type.

Exemples :

Le noyau le plus simple est le noyau linéaire. Supposons que les données à analyser sont de type vectoriel $E = R^P$. On définit le noyau linéaire par le produit scalaire pour chaque $x, x' \in R$: $k_L(x,x') = < x,x' > = \sum_{i=1}^{p} x_i x'_i$; $k_L(x,x')$ prend des valeurs d'autant plus grandes que *x* et *x'* se 'ressemblent'. Ce produit entre deux vecteurs est un noyau linéaire, symétrique et sémi-défini

positif. Dans le cas où les données ne sont pas des vecteurs, une manière plus générale existe pour construire des noyaux. D'abord chaque objet $x \in X$ est représenté comme un vecteur $\phi(x) \in R^p$ et ensuite, un noyau est défini pour chaque paire le produit scalaire $x, x' \in X$: $k(x, x') =< \phi(x), \phi(x') >$. Des noyaux encore plus généraux existent et correspondent de manière générale à des produits scalaires sur un espace de Hilbert F (*feature space*) au lieu de R^p. Utiliser un noyau revient à représenter chaque objet $x \in X$ comme un vecteur $\phi(x) \in F$. Néanmoins, le calcul de produits scalaires dans *F* ne necessite pas la connaissance explicite de $\phi(.)$ (Vert *et al.*, 2004). Ceci est à la base de l'astuce du noyau.

L'astuce du noyau :

Tout algorithme basé sur des données vectorielles qui peut être exprimé uniquement par des produits scalaires entre vecteurs, peut être réalisé implicitement dans le *feature space* associé à n'importe quel noyau, en remplaçant tout produit scalaire par la fonction du noyau (Vert *et al.*, 2004).

Par ailleurs, l'efficacité des méthodes à noyaux repose sur le théorème du représentant (*representer theorem*). Soit un noyau *k* définie sur un ensemble fini $S = \{x_1, ..., x_n\} \subset X$ d'objets et $\Psi : R^{n+1} \to R$ une fonction de *n+1* arguments, strictement croissante dans son dernier argument. Alors, chaque solution du problème $\min_{f \in H_k} \Psi(f(x_1), ..., f(x_n), \|f\|_{H_k})$, où H_k est le *feature space* associé à *k*, admet la representation suivante : $f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$.

Maintenant, intéressons nous à certains noyaux couramment utilisés :

- Un **noyau gaussien** est une fonction qui diminue avec la distance euclidienne d entre objets et reflète alors leur similarité:

  $k_G(x, x') = e^{-\frac{d(x,x')^2}{2\sigma^2}}$ où σ est un paramètre.

- Un autre noyau souvent utilisé est le **noyau polynômial** $k_P(x, x') = (x^T x' + c)^d$, où d est le degré du polynôme et c une constante.

- Pour la représentation des données provenant d'un graphe, un **noyau de diffusion** (Kondor & Lafferty, 2002) est géneralement utilisé. Il s'agit d'un noyau basé sur la diffusion de la

chaleur et l'équation de diffusion qui décrit comment la chaleur, les gaz, etc. diffusent avec le temps de manière homogène. Ces idées sont transposées dans un contexte discret, un graphe, en imaginant plusieurs pas très petits pour parcourir le graphe. Ce noyau prend une valeur grande quand la distance entre noeuds d'un graphe est courte : $k_D(x,x') = e^{\beta L}$ où L est la matrice laplacienne du graphe et β est la constante de diffusion. La matrice laplacienne est construite de la manière suivante :

$$L = \begin{cases} 1 & pour\ i \sim j \\ -d_i & pour\ i = j \\ 0 & ailleurs \end{cases}$$

où d_i est le degrée du noeud i et $i \sim j$ signifie que les noeuds sont liés.

Il est important de mentionner ici que de nombreuses opérations sur les noyaux conduisent encore à un noyau. Il en est ainsi de la combinaison linéaire des noyaux: $K = \sum_{p=1}^{P} w_p K_p$, lorsque $w_p > 0$ (Yamanishi *et al.*, 2003).

Méthodes à noyaux : classification svm

L'algorithme Endeavour (Aerts *et al.*, 2006) a été amélioré par la même équipe de recherche (De Bie *et al.*, 2007) en utilisant différents noyaux pour les différentes sources de données et un classifieur svm. Le but, est de trouver un hyperplan qui sépare les données de deux classes. Plus exactement l'hyperplan séparent les gènes (appartenant à une maladie ou une voie métabolique) avec une marge la plus large possible, tout en tenant compte de différentes sources de données utilisées aussi par Aerts et collaborateurs (De Bie *et al.*, 2007). Si les gènes du lot d'apprentissage sont représentés par la matrice *X*, contenant dans la ième ligne le vecteur d'attributs x_i du ième gène utilisé pour l'apprentissage, nous pouvons définir une fonction f du gène x comme $f(x) \triangleq x'w$. Nous cherchons un vecteur de poids w avec $\|w\|^2 \leq 1$ de telle manière que les deux classes soient le plus séparées possible (Figure 2-16). Plusieurs noyaux peuvent être utilisés. Ils sont fusionnés par la moyenne des noyaux: $K = \frac{1}{m}\sum_{j=1}^{m} \frac{K_j}{\beta_j}$ où β_j est une constante de normalisation positive. De plus un poids μ peut être utilisé pour tenir compte de l'importance de chaque noyau. La fonction de décision *f* peut être réécrite en fonction de α, β, μ : $f(x) = \frac{1}{\sqrt{\alpha' K \alpha}} \sum_{i=1}^{n} \alpha_i \left[\sum_{j=1}^{m} \mu_j \frac{k_j(x, x_i)}{\beta_j} \right]$. De Bie

et collaborateurs présentent un algorithme pour choisir les paramètres β et contrôler la quantité de faux positifs (De Bie *et al.*, 2007). Avec cet algorithme les auteurs dépassent les résultats obtenus par la même équipe avec Endeavour (Aerts *et al.*, 2006).

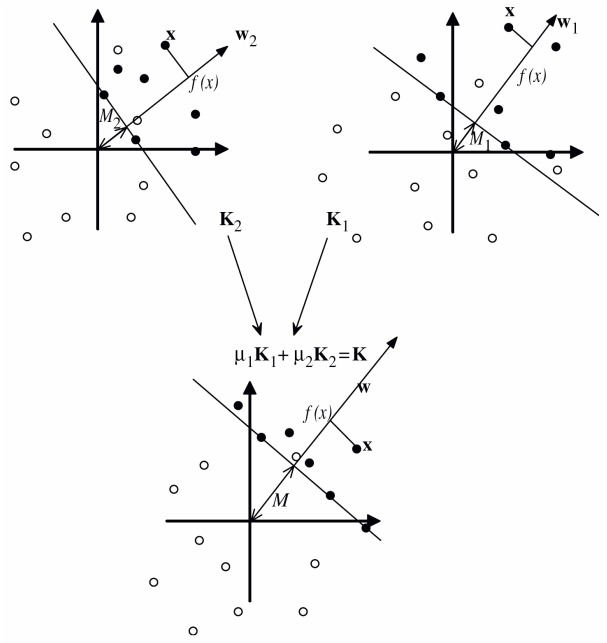


Figure 2-16 : Représentation schématique de l'hyperplan séparant les données positives (cercles noirs) de l'origine et les gènes négatifs (cercles blancs). La combinaison de deux noyaux conduit à une meilleure séparation des deux groupes de gènes.

Méthodes à noyaux : analyse de corrélation canonique

L'analyse de corrélation canonique (KCCA) proposée par Yamanishi et collaborateurs est une méthode supervisée qui s'appuie sur une partie connue du résau en inférant la partie manquante en fusionant différentes sources de données (Yamanishi *et al.*, 2004a; Yamanishi *et al.*, 2004b; Yamanishi *et al.*, 2003). Cette méthode est exposée en détail dans la section 2.3.1.4

Méthodes à noyaux : classification svm basée sur un modèle local

La méthode dite locale pour la reconstruction de réseaux biologiques a été appelée ainsi en opposition aux méthodes globales (SVM, KCCA ...) reposant sur la construction d'un seul classifieur construit à partir de toutes les données et utilisé indifféremment pour la recherche de liens entre n'importe quels gènes . Cette méthode a été proposée par Bleakley et collaborateurs dans le but de mieux reconstruire les réseaux biologiques quand il s'agit d'un graphe avec des arrêtes distribuées de manière non uniforme (Bleakley *et al.*, 2007). Contrairement à la KCCA, où un modèle global est construit pour inférer des nouveaux liens entre tous les noeuds (les gènes), la méthode locale évalue si une arrête existe entre un nouveau noeud et le graphe connu. Cependant, il n'y a pas la possibilité de trouver des nouveaux liens entre deux nouveaux noeuds (i.e. deux

protéines qui ne font pas partie du graphe connu). La méthode commence par un apprentissage individuel d'un sous-graphe associé avec chaque noeud (gène) du graphe connu. Pour chaque gène *v* de l'ensemble d'apprentissage *V* un classifieur svm est appris. Ce classifieur binaire sépare plus ou moins bien les gènes de l'ensemble d'apprentissage directement liés au gène *v.* Une fois la règle de classification pour ce noeud apprise, selon la procédure habituelle de validation croisée sur l'ensemble de paramètres du classifieur et de construction de noyaux, les liens des nouveaux noeuds (nouveaux gènes) associés à ce seul noeud sont prédits. Chaque gène candidat est soumis à chacun de ces classifieurs pour la recherche de liens avec les gènes de *V.* Plus précisément, chaque noeud dans le lot d'apprentissage $v \in V_1$ est considéré indépendamment. Chaque autre noeud du lot d'apprentissage $u \in V_1$ est lié ou pas au noeud *v.* A chaque noeud $u \in V_1$ un label est associé $Y_v \in \{-1,+1\}$, +1 si une arrête existe de *v* à *u* et -1 si l'arrête n'existe pas. Ensuite un algorithme svm est utilisé pour apprendre une fonction f_v capable d'associer le label -1 ou +1 à chaque nouveau noeud faisant partie du lot test V_2. Ce processus est répété pour chaque noeud $v \in V_1$ pour obtenir une prédiction pour tous les nouveaux noeuds (protéines candidates) et le set d'apprentissage (Bleakley *et al.*, 2007). Lors de l'application de cette procédure aux données standards chez la levure, Bleakley et al. (2007) ont obtenu des résultats qui surpasse plusieurs méthodes à noyaux au niveau de l'aire sous la courbe ROC et du FDR. Les auteurs soulignent la qualité des premières détections (20%) avec un FDR très faible. L'approche locale, qui paraît séduisante par ces bons résultats, présente quelques inconvénients qui ont été abordés récemment lors d'un stage à MIA (Dridi, 2008) et différentes solutions ont été proposées. Le premier inconvénient concerne la prédiction des liens entre gènes de l'ensemble test (l'ensemble de candidats). La méthode locale ne peut pas retrouver de liens directs sans modification. Le second inconvénient concerne l'estimation de paramètres de la méthode et des noyaux. Les gènes utilisés pour cette estimation forment un ensemble test additionnel. De fait, la méthode proposée intitalment ne peut prédire des liens contre un ensemble test et une partie de l'ensemble d'apprentissage. Enfin, le troisième inconvénient concerne la construction de classifieurs binaires qui ne prennent en compte que le voisinage du première ordre d'un gène.

Méthodes à noyaux : complétion de noyau

Kato et collaborateurs envisagent la reconstruction de réseau sous l'angle de la complétion d'un noyau de diffusion. Chaque source de données utilisée a un poids différent selon son caractère (Kato *et al.,* 2005). Un noyau de diffusion est construit avec le réseau connu et un autre noyau est

bâti avec les données génomiques. Ce qui est différent dans cette méthode est qu'un algorithme EM (expectation-maximization) est utilisé pour estimer la partie inconnue du noyau. On dispose donc, d'une part, d'un graphe pour l'apprentissage à partir duquel est construit le noyau de diffusion K_I. D'autre part, de n_K noyaux pour les n_K autres sources de données. Les n_K noyaux sont définies pour tous les gènes (apprentissage et candidats) et combinés : $P = \sum_{k=1}^{n_K} b_K P_K$. On cherche alors à compléter le noyau K_I et donc à construire un noyau $Q = \begin{bmatrix} K_I & Q_{vh} \\ Q_{vh}^T & Q_{hh} \end{bmatrix}$ en tenant compte de toutes les données. Pour cela on se ramène dans le cadre de la statistique inférentielle et on associe à Q (resp. à P) une distribution gaussienne $q(x) \sim \mathrm{N}(0, Q)$. Les distributions p et q correspondent aux noyaux P et Q. La moyenne de p est nulle et donc $E[x] = 0$. La covariance est P, $E_P[xx^T] = P$. La moyenne de q est également nulle et la covariance est $E_q[vv^T] = K_I$. Le but est de déterminer la covariance de la deuxième partie h : $Q_{vh} = E_q[vh^T]$ et $Q_{hh} = E_q[hh^T]$. Dans le cas d'une seule source de données *(n_K=1)*, la relation entre v et h est apprise à partir de p en considérant la distribution conditionnelle $p(h|v)$. La distribution jointe de q est estimée comme $\hat{q}(v,h) = p(h|v)q(v)$ et les matrices de covariances sont estimées à partir de $\hat{q}$: $Q_{vh} = K_I P_{vv}^{-1} P_{vh}$, $Q_{hh} = P_{hh} - P_{vh}^T P_{vv}^{-1} P_{vh} + P_{vh}^T P_{vv}^{-1} K_I P_{vv}^{-1} P_{vh}$ où P_{vv}, P_{vh} et P_{hh} sont des sous-matrices de P: $P = \begin{bmatrix} P_{vv} & P_{vh} \\ P_{vh}^T & P_{hh} \end{bmatrix}$. Dans le cas où plusieurs sources de données $P_1, ..., P_{n_K} \in R^{l \times l}$, sont utilisées, la combinaison de ces matrices est utilisée (avec un poids pour chacune). Le noyau combiné avec la somme des poids $b_{n_K} = 1$ est donné par : $P(b) = \sum_{k=1}^{n_K} b_k P_k + \sigma^2 I$ où $\sigma^2 I$ est un terme de régularisation de P. Les poids de chaque noyau est estimé au même temps que les parties inconnues Q_{vh} et Q_{hh} en minimisant la divergence de Kullback-Leiber (KL) : $\min_{Q_{vh}, Q_{hh}, b} D[Q, P(b)]$. Ce problème d'optimisation n'est pas convexe et uniquement des minima locaux peuvent être trouvés. Néanmoins, Kato et collaborateurs proposent de contourner le problème de minima locaux par l'algorithme EM de Dempster et collaborateurs (1977) (Kato *et al.*, 2005). La minimisation se fait en deux étapes: i) *E-step* : fixer b, minimiser $D(Q,P)$ par rapport à Q_{vh} et Q_{hh} et ii) *M-step* : fixer Q, minimiser $D(Q,P)$ par rapport à b. Néanmoins le *M-step* n'est pas convexe et les auteurs proposent d'additionner des variables cachées et de reformuler l'algorithme EM. Le vecteur des variables est étendu à $c = (v^T, h^T, z^T)^T$ où z est

supposée gaussienne, centrée, de covariance *R(b)*. Le problème a résoudre est : $\min_{Q_{vh},Q_{hh},Q_{xz},Q_{zz},b} D\left[\tilde{Q},R(b)\right]$ où : $\tilde{Q}=\begin{bmatrix} Q & Q_{xz} \\ Q_{xz}^T & Q_{zz} \end{bmatrix}$; il possède le même optimum global que le problème initial. Les deux étapes de ce nouveau problème conduisent à des problèmes convexes. L'algorithme a été testé sur les données de la levure utilisées aussi par Yamanishi et collaborateurs (2004b). Les auteurs obtiennent des résultats légèrement meilleurs que les résultats obtenus avec la KCCA.

Méthodes à noyaux : recherche des interactions à partir de la littérature scientifique

Récemment, une méthode à noyau a été utilisée pour extraire des relations à partir de la littérature scientifique ; elle repose sur l'analyse de phrases qui conduit à un graphe et permet la détection d'interactions entre gènes, même dans le cas de structures de syntaxe complexes (Kim *et al.*, 2008). Cette méthode propose un modèle plus complexe que la détection de relations entre gènes simplement par rapport à la co-occurrence dans un article. Basés sur la structure d'une phrase sous forme de graphe les auteurs proposent quatre noyaux capables de représenter les textes comme un graphe : *predicate kernel, walk kernel, dependency kernel, hybrid kernel.* Les meilleurs résultats ont été obtenus avec le *walk kernel.* La méthode a été développée sur des articles concernant la transcription de *B. subtilis.* D'abord, une analyse de la syntaxe de la phrase est réalisée. Elle permet de représenter des dépendances entre gènes par des catégories morphosyntaxiques et des relations de fonction. Ensuite, des noyaux sont construits et doivent représenter, en termes de similarité, la voie la plus courte entre deux gènes sur ce graphe. Si nous prenons l'exemple du *walk kernel,* la phrase montrée en (Figure 2-17) est représentée par l'analyse syntaxique en Figure 2-18. Ensuite le *walk kernel* est construit pour tenir compte de la relation entre 'ywhE' et 'sigF', même si aucune relation de syntaxe existe, parce que le prédicat 'control' fait référence à 'expression'. C'est à dire que 'ywhE', qui n'est pas relié par le prédicat à quand même une relation avec 'sigF'. Dans ce noyau la relation est définie par un parcours dans la représentation graphique de la phrase. Si $P=(V,E)$ est un graphe et V et E sont des noeuds et des arêtes. Un parcours (*walk*) est défini comme des noeuds et arêtes en alternance, commençant par un noeud et terminant par un noeud. Si $v \in V$ et $e \in E$, un parcours est v_i, $e_{i,i+1}$, v_{i+1}, e_{i+2},...,v_{i+n-1}. De la même manière, des parcours commençant et terminant par une arête sont pris en compte (Kim *et al.*, 2008).

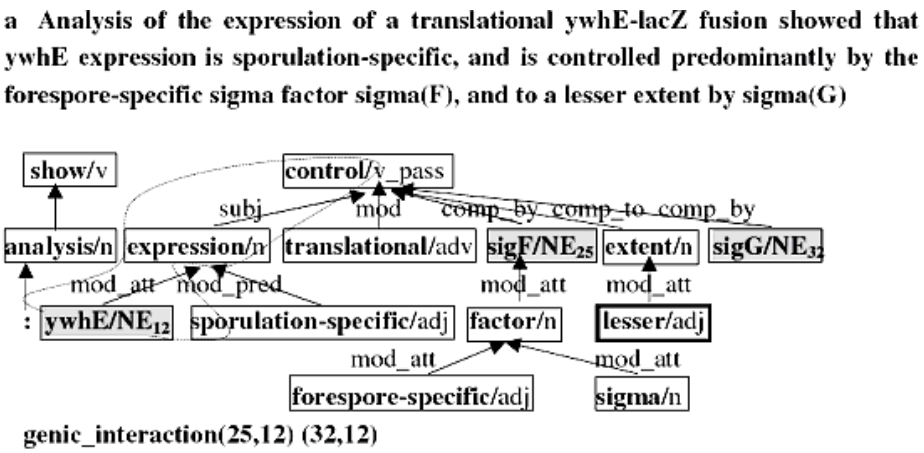


Figure 2-17 : Exemple d'une phrase représentée comme un graphe utilisée pour construire un noyau qui reflète la relation entre gènes (Kim *et al.*, 2008).

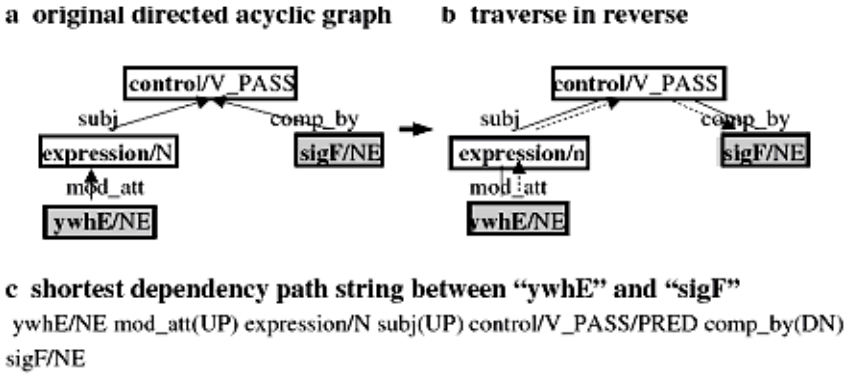


Figure 2-18 : Représentation de l'interaction entre les gènes *ywhE* et *sigF* (Kim *et al.*, 2008).

Les noyaux ont été utilisés pour réaliser un apprentissage svm et les résultats se sont avérés fiables avec une reconnaissance de 70% des paires interagissantes connues (Kim *et al.*, 2008).

2.3.1.4 La KCCA : une méthode à noyaux supervisée qui permet l'intégration de données

La KCCA présente deux caractéristiques avantageuses : c'est une analyse qui permet l'intégration de différents types de données et permet de faire un apprentissage par rapport à un jeu de données standard pour lequel les relations entre objets sont connues. De plus, elle permet l'inférence de liens entre tous les objets utilisés (ceux faisant partie du jeu d'apprentissage et ceux appartenant au lot des candidats).

La KCCA est basée sur l'analyse de corrélation canonique (CCA). Soient deux groupes de variables Y_1 et Y_2 décrivant un même objet *x*. L'analyse de corrélation canonique consiste à trouver des repères qui se correspondent pour représenter l'objet dans chacun de ces repères. Ces derniers sont obtenus en recherchant des combinaisons linéaires des variables (canoniques) de chaque groupe. Ainsi on cherche a_1 et a_2 le plus corrélées possibles : $(a_1, a_2) = \arg\max_{a_1,a_2} \left| corr(a_1^T Y_1, a_2^T y_2) \right|$. Le premier couple d'axes canoniques correspond à la valeur la plus élevée de la corrélation, les autres maximisent aussi cette corrélation avec des contraintes d'orthogonalité par rapport aux axes déjà trouvés.

La KCCA est une CCA généralisée utilisant l'astuce du noyau (Bach & Jordan, 2002). Il s'agit de réaliser une CCA régularisée dans le *feature space* défini par deux noyaux différents sur les mêmes objets (les mêmes gènes). Néanmoins, deux difficultés techniques doivent être surmontées. La première chose à faire c'est représenter les objets par des produits scalaires pour pouvoir se servir de l'astuce du noyau, cela revient à construire un noyau semi-défini positif. La deuxième difficulté est que la CCA n'est pas adaptée à des espaces de dimension très élevée. C'est le cas de la KCCA où le *feature space* est de dimension élevée. Pour cette raison là, une régularisation doit être faite.

Le but de la KCCA est de détecter des corrélations entre deux jeux de données $x_1 = (x_1^{(1)},..., x_1^{(n)})$ et $x_2 = (x_2^{(1)},..., x_2^{(n)})$ où n est le nombre d'objets et chaque jeu de données appartient à X_1 / X_2 pour $i = 1,..., n$. Pour détecter des corrélations entre les deux jeux des données, ils sont projetés dans le *feature space H* par $\phi_1 : X_1 \to H_1$ / $\phi_2 : X_2 \to H_2$. Un analyse de corélation canonique classique est réalisée entre les images $\phi_1(x_1)$ et $\phi_2(x_2)$. Pour chaque paire d'axes $f_1 \in H_1$ et $f_2 \in H_2$ des projections $u_1 = (u_1^{(1)},..., u_1^{(n)}) \in R^n$ et $u_2 = (u_2^{(1)},..., u_2^{(n)}) \in R^n$ de x_1 et x_2 sur f_1 et f_2 peuvent être

définies par $u_1^{(i)} := \left\langle f_1, \phi_1(x_1^{(i)}) \right\rangle_{H_1}$, $u_2^{(i)} := \left\langle f_2, \phi_2(x_2^{(i)}) \right\rangle_{H_2}$ pour $i = 1,...,n$; $\langle \ \rangle_H$ indique le produit scalaire dans l'espace *H*. La moyenne (*m*), la variance (*v*) et la covariance (*c*) de u_1 et u_2 sont définies par :

$$\hat{m}(u_j) := \frac{1}{n}\sum_{i=1}^{n} u_j^{(i)}$$

$$\hat{v}(u_j) := \frac{1}{n}\sum_{i=1}^{n} (u_j^{(i)} - \hat{m}(u_j))^2$$

$$\hat{\text{cov}}(u_1, u_2) := \frac{1}{n}\sum_{i=1}^{n} (u_1^{(i)} - \hat{m}(u_1))(u_2^{(i)} - \hat{m}(u_2))$$

Un CCA kernelisée conduirai à maximiser par rapport à $f_1 \in H_1$ et $f_2 \in H_2$ la corrélation entre u_1 et u_2 :

$$\hat{corr}(u_1, u_2) := \frac{\hat{\text{cov}}(u_1, u_2)}{(\hat{\text{var}}(u_1)\,\hat{\text{var}}(u_2))^{1/2}}$$

Les espaces H_1 et H_2 peuvent être très riches et la maximisation précedente conduirai à des aves trop adaptées aux données (sur-ajustement). Pour cette raison, la régularisation de f_1 et f_2 est importante. Comme la régularité de f_1 et f_2 est liée à la norme dans H_1 , H_2 Yamanishi propose de maximiser la quantité suivante (Yamanishi *et al.*, 2004a) :

$$\gamma(f_1, f_2) := \frac{\hat{\text{cov}}(u_1, u_2)}{(\hat{\text{var}}(u_1) + \lambda_1 \|f_1\|^2)^{1/2} (\hat{\text{var}}(u_2) + \lambda_2 \|f_2\|^2)^{1/2}}$$

où λ_1 et λ_2 sont des paramètres de régularisation, qui permettent de contrôler la régularité de f_1 et f_2.

Le lagrangien de ce problème est :

$$L(f_1, f_2, \rho_1, \rho_2) = \hat{\text{cov}}(u_1, u_2) + \frac{\rho_1}{2}(1 - \hat{\text{var}}(u_1) - \lambda_1 \|f_1\|^2) + \frac{\rho_2}{2}(1 - \hat{\text{var}}(u_2) - \lambda_2 \|f_2\|^2)$$

où ρ_1 et ρ_2 sont des multiplicateurs de Lagrange. Par l'annulation de la dérivée de L par rapport à f_1 et f_2 à l'optimum, nous déduisons : $f_1 = \sum_{j=1}^{n} \alpha_1^{(i)} \phi_1(x_1^{(j)})$, $f_2 = \sum_{j=1}^{n} \alpha_2^{(i)} \phi_2(x_2^{(j)})$ pour α_1 et $\alpha_2 \in R^n$.

Supposons maintenant les points centrés dans le *feature space* et donc les moyennes de u_1 et u_2 nulles. La variance et la covariance de u_1 et u_2 s'expriment alors simplement en fonction de α_1 et α_2 :

$$\hat{\text{var}}(u_1) = \frac{1}{n}\alpha_1^T K_1^2 \alpha_1 \qquad \hat{\text{var}}(u_2) = \frac{1}{n}\alpha_2^T K_2^2 \alpha_2 \qquad \hat{\text{cov}}(u_1, u_2) = \frac{1}{n}\alpha_1^T K_1 K_2 \alpha_2$$

où K_1 et K_2 sont les noyaux $n \times n$: $K_1(i,j) = k_1(x_1^{(i)}, x_1^{(j)})$ et $K_2(i,j) = k_2(x_2^{(i)}, x_2^{(j)})$. Les normes de f_1 et f_2 s'écrire également en fonction de α_1 et α_2 : $\|f_1\|^2 = \alpha_1^T K_1 \alpha_1$, $\|f_2\|^2 = \alpha_2^T K_2 \alpha_2$. En réécrivant le Langrangien aussi en fonction de α_1 et α_2 on obtient :

$$L(\alpha_1, \alpha_2, \rho_1, \rho_2) = \frac{1}{n}\alpha_1^T K_1 K_2 \alpha_2 + \frac{\rho_1}{2n}(n - \alpha_1^T K_1^2 \alpha_1 - n\lambda_1 \alpha_1^T K_1 \alpha_1^{\ 2}) + \frac{\rho_2}{2n}(n - \alpha_2^T K_2^2 \alpha_2 - n\lambda_2 \alpha_2^T K_2 \alpha_2)$$

nous obtenons que les valeurs (α_1, α_2) et (ρ_1, ρ_2) qui donnent la solution du Lagrangien et sont la solution du problème des valeurs propres généralisés suivants :

$$\begin{pmatrix} 0 & K_1 K_2 \\ K_2 K_1 & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \rho \begin{pmatrix} (K_1 + (n\lambda_1/2)I)^2 & 0 \\ 0 & (K_2 + (n\lambda_2/2)I)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$ (Yamanishi *et al.*, 2004a).

Les corrélations canoniques sont en réalité les valeurs propres les plus larges de ce problème (Bach & Jordan, 2002). Les variables canoniques u_1 et u_2 peuvent être obtenues de la manière suivante : $u_1 = K_1\alpha_1$, $u_2 = K_2\alpha_2$. Ceci est valable si les points sont centrés dans le *feature space.* Ce centrage peut être obtenu par $K_{centré} = N_0 K N_0$ où $N_0 = (I - \frac{1}{n}11^T)$ avec $1 = (1,...,1)^T \in R^n$. En pratique, on retiendra alors pour réprésenter les données un certain nombre de composants canoniques associées aux plus grands valeurs propres.

Les noyaux utilisés (K_1 ou K_2) peuvent être des noyaux intégrés (Yamanishi *et al.*, 2003). Ainsi par exemple, le noyau K_2 peut être obtenu par l'addition de plusieurs noyaux $K_2 = K_A + K_B + K_C$ où les trois noyaux doivent être de la même taille mais peuvent avoir un poids différent. Le noyau K_1 est vu comme le noyau standard qui est utilisé comme apprentissage. L'apprentissage est fait lorsque les repères sont calculés sous la contrainte de corrélation entre les deux noyaux. Les objets (les gènes) sont représentés par u_1 et u_2. Finalement, la distance entre deux gènes i et j sera la

distance euclidienne entre $u_2(i)$ et $u_2(j)$.

La KCCA a été utilisée pour retrouver des liens connus entre protéines de *E. coli* (Yamanishi *et al.*, 2004a; Yamanishi *et al.*, 2004b), et de la levure (Yamanishi *et al.*, 2003). Dans ces travaux il a été démontré que c'est l'intégration de données qui permet d'obtenir les meilleurs résultats. Ainsi, Yamanishi et collaborateurs ont utilisé quatre types de données disponibles chez la levure : des données transcriptomiques, des données d'interactions entre protéines, des données de localisation et des profils phylogénétiques (Yamanishi *et al.*, 2003). Les données de référence correspondent au graphe des voies métaboliques, disponible pour une partie des gènes. Dans ces études le noyau standard (bâti sur les voies métaboliques) est un noyau de diffusion. Cette méthode permet de trouver des distances entre tous les gènes participant à l'analyse, c'est à dire entre gènes faisant partie des données standard (le graphe des voies métaboliques) et entre gènes ne faisant pas partie de ce graphe (gènes candidats), ainsi qu'entre deux gènes candidats.

En résumé, la KCCA est une méthode simple à mettre en œuvre, qui permet l'intégration de différents types de données, permet de faire un apprentissage par rapport aux relations connues et sert à retrouver des liens entre tous les gènes pour lesquels des données sont disponibles.

2.3.2 Analyse de sensibilité

Après avoir décrit plusieurs méthodes de prédiction intégrant des données nombreuses et variées, on peut se poser la question de l'importance de chacune de ces différentes données dans la prédiction, et plus généralement de la sensibilité des prédictions en regard des données d'une part, des différents paramètres de la méthode KCCA, d'autre part. Classiquement, les études d'analyse de sensibilité sont souvent utilisées pour simplifier des modèles complexes et éventuellement émuler ces modèles par des modèles plus simples et plus faciles à mettre en œuvre. Dans cette section seront presentés les principaux ingrédients de l'analyse de sensibilité, ingrédients que nous avons combinés pour répondre à des questions plus précises dans notre contexte.

L'analyse de sensibilité (AS) étudie la façon dont des perturbations sur les entrées d'un modèle engendrent des perturbations sur sa réponse. L'analyse de sensibilité globale s'intéresse à la variabilité de la sortie du modèle sur le domaine de variation des entrées. Elle étudie la répercussion de la variabilité des entrées sur celle de la sortie, en déterminant la proportion de la variance de la sortie attribuée à chacune des entrées ou des ensembles d'entrées (Saltelli *et al.*, 2000). Cette

analyse peut être effectuée par simulation, ce qui permet non seulement d'étudier les effets des entrées, mais aussi les effets des interactions entre entrées sur la sortie. L'interprétation des résultats des simulations est liée à un plan d'expériences adéquat pour les simulations.

Aussi bien les entrées que les sorties des modèles complexes peuvent revêtir plusieurs formes (variables réelles, qualitatives, images, graphes, fonctions multivariées, etc.). Nous considérons seulement un modèle opérant sur un vecteur de variables et produisant une sortie ou variable réponse scalaire. Parmi les méthodes d'AS existantes, la plus simple estime la part de la variance de la réponse due à la variance de chaque variable d'entrée (Saltelli *et al.*, 2000). Pour que cette étude de variance soit efficace et répondre aux objectifs de l'analyse de sensibilité, il est très important de faire un bon échantillonnage. En général, des plans d'expériences sont donc construits pour indiquer les valeurs des variables d'entrée à choisir pour étudier les perturbations sur la sortie.

2.3.2.1 Plan d'expériences

La manière la plus complète consiste bien sûr à tester toutes les combinaisons possibles des valeurs que chaque variable d'entrée peut prendre. Cependant, l'espace de données est très grand (2^p simulations pour une collection de p facteurs avec 2 niveaux chacun). Il est clair que plus le nombre de variables augmente plus il est difficile de réaliser toutes ces simulations. Le facteur fait réference à une variabe d'entrée ou à un paramètre du système étudié.

Lorsque les facteurs sont nombreux ou que la simulation (le calcul qui conduit à une réponse ou sortie obtenu pour un groupe de variables d'entrée données) est coûteuse en temps et/ou en argent, on est contraint de limiter le nombre d'expériences. Plusieurs possibilités s'ofrent à nous pour l'échantillonage des entrées : purement aléatoire, perturbation tour à tour d'un seul facteur. L'inconvénient de ce dernier type d'échantillonnage est que les interactions ne sont pas prises en compte. La perturbation d'un facteur peut ne pas changer le résultat parce qu'il n'est pas très important ou parce qu'un autre facteur compense l'information. Cependant, si deux facteurs sont perturbés dans une même simulation, le résultat pourrait changer s'il n'y a pas de compensation. Il est alors nécessaire de construire un plan d'expériences. Ce plan doit fournir un ensemble raisonnable de simulations informatives. Pour cela et pour l'interprétation des résultats le plan doit être adossé à un modèle paramètrique dont les coefficients seront interprétables et qui en outre permettra, éventuellement, d'émuler le système étudié. Les plans factoriels fractionnaires peuvent répondre à ces impératifs (Box *et al.*, 1978).

Il faut alors choisir une matrice d'échantillonnage *X* qui permet d'estimer les facteurs β. Si *V* est la variance de β, cette variance sera minimale quand la moitié des simulations est conduite avec $x = x_{\max}$ et l'autre moitié avec $x = x_{\min}$. Une matrice d'échantillonnage à plusieurs facteurs *p* est alors donnée par (si la variable peut prendre deux niveaux):

$$X = \begin{pmatrix} 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

La première colonne ne consitue pas une expérience mais représente la moyenne des facteurs. Les colonnes de cette matrice sont orthogonales. Dans une situation avec $p > 2$ facteurs et pour chaque facteur x_k, le nombre de simulations pour lesquelles $x_k = 1$ doit être égal au nombre de simulations pour lesquelles $x_k = -1$ pour avoir l'orthogonalité avec la colonne de 1 qui correspond au facteur β_0. Un plan factoriel complet consiste alors à effectuer 2^p expériences.

Les interactions peuvent être testées par des termes additionnels obtenus par la multiplication des facteurs dont les interactions vont être évaluées. Ainsi, pour un plan d'expériences de 3 facteurs (A), les interactions entre les facteurs $x_1 x_2$ sont testés par la dernière colonne en (B), qui est une multiplication des colonnes 1 et 2 de (A). (B) est la matrice utilisée pour estimer le modèle linéaire avec un treme d'interaction (x_1, x_2) qui n'est pas confondu avec x_1, x_2, x_3. Les colonnes de (B) sont encore orthogonales.

(A)

$$X = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

(B)

$$X = \begin{pmatrix} 1 & -1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Si le nombre de facteurs est élevé, des plans fractionnaires avec $n = 2^{p-k}$ simulations peuvent être construits (Box *et al.*, 1978). Pour le cas de quatre facteurs, par exemple, on peut utiliser un plan d'expériences complet sur 3 facteurs et échantillonner le quatrième facteur en rajoutant une colonne multiple de deux premières colonnes. On ne pourra pas distinguer l'effet d'interaction des deux

premiers facteurs et du quatrième facteur. Il s'agit d'un plan de résolution III. Pour pouvoir estimer les facteurs principaux sans confusion avec les interactions un plan de résolution IV est nécessaire. Pour l'obtenir il faut choisir la colonne à ajouter comme un terme d'interaction d'ordre 3. Dans ce plan uniquement les interactions d'ordre 2 seront confondues entre elles.

2.3.2.2 Analyse de variance sur la sortie

Si la matrice d'échantillonnage *X* de taille $n \times p$ est correctement choisie (où *n* est le nombre de simulations), par exemple avec un plan d'expérience fractionnaire, il sera possible d'estimer des effets principaux et les interactions des *p* facteurs. Cette matrice est choisie de manière à obtenir une estimation des coefficients la plus précise possible.

Appelons $x_1,...,x_p$ les *p* facteurs correspondants aux variables d'entrée et Y la réponse ou variable de sortie. L'estimation des ces effets est réalisée à l'aide d'un modèle linéaire : $Y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \sum_{i=1}^{p} \sum_{j=1}^{p} \beta_{i,j} x_i x_j + \in$ où $\in$ est l'erreur. Lorsque les facteurs prennent des valeurs réelles dans un intervalle $[x_{\min}, x_{\max}]$, l'échantillonnage qui convient minimise la variance des estimateurs $\hat{\beta}_i$ des facteurs $\beta_i \quad i = 1,..., p$ et consite à échantillonner les extêmités de ces intervalles.

L'estimateur des moindres carrés correspondant est obtenu par $\hat{\beta} = (X^T X)^{-1} X^T Y$. X est complétée par des colonnes correspondant aux termes d'interaction. Dans le cas d'une matrice d'échantillonnage de résolution IV, X est orthogonale et $\hat{\beta} = D^{-1} X^T Y$ où D est diagonale. Les effets principaux ne sont pas confondus avec les interactions d'ordre 2. Par contre, il y a confusion entre certains effets d'interactions à l'ordre 2.

2.4 Thèmes concernant les rôles prédits pour les peptidases d'intérêt

2.4.1 Les protéines exportées

La cellule a développé des systèmes dans le but de différencier les protéines cytoplasmiques de celles devant être guidées à l'extérieur ou vers un autre compartiment cellulaire. Une protéine est exportée si sa destination finale est extra-cytoplasmique (membranaire ou associée à l'enveloppe bactérienne). Elle est sécrétée si elle est libérée dans le milieu extérieur. La plupart des protéines destinées à être exportées ou sécrétées sont synthétisées sous forme de précurseurs. Le précurseur est composé de la partie mature de la protéine, à laquelle est associée en N-terminal une séquence signature, le peptide signal (PS). Ce dernier est reconnu par la machinerie cellulaire et pris en charge par un complexe protéique membranaire, le translocon, qui assure sa translocation à travers la membrane. Quatre localisations extracellulaires de protéines nécessitent la présence d'un PS : les protéines sécrétées dans le milieu extérieur, les lipoprotéines, les protéines ancrées à la membrane cellulaire et enfin les protéines sécrétées associées au peptidoglycane par des motifs particuliers. Ici ne seront abordés que les trois premiers types : les protéines sécrétées (libérées dans le milieu) et deux classes de protéines, exportées à l'extérieur de la cellule par la voie Sec, qui ne sont pas libérées dans le milieu: les lipoprotéines et les protéines ancrées à la paroi via un motif LPXTG.

La machinerie Sec est une des 16 machineries de sécrétion qui existent chez les bactéries. Elle est la seule à être présente chez toutes les bactéries et à être essentielle pour la viabilité (Papanikou *et al.*, 2007). Néanmoins, d'autres systèmes dédiés à l'export des protéines existent. Ils sont souvent appelés systèmes de sécrétion indépendants de la voie Sec. Chez *B. subtilis*, ces systèmes sont de moindre importance que la voie Sec mais il existe au moins trois autres voies alternatives : i) la voie *twin arginine translocation* (TAT), ii) une voie dédiée à la sécrétion des prépilines iii) des ABC transporteurs (Tjalsma *et al.*, 2000). De plus il y a des protéines qui, en absence de peptides signaux, se trouvent néanmoins à l'extérieur du cytoplasme. Comme il s'agit le plus souvent de protéines qui possèdent une fonction connue à l'intérieur de la cellule on les appelle des *moonlighting proteins*, parce qu'elles possèdent une deuxième fonction[1] (Bendtsen *et al.*, 2005).

En résumé, l'export peut être divisé en quatre grandes étapes :

i) **La synthèse** : La synthèse des protéines qui possèdent un peptide signal est faite de la même manière que la synthèse d'autres protéines. Elle peut être réalisée en même temps que la translocation à travers la membrane dans le cas de la sécrétion co-traductionnelle.

ii) **La reconnaissance** : pendant l'étape initiale, le précurseur est reconnu par des systèmes spécialisés, pris en charge et guidé vers la membrane cytoplasmique.

ii) **La translocation** : cette étape correspond au passage actif du précurseur au travers de la membrane cytoplasmique. C'est un processus actif qui fait intervenir un grand nombre de protéines, SecYEG constituant le canal de translocation et SecDF, YajC, YidC facilitant la translocation du précurseur.

[1] To moonlight : avoir un deuxième travail, en plus du travail à plein temps, souvent la nuit.

iii) **Le clivage du peptide signal** : pendant l'étape finale de l'exportation, le PS du précurseur est clivé puis hydrolysé par des protéases spécifiques. Chez les bactéries Gram positives, les protéines sécrétées sont libérées dans le milieu, chez les bactéries Gram négatives elles sont libérées dans le périplasme puis prises en charge pour traverser la membrane externe.

iv) **L'hydrolyse du peptide signal** : cette étape est réalisée par des peptidases spécialisées, dites signal peptide peptidases (SPPases).

Maintenant nous aborderons une à une les étapes de l'export de protéines, en mettant l'accent sur les points essentiels pour la discussion des résultats concernant les rôle des peptidases d'intérêt de ce travail.

2.4.1.1 La synthèse des protéines exportées

La synthèse de toutes les protéines consiste en plusieurs étapes commençant par l'activation des acides aminés et se terminant par les modifications post-traductionnelles des protéines (Nelson & Cox, 2000), voir Figure 2-19.

I)	Activation	Cette étape se caractérise par l'activation de la partie C-terminale des acides aminés pour faciliter la formation de la liaison peptidique et la création d'un lien entre un acide aminé et l'ARN messager (ARNm). L'ARNm est reconnu grâce à la séquence Shine Delgrano qui se trouve juste avant le codon start.
II)	Initiation	Le premier ARNt est attaché à la sous unité ribosomale 30S. Une fois ce complexe formé, la sous unité plus grande 50S s'attache. Il s'agit d'une étape nécessitant de l'énergie fournie par le GTP et régulée par les facteurs d'initiation.
III)	Elongation	Le polypeptide est rallongé par l'ajout des acides aminés correspondant aux codons du mARN. Dans cette étape les facteurs d'élongation interviennent et une molécule de GTP est utilisée par ajout de chaque acide aminé.
IV)	Terminaison et libération	Le codon stop provoque l'arrêt de la traduction et le polypeptide est libéré par les facteurs de libération (*release factors*). Les *release factors* i) hydrolysent la dernière liaison peptidyl-ARNt, ii) libèrent le dernier ARNt et finalement, iii) les deux sous unités du ribosome se séparent.
V)	Repliement et modifications post-traductionnelles	Selon la protéine, il est possible que certains acides aminés soient enlevés et/ou des groupes acétyle, phosphoryle, méthyle, carboxyle soient rajoutés.

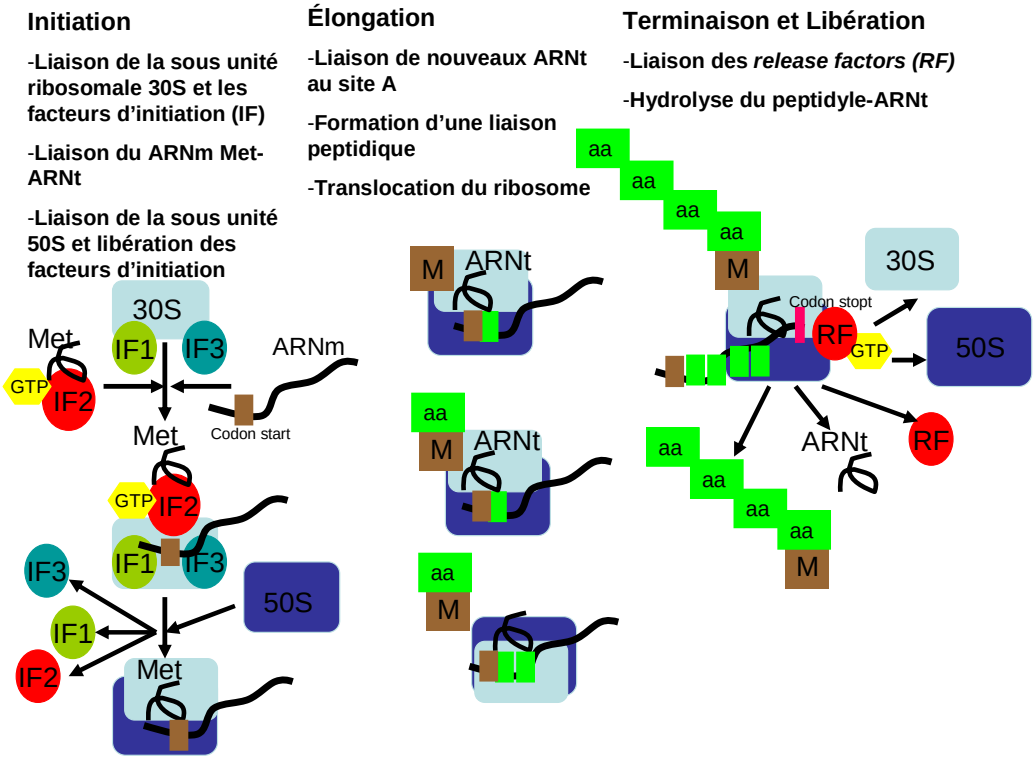


Figure 2-19 : Schéma des étapes II à IV de la synthèse des protéines.

2.4.1.2 La reconnaissance des protéines exportées

Dans le cas d'une translocation co-traductionnelle, toujours pendant la synthèse, plus exactement pendant l'étape d'élongation, le peptide signal de la pré-protéine est reconnu par la particule de reconnaissance du peptide signal (SRP) composé de l'ARN 4,5 S et de la protéine Ffh, une GTPase. Ffh est un protéine bifonctionnelle qui possède deux domaines, le domaine G (dans la partie N-terminale) qui lie le GTP et un domaine de alpha-hélice (dans la partie C-terminale) qui lie la sous-unité ARN et le peptide signal du polypeptide (Gutierrez *et al.*, 1999). La particule SRP existe chez tous les organismes entièrement séquencés (bactéries, archaeas, eucaryotes). Le peptide signal de la pré-protéine, ainsi que le ribosome, attaché à SRP, capturent une molécule de GTP et l'élongation du polypeptide est arrêtée à une longueur d'environ 70 acides aminés. Le complexe ribosome-polypeptide-SRP-GTP est dirigé à la membrane cellulaire où se poursuit l'élongation. Le polypeptide est délivré, pendant sa synthèse, à un complexe de translocation. Le récepteur de SRP dans la membrane est FtsY, cette protéine transfère le polypeptide en cours de synthèse à la machinerie de translocation. Une fois la pré-protéine prise en charge par la machinerie de translocation, SRP est libéré et la traduction du peptide continue passant directement dans la membrane (Gutierrez *et al.*, 1999), Figure 2-20.

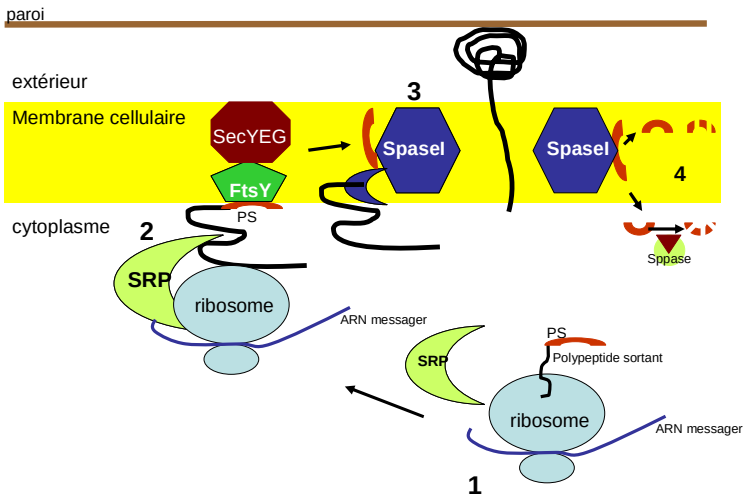


Figure 2-20 : Synthèse et sécrétion des protéines reconnues par SRP et possédant un peptide signal clivé par la signal peptidase I (SPase I). 1) SRP reconnaît le polypeptide. 2) Le complexe ribosome-polypeptide est conduit à la membrane cellulaire vers le récepteur de SRP : FtsY. 3) La protéine synthétisée est prise en charge par la machinerie de sécrétion. Le peptide signal est clivé. 4) La signal peptidase I hydrolyse le peptide signal en séparant la partie hydrophobe de la partie hydrophile et le peptide signal est dégradé.

Dans le cas où une protéine n'est pas reconnue par SRP (les protéines cytoplasmiques ou les protéines transloquées de manière post-traductionnelle), c'est la chaperonne de repliement (*trigger factor*) qui lie les polypeptides sortant du ribosome. Apparemment, c'est l'hydrophobicité globale du polypeptide sortant qui est à l'origine de cette différence de prise en charge par SRP ou par le trigger factor. Dès que le peptide signal est visible, SRP se lie au peptide. Une fois que les sites de reconnaissance pour le *trigger factor* sont apparus il y a une concurrence entre SRP et *trigger factor* (Eisner *et al.*, 2006; Scott & Barnett, 2006).

Dans les cas où la translocation est réalisée après la fin de la synthèse de la protéine, les protéines sécrétées ne semblent pas être reconnues par un système spécifique chez les bactéries Gram positives quand elles émergent du ribosome. Chez les bactéries Gram négatives elles sont conduites après la traduction au translocon SecA par SecB. SecB reconnaît et lie les protéines qui ont un peptide signal d'au moins 20 acides aminés et les transfère au translocon évitant aussi le repliement de la protéine avant sa sécrétion. Les bactéries Gram-positives ne possèdent pas de SecB. Chez *B. subtilis,* CsaA pourrait accomplir le rôle de SecB (Tjalsma *et al.*, 2000).

La voie co-traductionnelle, qui fait intervenir SRP, est la mieux caractérisée chez les bactéries Gram-positives. Néanmoins, certaines données font penser que des voies alternatives, post-traductionnelles existent aussi chez ces bactéries. Ainsi, la voie SRP n'est pas essentielle chez les streptocoques en condition de croissance sans stress. Dans le cas d'un double mutant SRP-YidC2, la croissance diminue significativement, ce qui est en faveur d'une voie alternative à SRP dans laquelle interviendrait YidC2 (Scott & Barnett, 2006). Il est connu chez *E. coli* que YidC participe à l'insertion de protéines intégralement membranaires dans la membrane, telles que les protéines nécessaires à la respiration en aérobiose ou en anaérobiose (de Gier & Luirink, 2001; Price & Driessen, 2008; Samuelson *et al.*, 2000). YidC ne serait alors pas responsable de la sécrétion de protéines en soi, mais plutôt de l'insertion de protéines dans la membrane. Il peut s'agir de protéines transloquées par la voie Sec mais aussi de protéines indépendantes de cette voie (Samuelson *et al.*, 2000). Il existe des homologues de YidC chez *L. lactis* SK11 et de MG1363 (YidC). L'homologue du gène codant YidC est *ybdC* dans la souche IL1403. Néanmoins, son rôle chez le lactocoque n'a pas été étudié et aucune voie de translocation post-traductionelle n'a été identifiée chez *L.lactis* à ce jour.

2.4.1.3 La translocation de protéines

Parmi les bactéries, la machinerie Sec la mieux caractérisée est celle de *E. coli* suivie de celle de *B. subtilis.* Chez les bactéries lactiques, le séquençage du génome de *L. lactis* IL1403 (Bolotin *et al.*, 2001) a révélé la présence de presque tous les gènes *sec*, mais révèle également quelques particularités (absence de SecB, SecDF), Tableau 2-4. Par analogie, plusieurs fonctions peuvent être transposées aux bactéries lactiques mais les mécanismes détaillés restent inconnus.

Tableau 2-4 : Protéines faisant partie de la machinérie de translocation

Protéine	***E. coli***	***B. subtilis***	***L. lactis***
SecA	oui	oui	oui
SecB	oui	-	-
SecY	oui	oui	oui
SecE	oui	oui	oui
SecG	oui	oui	oui
SecD	oui	oui	-
SecF	oui	oui	-
YajC	oui	oui	oui (YwaB)
YidC	oui	oui	oui (YbdC)
CsaA	-	oui	-

Chez les bactéries Gram positives, il a longtemps été considéré que l'exportation se faisait uniquement selon un mode co-traductionnel. La découverte de CsaA permet de postuler la présence d'autres voies pour la sécrétion de protéines. Le mécanisme de prise en charge des précurseurs naissants par le système SRP est très similaire à celui décrit chez *E. coli*, notamment pour la prise en charge des protéines membranaires (Bunai *et al.*, 1999). SRP fonctionne alors comme un chaperon conduisant directement le précurseur au translocon, par association directe entre Ffh et SecA, alors que pour les protéines membranaires, le SRP délivre les précurseurs via FtsY (Bunai *et al.*, 1999). Chez *L. lactis*, aucune donnée expérimentale n'est jusqu'à présent disponible. Cependant, l'analyse de son génome a permis de détecter des homologues de l'ARN 4,5S, ARN et de Ffh (54 % et 48 % d'identité avec Ffh de *B. subtilis* et *E. coli*, respectivement). Il y a également un homologue de FtsY (58 % et 38 % d'identité avec FtsY de *B. subtilis* et *E. coli*, respectivement). En l'absence d'un homologue de séquence de SecB et de CsaA, le mécanisme qui permet au précurseur d'emprunter spécifiquement la voie SRP n'est pas établi en détail. Plusieurs travaux suggèrent que le chevauchement de substrats sécrétés par SRP et d'autres voies est très réduit. Apparemment, SRP serait chargé de la translocation de protéines codées par des gènes régulés et d'autres voies existeraient pour les protéines des gènes constitutifs (Gutierrez *et al.*, 1999). Par ailleurs, Zanen et collaborateurs ont étudié l'importance de l'hydrophobicité du peptide signal par le remplacement d'acides aminés hydrophobes (Leu) ou moins hydrophobes (Ala) dans le peptide signal de l'α-amylase (Zanen *et al.*, 2005). En effet, les précurseurs des protéines possédant un peptide signal hydrophobe sont de préférence pris en charge par SRP et non pas par d'autres voies chez *B. subtilis.* Ces résultats, obtenus chez *B. subtilis*, confirment les observations faites chez *E. coli* par Lee et collaborateurs, qui avaient déjà observé que les modifications des peptides signaux qui augmentent leur hydrophibicité conduisent les précurseurs à emprunter la voie SRP au lieu de

SecB (Lee & Bernstein, 2001).

En plus de l'hydrophobicité il a été aussi montré que le peptide signal peut diriger les protéines exportées à un endroit précis de la membrane en vue de leur export (Carlsson *et al.*, 2007). Les auteurs ont étudié les protéines M et F chez *S. pyogenes.* Il s'agit de protéines ancrées dans la paroi. La sécrétion de la protéine M se fait à proximité du septum. Au contraire, la protéine F est localisée dans les vieux pôles de la cellule, opposés aux septa. Carlsson et collaborateurs ont étudié l'effet de l'échange des peptides signaux de ces deux protéines sur leur localisation. Les résultats montrent que la protéine F possédant le peptide signal de la protéine M est sécrétée proche du septum et que la protéine M possédant le peptide signal de la protéine F est préférentiellement sécrétée dans les vieux pôles de la cellule, démontrant que c'est le peptide signal qui dirige les protéines au bon endroit (Carlsson *et al.*, 2007).

2.4.1.4 Le peptide signal et son clivage

Le peptide signal (PS) est la séquence reconnue par la machinerie cellulaire pour diriger l'exportation et la sécrétion d'une protéine. De nombreuses études ont permis de déterminer les caractéristiques principales des PS. Globalement, les PS des eucaryotes et des bactéries à Gram-négatives ont une longueur moyenne de 20 acides aminés, alors que ceux des bactéries Gram-positives ont une longueur moyenne de 28 acides aminés (Tjalsma *et al.*, 2000). Chez *L. lactis*, 146 des 2310 protéines prédites possèdent un PS potentiellement clivé par la SPase I, ces peptides signaux ont une longueur moyenne de 33 acides aminés (Buist *et al.*, 2006).

Tous les PS sont constitués de trois domaines structuraux caractéristiques : un domaine N-terminal chargé positivement (N), un domaine central H hydrophobe (H) et un domaine C-terminal (C), qui contient le site de clivage du PS, constitué de petits résidus polaires (von Heijne, 1989). Pour que sa prise en charge soit efficace, il faut que la charge de la région N soit positive par rapport à la charge de la région entourant le site de clivage (von Heijne, 1989). Une seule région est conservée dans les PS de tous les organismes : celle qui entoure le site de clivage par la signal peptidase (SPase). Deux acides aminés sont conservés aux positions -1 et -3 relatives au site de clivage du PS. Le motif le plus couramment observé est Ala-X-Ala. L'acide aminé en position -2 est généralement un résidu de petite taille. De façon plus générale, les précurseurs contenant soit Ala, Gly, Ser, Cys ou Pro à la position -1 et Ala, Gly, Ser, Cys, Thr, Val, Ile, Leu ou Pro à la position -3, sont efficacement clivés par la SPase. Presque tous les résidus sont tolérés aux positions -2, -4 et -5 (von Heijne, 1989). Chez les bactéries Gram positives, la coupure du peptide signal se trouve à 7-9 acides aminés de la région C-terminal de la fin de la région H. Au contraire, chez *E. coli*, la coupure se trouve à 3-7 acides aminés de la même position (von Heijne & Abrahamsen, 1989).

Une fois que la protéine a traversé la membrane, la dernière étape de l'exportation se produit. Dans cette étape les séquences signal sont enlevées par des peptidases signal pendant le processus de translocation (Novak *et al.*, 1986). Il s'agit d'enzymes chargées de cliver le peptide signal des protéines qui sont exportées quand leur extrémité C-terminale arrive à l'extérieur de la membrane. Ce clivage est une condition pour la libération de la protéine mature dans le milieu extérieur. Chez *Bacillus subtilis* plusieurs SPases sont présentes. Au contraire, chez la plupart de eubactéries et archaebactéries, ainsi que chez la levure, une seule SPase semble être suffisante (Tjalsma *et al.*, 2000). Comme pour *E. coli* (Dalbey & Wickner, 1992) la SPase est une protéine avec une activité essentielle chez *Bacillus subtilis* (Tjalsma *et al.*, 2000). Chez les bactéries, le site actif des SPases est prédit comme étant localisé sur la partie extérieure de la membrane cytoplasmique (Tjalsma *et al.*, 2000). Les peptidases signal reconnaissent apparemment la structure des peptides signaux

mentionnée plus haut (section 2.4.1.4) et non pas une séquence spécifique (Dev & Ray, 1990). Les signal peptidases de type II (SPases II) sont spécifiquement chargées de cliver les peptides signaux de lipoprotéines. De la même manière que les signal peptidases de type I (SPases I), leur site actif est prédit comme étant à l'extérieur de la membrane (Tjalsma *et al.*, 2000).

2.4.1.5 Hydrolyse du peptide signal

Une fois le peptide signal libéré de la protéine, il est dégradé très rapidement par les SPPases (Dev & Ray, 1990). Chez *E. coli*, les peptides signaux sont dégradés par leur extrémité carboxy-terminale dans les extraits cellulaires. Sans que la raison soit connue, on sait que les peptides signaux sont dégradés 300 fois plus rapidement que les autres peptides. 90% de l'activité de dégradation a lieu dans le cytoplasme (Novak *et al.*, 1986).

Au moins deux SPPases ont été décrites chez *E. coli*: i) une SPPase identique à la protease So qui représente moins de 10% de l'activité et ii) une autre similaire à l'oligopeptidase OpdA de *S. thyphimurium* décrite par (Vimr *et al.*, 1983) et responsable de 90% de l'activité de dégradation cytoplasmique chez *E. coli* (Chen *et al.*, 1987). Les mutants des gènes codant cette deuxième enzyme présentent une croissance normale. L'activité de l'enzyme est stimulée par le Co^{+2} et le Mn^{+2} et inhibée par l'EDTA. L'enzyme hydrolyse des peptides d'une taille de 5 acides aminés minimum avec 2 à 3 résidus de chaque côté de la coupure. Chez *E. coli* il a été observé que la dégradation des peptides signaux est cruciale, puisqu'une inhibition de la translocation de protéines par des peptides signaux synthétiques a été observée in vitro (Chen *et al.*, 1987; Wickner *et al.*, 1987). Les peptides signaux synthétiques de la protéine LamB ont un effet négatif sur la translocation de cette protéine ainsi que sur d'autres protéines hétérologues (Chen *et al.*, 1987). Les auteurs ont également testé si quatre autres peptides signaux synthétiques dérivés du peptide signal de LamB avaient un effet sur la translocation. Il s'est avéré que le peptide sauvage avait la capacité d'inhibition la plus élevée. Les auteurs ne concluent pas si les peptides inhibent le translocon (SecYEG), SecA ou le recepteur FtsY. La dégradation rapide et efficace des peptides signaux serait donc nécessaire pour maintenir la sécrétion de protéines. OpdA a été étudiée également par Dev et Ray (1990) chez *E. coli*. Elle clive le peptide signal en oligopeptides qui sont après hydrolysés par d'autres peptidases inconnues. Les peptides signaux ne s'accumulent pas chez les mutants des hydrolases des peptides signaux, ce qui indique qu'une seule peptidase n'est pas responsable de la dégradation complète des peptides (Dev & Ray, 1990).

Chez *B. subtilis* plusieurs SPPases potentielles existent. La mieux caractérisée est SppA, l'homologue de la protéase IV de *E. coli*. Un mutant de *sppA* chez *B. subtilis* provoque une diminution de la maturation de AmyQ. Pour Bolhius et collaborateurs ce résultat suggére que l'activité de la SPase I est affectée, probablement par l'accumulation de peptides signaux non dégradés (Bolhius *et al.*, 1999). Deux autres SPPase potentielles existent chez *B. subtilis*, il s'agit de TepA et YmfB. En plus d'une certaine homologie de ces deux protéines avec SppA, TepA présente des homologies avec la protéase Clp. Dans un mutant *tepA* le taux de translocation de protéines sécrétées est fortement affecté. Bolhius et collaborateurs (1999) ont proposé trois rôles potentiels pour TepA dans la sécrétion de protéines. D'abord elle pourrait être un analogue de OpdA et dans ce cas la présence de peptides signaux non dégradés affecterait la sécrétion. Le deuxième rôle attribué à TepA est une fonction régulatrice semblable à celle de Clp chez *B. subtilis*, et finalement les auteurs proposent qu'elle pourrait être la chaperonne des protéines sécrétées. Néanmoins, son rôle précis dans la sécrétion n'est pas déterminé.

2.4.1.6 Les lipoprotéines

Les bactéries Gram négatives possèdent une membrane externe qui confine plusieurs protéines à l'espace périplasmique. *E. coli* possède environ 90 lipoprotéines, localisées pour la plupart à la surface de la membrane externe, d'autres se trouvent aussi sur la membrane interne. Elles sont attachées à la membrane par leur partie lipidique et la partie protéique est exposée dans le périplasme quand elles possèdent un signal de rétention ou sont transportées jusqu'à la membrane externe par les protéines Lol quand elles ne possèdent pas ce signal (Narita *et al.*, 2004). D'abord, l'importance d'un aspartate en position 2 a été soulignée pour le maintien des lipoprotéines dans la membrane interne (Yamaguchi *et al.*, 1988). Cette règle de rétention a été élargie ensuite en utilisant une protéine MalE modifiée en lipoprotéine et exprimée dans un mutant *malE*. Elle est fonctionnelle uniquement si elle reste sur la membrane interne, où elle permet le transport du maltose et la croissance dans un milieu contenant comme seule source de sucre le maltose. En remplaçant systématiquement l'aspartate en deuxième position par d'autres acides aminés, les auteurs ont conclu que la phénylalanine, le tryptophane, la tyrosine, la glycine et la proline ont le même effet et permettait de retenir la lipoprotéine sur la membrane interne (Seydel *et al.*, 1999).

Chez les bactéries Gram-positives, la modification des lipoprotéines permet de prévenir leur perte dans l'environnement parce qu'elles restent ancrées à la membrane cytoplasmique. Cela peut expliquer pourquoi 32 lipoprotéines de *B. subtilis* sont homologues aux protéines périplasmiques de *E. coli* (Sutcliffe & Russel, 1995) (Tjalsma *et al.*, 1999a).

En 1981 la présence d'une lipoprotéine chez les bactéries Gram-positives a été montrée pour la première fois : deux laboratoires de recherche ont montré indépendamment que la penicillinase extracellulaire de *Bacillus licheniformis* existe sous une forme lipidique associée à la membrane (Sutcliffe & Russel, 1995). Parmi les lipoprotéines les plus importantes nous trouvons la protéase de paroi (PrtM) chez *L. lactis* et PrtsA, une protéine similaire à 30% à PrtM, chez *Bacillus subtilis.*

Les précurseurs des lipoprotéines possèdent un peptide signal qui leur permet de traverser la membrane et qui est très similaire au peptide signal des protéines sécrétées : une partie amino-terminale chargée positivement, une région hydrophobe centrale et une partie carboxy-terminale contenant le site de clivage pour la signalpeptidase II. La différence majeure entre ce peptide signal et le peptide signal des protéines sécrétées est la lipobox (Tjalsma *et al.*, 1999a). La séquence consensus de la lipobox est : Leu-(Ala/Ser)-(Gly/Ala)-Cys (Narita *et al.*, 2004).

La maturation de la lipobox qui conduit à l'attachement à la membrane, consiste en i) la formation d'un lien thioéther entre la cystéine et le diacylglycérol (deux chaînes d'acides gras liées à une molécule de glycérol) catalysée par la phosphatidylglycerol/prolipoprotein diacylglyceryl transferase (Lgt) et ii) le clivage du peptide signal par la signal peptidase II (Narita *et al.*, 2004) (Figure 2-21).

Chez *Staphylococcus aureus* un mutant *lgt* a été construit. La modification introduite par Lgt est détectée par marquage avec le [^{3}H]palmitate. Chez ce mutant, aucune protéine n'a été détectée par ce marquage. Les lipoprotéines n'ont pas été ancrées correctement à la membrane et certaines se sont trouvées dans la fraction extracellulaire (Stoll *et al.*, 2005).

Les lipoprotéines sont reconnues par la signal peptidase II grace à la lipobox. La SPase II clive le peptide signal de ces protéines. Elle reconnaît la modification diacylglyceryl et clive avant la cystéine (Venema *et al.*, 2003). Une seule SPase II est présente chez *B. subtilis*, *E. coli* et *L. lactis*

(Tjalsma *et al.*, 1999a). La SPase II de *L. lactis*, codée par *lspA*, présente 60 % et 56 % d'homologie avec la SPase II de *B. subtilis* et *E. coli*, respectivement. La SPase II est essentielle chez *E. coli*. Au contraire, elle n'est pas essentielle chez *B. subtilis*. Cependant, au moins une lipoprotéine est essentielle chez *B. subtilis* : PrsA. Dans un mutant de SPase II, le précurseur mais aussi une forme mature sont retrouvés. Cela indique qu'une autre enzyme est capable de cliver le peptide signal de PrsA. Néanmoins, la sécrétion de certaines protéines, dépendantes de PrsA a diminuée. C'est à dire que la forme trouvée a une activité réduite (Tjalsma *et al.*, 1999a). L'absence de la SPaseII chez *B. subtilis* se manifeste par une sensibilité à des températures élevées et faibles. Par l'utilisation d'une alpha amylase biotinilée, utilisée comme marqueur du taux de translocation, les auteurs ont montré que l'absence de la signal peptidase II ne provoquait pas l'accumulation de précurseurs de lipoprotéines mais plutôt un mauvais fonctionnement de certaines d'entre elles (Tjalsma *et al.*, 1999a). Par contre, la signal peptidase II n'est pas nécessaire pour le développement de processus dans lesquels il est connu que des lipoprotéines participent. Il s'agit de processus comme la compétence, la sporulation et la germination. Ce dernier résultat montre que les précurseurs des lipoprotéines nécessaires à ces processus sont maturés de manière alternative ou sont actifs sans être maturés (Tjalsma *et al.*, 1999a). Des résultats semblables ont été obtenus chez *Streptococcus pneumoniae* où 13 lipoprotéines analysées ont gardé la taille de leur précurseurs (ne sont pas maturées) en absence de la SPase II, affectant uniquement la croissance en milieu minimum ou dans le sang mais pas dans un milieu riche. Les auteurs ont démontré que certaines fonctions impliquant des lipoprotéines (principalement des ABC transporteurs) étaient affectées (Khandavilli *et al.*, 2008).

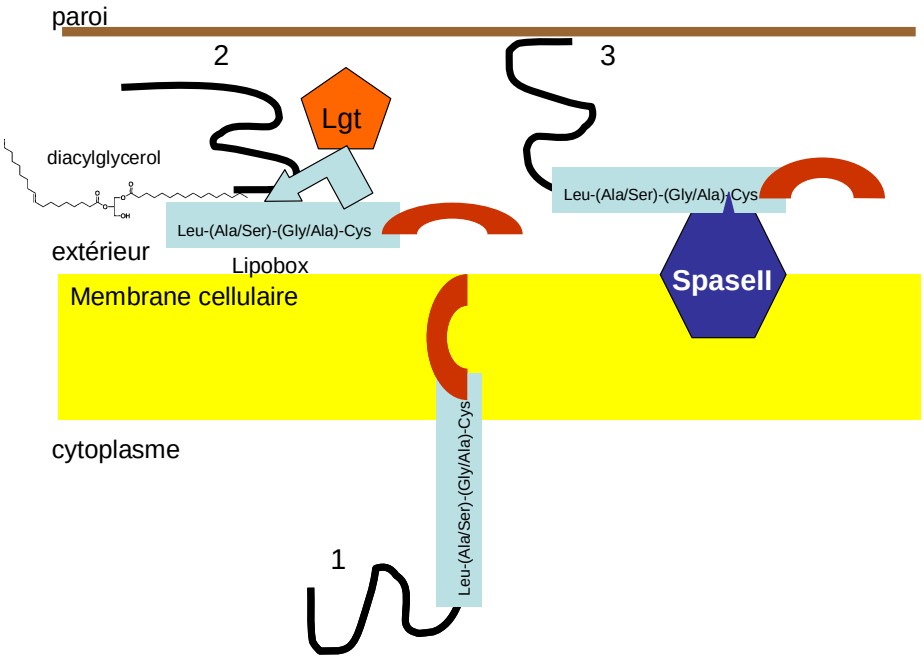


Figure 2-21 : Maturation des lipoprotéines. 1) Translocation 2) Modification par Lgt : ancrage à la membrane par la cystéine de la lipobox. 3) Clivage du peptide signal par la SPase II qui reconnaît la modification réalisée par Lgt **(Narita *et al.*, 2004)**.

2.4.1.7 Les protéines ancrées

Un famille de protéines est liée de manière covalente au peptidoglycane des bactéries Gram-positives. Le mécanisme d'ancrage a initialement été étudié chez la bactérie pathogène *Staphylococcus aureus*. Les protéines ancrées possèdent un PS et un domaine conservé à leur extrémité C-terminale qui est nécessaire et suffisant pour l'ancrage. Cette ancre est constituée de trois parties : un motif LPXTG, une séquence d'une vingtaine d'acides aminés majoritairement hydrophobes et une courte queue de résidus chargés, le plus souvent positivement. Il s'agit alors d'un peptide semblable au peptide signal. Les résidus chargés positivement à l'extrémité du domaine C-terminal empêchent la translocation de l'ancre et positionnent le motif LPXTG pour qu'il soit accessible à la sortase. Ayant accès à l'ancre elle peut cliver la protéine entre la thréonine et la glycine du motif LPXTG. Cela catalyse la formation d'une liaison entre le groupe carboxyle de la thréonine et le groupe amine des ponts du peptidoglycane (Mazmanian *et al.*, 2001). Cette action d'ancrage par la sortase est réalisée une fois les protéines exportées et le peptide signal clivé (Mazmanian *et al.*, 2001). Des mutants de *strA* (le gène codant pour la sortase chez *S. aureus*) ne sont pas capables d'ancrer certaines protéines de surface et de causer certaines infections animales (Mazmanian *et al.*, 2001).

Chez *L. lactis*, deux gènes spécifiant des sortases putatives, *srtA* et *srtB* existent. Les mécanismes d'ancrage semblent être très semblables entre les streptocoques et les lactocoques puisque les protéines hétérologues contenant le motif LPXTG (provenant de *Streptococcus pyogenes*) sont ancrées correctement chez différentes bactéries lactiques (Piard *et al.*, 1997).

2.4.2 Structure du peptidoglycane

Les parois cellulaires des bactéries sont constituées d'un hétéropolymère alternant la N-acétyleglucosamine (β1->4) et l'acide N-acétylemuramique. Ces molécules se trouvent l'une à côté de l'autre liées par des peptides courts (les muropeptides) dans la paroi cellulaire. Des acides téichoïques, des polysaccharides et des protéines sont souvent attachés au peptidoglycane. Chez les bactéries Gram-positives, plusieurs couches de peptidoglycane sont liées de manière covalente (Nelson & Cox, 2000).

Le peptidoglycane de *L. lactis* est constitué d'un monomère (GlcNAc-MurNAc-L-Ala-α-D-Glu-L-Lys-D-Ala-D-Ala). Un D-Asp du peptide qui forme le pont est attaché à la lysine de ce monomère (Schleifer & Kandler, 1972), voir Figure 2-22. Ces résultats ont été confirmés par Courtin *et.al.* lors de l'analyse de muropeptides par spectrométrie de masse (séparés auparavant par HPLC) (Courtin *et al.*, 2006). Les auteurs ont trouvé 35,4% de monomères, 44,6% de dimères, 17.5% de trimères, 2.5% de tétramères et une réticulation de la paroi (cross-linking-index) de 35,8. Le calcul de cet index a été proposé par (Glauner, 1988).

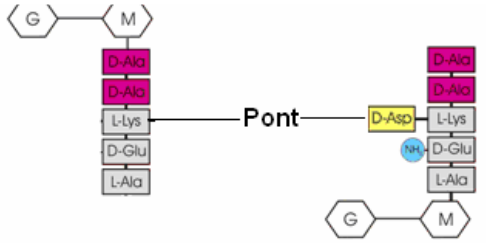


Figure 2-22 : Exemple d'un dimère trouvé dans le peptidoglycane de *L. lactis*.

La synthèse du peptidoglycane est réalisée en plusieurs étapes. Les deux étapes principales sont l'assemblage et la polymérisation. La première étape concerne l'assemblage de l'unité dissaccharide-peptide par une série de précurseurs uridine 5'phosphate (UDP) et des intermediaires lipidiques : MurA et MurF catalysent la formation des précurseurs UDP-Acide N-acétyle muramique (NAM)-pentapeptide à partir de UDP-NAM. Ensuite, le transfert de la partie phospho-pentapeptide-NAM à l'accepteur membranaire est fait par MraY pour obtenir le lipide I. L'addition de N-acétyl-glucosamine par MurG donne le lipide II. Cet intermédiaire lipidique est transloqué à travers la membrane et incorporé dans le peptidoglycane en croissance. La seconde étape concerne la polymérisation des monomères réalisée par des *penicillin binding proteins* et des transpeptidases qui catalysent respectivement la formation des chaînes et des ponts peptidiques (Bhasvar & Brown, 2006) voir Figure 2-23.

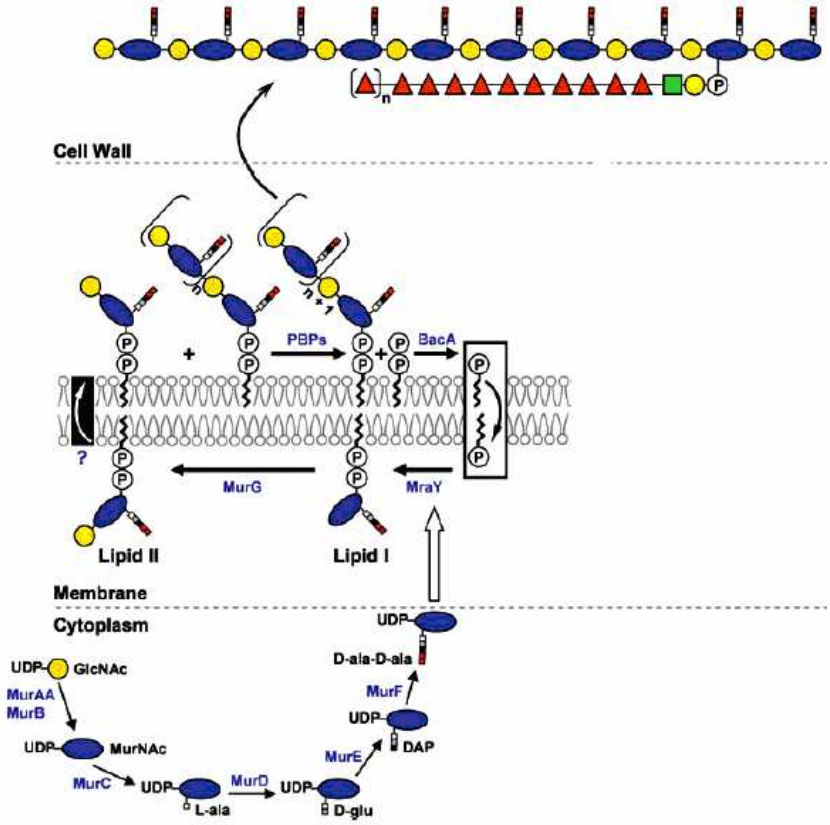


Figure 2-23 : biosynthèse du peptidoglycane (Bhasvar & Brown, 2006).

2.4.3 Métabolisme du pyruvate

Dans cette section, sont présentés uniquement les mécanismes concernant la transformation du pyruvate en différents produits afin de pouvoir comprendre des modifications de cette conversion quand le métabolisme du pyruvate est altéré.

Le pyruvate est le produit final de la deuxième phase de la glycolyse ; il a trois destins possibles (Nelson & Cox, 2000):

1) Dans des conditions d'anaérobiose, le pyruvate est oxydé, avec libération d'une molécule de CO_2 provenant de son groupement carboxyle donnant le groupe acétyle pour l'acétyl-coenzyme A. Ce groupement acétyle est oxydé complètement dans le cycle de l'acide citrique.

2) Son destin chez les levures est un métabolisme dont les produits finaux sont l'éthanol et le CO_2.

3) Il est utilisé pour la production de lactate chez les bactéries lactiques. La réduction du pyruvate est catalysée par la pyruvate déshydrogénase. Pendant la glycolyse, deux molécules de NAD sont converties en NADH par la déshydrogénation de deux molécules de glycéraldéhyde-3-phosphate pour chaque molécule de glucose transformée. La réduction des deux molécules de pyruvate régénère deux molécules de NAD. Le processus est balancé. La conversion d'une molécule de glucose en lactate libère deux ATP.

Lactococcus lactis est un organisme homofermentaire, c'est à dire que le principal produit de la conversion du pyruvate est le lactate (plus de 90%). La conversion du pyruvate en lactate est catalysée par la lactate déshydrogénase (Ldh) (Cocaign-Bousquet *et al.*, 1996). Quand le sucre est en excès, la glycolyse est utilisée pour produire des molécules d'ATP nécessaires pour la croissance cellulaire. La régulation de la glycolyse est faite par le fructose-1,6-diphosphate (FDP) qui active la pyruvate kinase et la Ldh (Thomas *et al.*, 1980). Le résultat est un métabolisme linéaire de la conversion du sucre au lactate (Cocaign-Bousquet *et al.*, 1996).

Sous certaines conditions, chez *L. lactis*, les produits de la fermentation qui généralement ne représentent pas plus de 10%, sont produits en plus grande quantité. Il s'agit de la fermentation mixte (Cocaign-Bousquet *et al.*, 1996). Durant cette fermentation, la glycolyse continue à fournir le pyruvate, mais la transformation en lactate diminue. Cette réponse a été observée pour la première fois dans des cultures en chémostats limitants en sources de carbone. Sous ces conditions, le pyruvate est transformé par la pyruvate-formate-lyase (Pfl) en formate ou par la pyruvate déshydrogénase (Pdh) pour donner des mélanges acétate-éthanol. Ce type fermentation mixte est alors un résultat de la concurrence entre Ldh, Pfk et Pdh (Cocaign-Bousquet *et al.*, 1996). La Pfl est très sensible à l'oxygène et n'est active que lors des conditions d'anaérobiose stricte. La Pdh catalyse une réaction de décarboxylation réductive du pyruvate produisant actétyl-CoA et CO_2 (et la réduction de la coenzyme CoA)˙. Une fois l'acétyl-CoA produit, il est transformé en acétate (par l'acétate kinase) ou en éthanol (par l'aldéhyde déshydrogénase). La formation d'acétate génère un ATP supplémentaire et la production d'éthanol permet de régénérer deux co-enzymes réduits (Cocaign-Bousquet *et al.*, 1996). D'autres produits mineurs sont aussi synthétisés à partir du pyruvate en présence d'oxygène : le diacétyle et l'acétoïne. Ces produits sont synthétisés dans une voie métabolique impliquant l'acétolactate-synthase, qui produit de l'acétolactate. Ce produit est instable et se transforme spontanément en diacétyle en présence d'oxygène. L'acétolactate peut être converti aussi en acétoïne par l'acétolactate décarboxylase (AldB). L'acétoïne peut encore être transformée en butanediol (Cocaign-Bousquet *et al.*, 1996) (voir Figure 2-24).

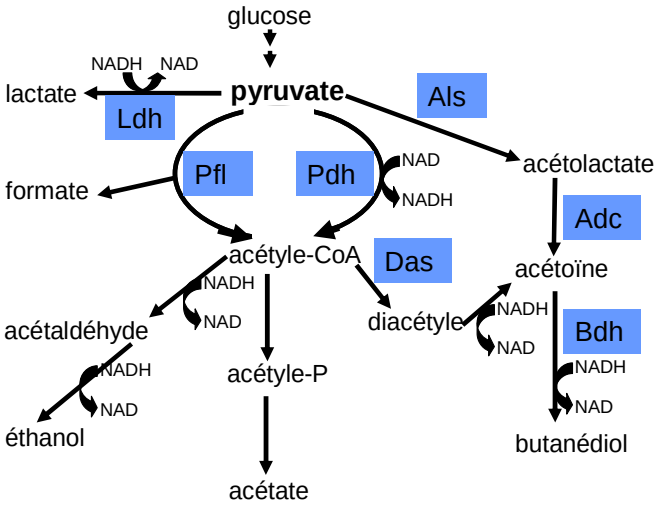


Figure 2-24 : Métabolisme du pyruvate chez *L. lactis*. Ldh : lactate déshydrogénase, Pdh : pyruvate déshydrogénase, Pfl : pyruvate-formate-lyase, Als : acétolactate synthase, Adc : acétolactate décarboxylase, Bdh : butanédiol déshydrogénase, Das : diacéthyl synthase.

3. Résultats de la KCCA

Les résultats obtenus par l'analyse de corrélation canonique bâtie sur des noyaux ou *kernel canonical correlation analysis* (KCCA) concernent i) la validation de la méthode par rapport aux données connues ii) des liens prédits par la méthode pour les protéines cibles PepF et YvjB. Ce dernier résultat est couplé à l'interprétation de ces liens et la proposition d'un rôle des protéines.

3.1 Reconstruction des liens connus des voies métaboliques

Les résultats de la reconstruction des liens connus concernant les voies métaboliques sont montrés dans l'article I. En résumé, le réseau connu est reconstruit avec un taux d'erreur très faible quand toutes les sources de données sont utilisées pour faire la KCCA. Néanmoins, la performance de chaque noyau a été étudiée indépendamment montrant que les deux noyaux les plus informatifs sont les noyaux construits sur les données transcriptomiques et avec les profils phylogénétiques. Ces résultats sont en accord avec les résultats rencontrés par Yamanishi et collaborateurs (Yamanishi *et al.*, 2004b; Yamanishi *et al.*, 2003).

Le réseau connu nous a également permis de réaliser une validation croisée de type *leave-one-out* pour déterminer les meilleurs paramètres des noyaux et de l'analyse de corrélation canonique. Les valeurs des paramètres testés et les valeurs optimales choisies sont explicitées dans les données supplémentaires de l'article I. Les valeurs des paramètres optimaux trouvés pour la combinaison des tous les noyaux ont été utilisées pour réaliser la prédiction de rôle des protéines PepF, YvjB et la prényl-peptidase CAAX. Pour les prédictions de liens de ces trois protéines, ainsi que pour les prédictions de liens d'autres protéines réalisées pour valider la KCCA, le seuil utilisé a été la plus grande distance entre protéines connues comme étant voisines sur le graphe des voies métaboliques. Ainsi, les listes de liens (voir annexe) contiennent uniquement les protéines avec une distance en dessous de ce seuil, bien que la distance avec toutes les protéines du lactocoque, ai été calculée.

3.2 Reconstruction des liens connus : régulons et interactions physiques entre protéines

Nous avons voulu évaluer la performance de la méthode en la comparant aux données déjà connues sur les régulons et sur d'autres méthodes permettant d'évaluer les interactions entre protéines comme le double hybride.

Nous avons voulu savoir si la méthode utilisée nous permettait de retrouver un régulon connu. Nous avons alors regardé les liens prédits pour CodY. C'est un régulon bien caractérisé qui régule plusieurs gènes impliqués dans la protéolyse répondant au *pool* d'acides aminés branchés libres dans la cellule (Guedon *et al.*, 2001b). Il comporte une quarantaine de gènes selon les résultats des études de Guedon *et al.* (2001) et den Hengst *et al.* (2005b). Nous n'avons pas retrouvé la majorité

des gènes du régulon CodY comme liens prédits. Par contre, un lien fort avec GuaB et avec le régulateur CtsR a été prédit. Des liens avec plusieurs protéines hypothétiques et ribosomales sont également prédits. La méthode indique alors un rôle global, en relation avec les réponses au stress, puisque CtsR est un régulateur négatif des gènes *clp* en réponse au stress (Derré *et al.*, 1999). De plus, une implication dans la synthèse de protéines est prédite grâce aux nombreuses protéines ribosomales faisant partie des liens retrouvés par la KCCA. Néanmoins, le régulon, c'est à dire les gènes identifiés comme étant régulés par CodY, ne font pas partie des liens prédits par notre méthode (voir liste complète en annexe). Un rôle régulateur global, impliqué dans la réponse au stress plus que dans la régulation de la protéolyse ressort de notre analyse statistique.

Ensuite, nous avons étudié les liens prédits pour PepN, une protéine connue et faisant partie du régulon CodY. Parmi ces liens nous ne retrouvons que trois gènes faisant partie du même régulon : PepDA1, PepDA2, et PepC. Cependant, le lien le plus proche est prédit avec le régulateur transcriptionnel YeeG (voir liste complète en annexe) et non pas avec le régulateur CodY. En conclusion, la KCCA n'identifie pas le régulon CodY comme étant le principal régulateur protéolytique ce qui suggère l'existence de régulations plus complexes pour les enzymes protéolytiques.

Un autre type d'interaction entre protéines est l'interaction physique directe, telle qu'elle est déterminée par la technique de double-hybride. Peu de données sur des interactions double hybride existent pour le lactocoque. Des expériences, dont les résultats n'ont pas encore été publiés, ont néanmoins été faites dans l'équipe d'Emmanuelle Maguin (Unité de Génétique Microbienne, INRA Jouy en Josas). Les protéines participant à la réplication d'ADN ont été étudiées. Des données sont également disponibles sur *B. subtilis* (Noirot & Noirot-Gros, 2004). Nous avons prédit par KCCA des liens pour les protéines participant à la réplication d'ADN suivantes : AtpA, AtpD, AtpG, AtpH, DnaA, DnaB, DnaD, DnaE, DnaH et DnaN. Le réseau obtenu entre ces protéines est montré en Figure 3-1. Les liens de ces protéines qui ne concernent pas les protéines faisant partie de la réplication ne sont pas montrés sur la Figure 3-1 (voir liste complète en annexe). Les liens obtenus par la méthode KCCA chez le lactocoque ne coïncident pas exactement avec les liens obtenus par double hybride (résultats non publiés pour SigX) ni avec les liens déterminés chez *B. subtilis* (Figure 3-1) mais permettent de retrouver globalement un réseau de protéines dont plusieurs sont impliquées dans la réplication prédisant bien le rôle principal de ces protéines.

Lors de la mise à l'épreuve du KCCA les résultats de la reconstruction des liens du réseau métabolique donnent les meilleurs résultats. La capacité de la méthode à prédire des liens connus des régulons est très faible. Les liens physiques entre protéines ne sont pas directement retrouvés, mais le réseau global de réplication d'ADN est correctement prédit par KCCA.

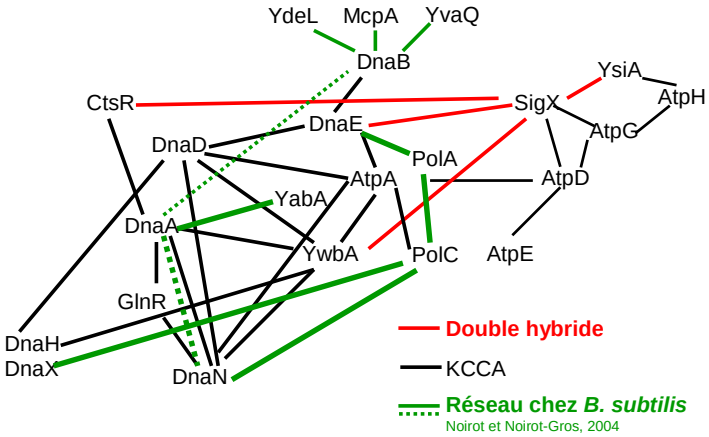


Figure 3-1:
A : Réseau des liens prédits entre protéines participant à la réplication par KCCA (noir) en comparaison avec le réseau trouvé par double hybride chez *L. lactis* (rouge) et le réseau d'interaction de protéines impliquées dans la réplication d'ADN chez *B. subtilis* (vert) (Noirot & Noirot-Gros, 2004). Ligne en pointillée : liens comprenant 1 à 2 intermédiaires.

3.3 Liens prédits et hypothèses sur le rôle des protéines cibles

La KCCA a fourni une matrice de distances entre toutes les protéines pour lesquels l'analyse a été réalisée. Cette liste est constituée de 1756 protéines de *L. lactis* pour des raisons d'harmonisation et de mise en forme des données expliquées en détail dans la section Matériels et Méthodes. A partir de cette matrice il est possible de reconstruire un graphe de relations entre les 1756 protéines. Utilisant comme seuil la plus grande distance entre protéines connues comme étant voisines sur le graphe des voies métaboliques, nous nous sommes intéressés en particulier à trois protéines PepF, YvjB et la prényl-peptidase CAAX (voir Figure 3-2). La liste complète des liens prédits est disponible en annexe.

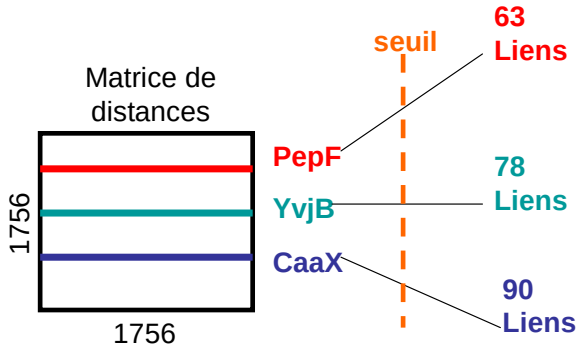


Figure 3-2 : Obtention des liens probables des trois peptidases d'intérêt à partir des distances entre gènes obtenus par KCCA.

La liste de liens prédits est une liste ordonnée, dont nous savons que les premières prédictions (les distances les plus faibles) sont les plus sûres. Compte tenu des protéines prédites comme étant très proches des protéines cibles (dix premières positions) et des fonctions cellulaires représentées par ces protéines, des diagrammes résumant les prédictions ont été établis pour PepF et YvjB (Figure 3-3).

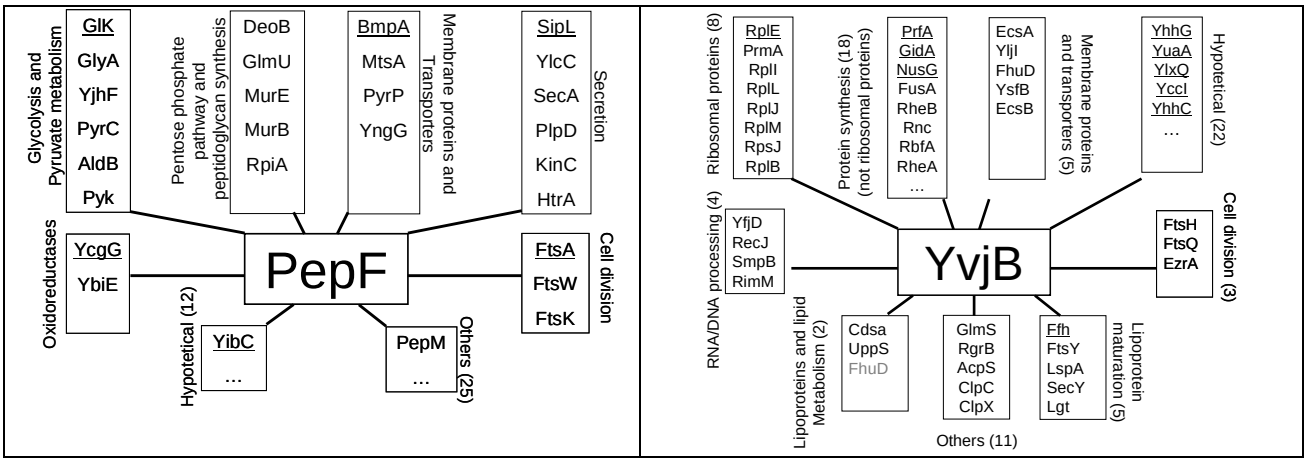


Figure 3-3 : Résumé des liens prédits pour PepF et YvjB par la méthode KCCA.

Ces résultats impliquent la participation probable de PepF dans i) la sécrétion de protéines, ii) le métabolisme du pyruvate, ii) la synthèse du peptidoglycane et iv) la division cellulaire. Compte tenu des connaissances sur PepF comme une oligoendopeptidase (Monnet *et al.*, 1994) et le fait que SipL, la signal peptidase I de *L. lactis*, est la protéine prédite comme étant la plus proche de PepF nous avons posé l'hypothèse que PepF participerait dans la sécrétion de protéines en dégradant le peptide signal clivé par SipL, c'est à dire comme une signal peptide peptidase intracellulaire chargée de dégrader la partie hydrophile du peptide signal (Figure 3-4).

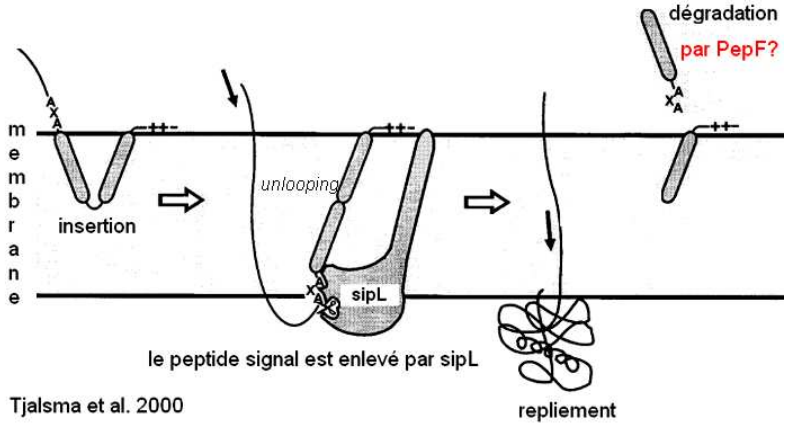


Figure 3-4 : Hypothèse du rôle possible de PepF dans la sécrétion cellulaire compte tenu du modèle existant chez la bactérie Gram positive *B. subtilis* (Tjalsma *et al.*, 2000).

Pour YvjB la protéine prédite comme étant la plus proche est Ffh, qui fait partie de la *signal recognition particle* (SRP) et qui prend en charge les protéines possédant un peptide signal. Le récepteur de SRP dans la membrane se trouve également parmi les liens de YvjB. Compte tenu qu'il s'agit de la voie de sécrétion co-traductionnelle plusieurs protéines participant dans la synthèse de protéines font partie des liens prédits pour YvjB. Ces résultats et le fait que YvjB est une protéine membranaire nous ont permis de poser l'hypothèse que YvjB participe dans la maturation de protéines prises en charge par SRP (Figure 3-5). Compte tenu du fait que la signal peptidase II, ainsi que Lgt (la prolipoprotéine diacylglycerol transférase) chargées de maturer les lipoprotéines, se trouvent liées à YvjB, et des connaissances chez *E. faecalis* (An *et al.*, 1999; Antiporta & Dunny, 2002) nous avons favorisé l'hypothèse que YvjB participerait à la maturation des lipoprotéines.

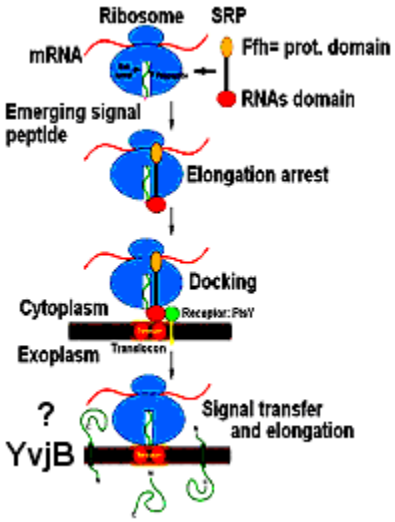


Figure 3-5 : Modèle de la voie de sécrétion co-traductionnelle et possible participation de YvjB (White & von Heijne, 2004).

En résumé, une fois la méthode KCCA validée sur des données connues et les paramètres optimaux déterminés, elle a été utilisée pour obtenir des distances qui ont permit d'établir une liste ordonnée de protéines possiblement liées aux peptidases d'intérêt. Cette liste et les connaissances existantes sur les peptidases ont servi à établir des hypothèses de travail sur le rôle probable de ces peptidases. L'article I détaille les résultats sur la validation de la méthode KCCA. Les articles I et II présentent brièvement les hypothèses biologiques et sont consacrés ensuite à la validation expérimentale de ces hypothèses.

4. Résultats des tests expérimentaux des liens prédits pour PepF

Les résultats des tests expérimentaux des liens prédits de PepF concernent en premier lieu sa participation dans la sécrétion des protéines. Les prédictions ont été faites pour des données de *L. lactis* IL1403. Un mutant *pepF* de cette souche a été construit par délétion d'une partie de la protéine contenant le site actif. Ce mutant a été utilisé pour démontrer l'implication de PepF dans la synthèse de la paroi cellulaire et le cycle du pyruvate. Néanmoins, *L. lactis* ne secrète pas naturellement une grande quantité de protéines. Pour cette raison là, une souche, dans laquelle la sécrétion d'une protéine pouvait être induite (NZ9000-pSEC1) a été utilisée pour étudier sa participation dans la sécrétion de protéines. Nous avons alors construit un deuxième mutant *pepF* dans cette souche. Les résultats de ces expériences sont illustrés dans l'article I. Ils montrent que lors d'une surproduction de protéines exportées (sécrétées ou ancrées à la paroi cellulaire), la sécrétion est affectée dans un mutant *pepF* et que cela est sûrement dû à l'accumulation du peptide signal non-degradé dans le cytoplasme qui encombre toute la machinerie de sécrétion.

Parmi les liens prédits de PepF nous avons aussi retrouvé une des sortases de *L. lactis.* Les protéines substrat de sortases possèdent aussi un peptide signal N-terminal, ainsi qu'un ancre dans la partie C-terminal. Ces deux peptides sont de nature similaire et nous avons voulu déterminer si PepF participe dans la dégradation des ces deux peptides. Le plasmide pVE5547 a été éléctroporé dans le mutant *pepF* de la souche NZ9000 pour induire la surproduction d'un substrat de la sortase. Les résultats de ces expériences n'ont pas été inclus dans l'article I et sont exposés dans la section 4.2.

4.1 Article I: Role of bacterial peptidase F inferred by statistical analysis and further experimental validation

HFSP Journal, 2008, vol. 2, pages 29 à 41.

Role of bacterial peptidase F inferred by statistical analysis and further experimental validation

Liliana Lopez Kleine,[1,2] Véronique Monnet,[1] Christine Pechoux,[3] and Alain Trubuil[2]

[1]INRA Unité de Biochimie Bactérienne, UR477. F-78350 Jouy en Josas, France
[2]INRA Unité de Mathématiques et Informatique Appliquées, UR341. F-78350 Jouy en Josas, France
[3]INRA Plateforme de microscopie électronique, MIMA2. F-78350 Jouy en Josas, France

(Received 28 September 2007; accepted 9 November 2007; published online 7 January 2008)

Despite the quantity of high-throughput data available nowadays, the precise role of many proteins has not been elucidated. Available methods for classifying proteins and reconstructing metabolic networks are efficient for finding global categories, but do not answer the biologist's specific and targeted questions. Following Yamanishi *et al.* [Yamanishi, Y, Vert, JP, Nakaya, A, and Kaneisha, M (2003). "Extraction of correlated clusters from multiple genomic data by generalized kernel canonical correlation analysis." *Bioinformatics* 19, Suppl. 1, i323–i330] we used a kernel canonical correlation analysis (KCCA) to predict the role of the bacterial peptidase PepF. We integrated five existing data types: protein metabolic networks, microarray data, phylogenetic profiles, distances between proteins and incomplete two-dimensional-gel data (for which we propose a completion strategy), available for *Lactococcus lactis* to determine relationships between proteins. The predicted relationships were then used to guide our laboratory work which proved most of the predictions correct. PepF had previously been characterized as a zinc dependent endopeptidase [Nardi, M, Renault, P, and Monnet, V (1997). "Duplication of the *pepF* gene and shuffling of DNA fragments on the lactose plasmid of *Lactococcus lactis*." *J. Bacteriol.* 179, 4164–4171; Monnet, V, Nardi, M, Chopin, MC, and Gripon, JC (1994). "Biochemical and genetic characterization of PepF on oligoendopeptidase from *Lactococcus lactis*." *J. Bio. Chem.* 269, 32070–32076]. Analyzing a PepF mutant, we confirmed its participation in protein secretion through a strong relationship between the signal peptidase I and PepF predicted by the KCCA. The global nature of our approach made it possible to discover pleiotropic roles of the protein which had remained unknown using classical approaches.
[DOI: 10.2976/1.2820377]

CORRESPONDENCE
Alain Trubuil:
alain.trubuil@jouy.inra.fr

Although a lot of high-throughput data are accumulated in databases, a large proportion of known proteins remains uncharacterized until targeted experiments prove their role. In place of analyzing global data, the biologists usually run experiments based on their own knowledge. When the role of a protein is difficult to identify due to the absence of clue or due to inconclusive laboratory results, two different approaches can indeed be considered: run experiments to detect protein interactions or use existing data to predict relationships that guide experiments.

Besides classical experiments to determine the role of single proteins, methodologies based on two-hybrid approaches (Ito *et al.*, 2001) and mass spectrometry of multiprotein complexes (Ho *et al.*, 2002) have been developed to detect protein-protein interactions in yeast. The two-hybrid method requires multiple experimental steps (cloning of the genes into a prey and a bait vector, transformation of

two-hybrid strains with both type of plasmids, mating reactions of all possible combinations and PCR of positive colonies to decode interactions) in order to obtain a satisfactory result (Ito *et al.*, 2001). Knowledge in protein interactions in bacteria has not reached the same level as in yeast (Noirot and Noirot-Gros, 2004). One reason for this difference is certainly due to the fact that this method is time consuming and the results contain many false positives that have to be eliminated through further experiments or analysis. On the other hand, identification of multiprotein complexes needs to set up and run quite heavy experiments, i.e., immunoaffinity purification followed by SDS-PAGE electrophoresis and mass spectrometry (Ho *et al.*, 2002).

Instead of using a technique like the two-hybrid or performing targeted experiments without a clear line of action, scientists can explore existing data, available in databases and unsupervised or supervised approaches to reconstruct protein networks. Several methods based on utilization of different data sources of high-throughput data have been proposed so far (Akesson *et al.*, 2004; Qi *et al.*, 2005; Covert *et al.*, 2004; Aerts *et al.*, 2006; Werhli and Husmeier, 2007).

The protein we are interested in is the zinc dependent oligoendopeptidase PepF. Its possible participation in protein turnover and sporulation had been evoked (Monnet *et al.*, 1994; Nardi *et al.*, 1997; Kanamaru *et al.*, 2002), but no precise role had been determined. Nevertheless, the importance of the protein due to a double copy found in *L. lactis* NCDO 763 and *L. lactis* SK11 (Monnet *et al.*, 1994; Siezen *et al.*, 2005) is intriguing. PepF is found in nearly all low GC bacterial species: (*Staphylococcus* sp., *Bacillus* sp., *Streptococcus* sp., *Lactobacillus* sp., *Mycoplasma* sp., *Clostridium* sp., etc), Spyrochetes, Proteobacteria (*Agrobacterium* sp., *Escherichia coli, Salmonella* sp., *Yersinia* sp., etc.), Archaea (*Halobacterium* sp., *Methanosarcina* sp., etc), Protozoa (*Plasmodium* sp.), and others (*Thermus* sp., *Deinococcus* sp., *Rhodopirellula* sp., etc). The different studies done on PepF in several bacteria have shown it interfering in important cell functions and its inactivation having pleiotropic effects. It is important to specify the role of this widespread bacterial protein with an apparent global and pleiotropic function, in order to control and improve strains used as model organisms as well as in many industrial applications.

To determine the role of the oligoendopeptidase PepF we conducted a study in two parts: (1) inferring possible partners of the protein by a global statistical analysis of existing high throughput data and (2) validating the predicted possible relationships by experimental work. The inference of possible partners of PepF was obtained by constructing a network for all potential proteins coded in the *Lactococcus lactis* IL1403 genome, based on different types of data. The kernel canonical correlation analysis (KCCA) (Yamanishi *et al.*, 2003) we used allowed us to obtain distances between all proteins of the bacterium and place PepF as well as other proteins of the organism in this network. To do this, we integrated four types of data available for all proteins from *L. lactis*: microarray data, phylogenetic profiles, distances between genes (coding for all potential proteins of *L. lactis*) on the chromosome and two-dimensional (2D)-gel data, our standard data set being the protein metabolic network. For 2D-gel data, we designed a new kernel taking care of missing data in such a way that no proteins, available for the other types of data but not present in 2D gel data, have to be discarded from the learning set.

Using KCCA (Yamanishi *et al.*, 2003) we defined 63 possible partners of PepF, belonging to numerous functional categories. Four of them were predominant: protein secretion, pyruvate metabolism, peptidoglycan synthesis and cell division. We experimentally validated PepF's implication in most of these functions. The study of PepF negative mutants confirmed its participation in protein secretion as well as in other predicted functions.

MATERIALS AND METHODS

During the first phase of our study, we did a kernel canonical correlation analysis to predict possible relationships of the peptidase PepF with other proteins of *L. lactis* using the method of Yamanishi *et al.* (2003). In subsequent phases of the work, we compared *pepF* mutants with the wild type strain with special attention to the main functional categories the potential partners of PepF, inferred by the KCCA, belong to.

Inference of possible relationships by kernel canonical correlation analysis

The KCCA is based on classical canonical correlation analysis (CCA) used to measure linear relationships between two groups of variables y and z. The goal is to find linear combinations a_1 and a_2 of y and z that are maximally correlated: $(a_1, a_2) = \arg\max |corr_{\|a_1\|=\|a_2\|=1}(a_1^T y, a_2^T z)|$. The linear combinations are found by eigenvector decomposition and are ordered decreasingly. The KCCA is a regularized CCA based on kernels. This means that the two groups of variables y and z used in CCA are replaced by kernels, inner products between objects, the objects being in our case the proteins. More precisely, the data set S is represented by a square matrix of pair wise comparisons $K=[k(x,x')]_{x,x' \in S \times S}$ (Schölkopf *et al.*, 2004). Then a classical CCA is done between the images of y and z. The advantage of using kernels is that many different data types can be represented as a comparison between objects and that once the kernel obtained different data types are represented in the same way and can be integrated into the same analysis.

We used five existing data types: protein metabolic networks, microarray data, phylogenetic profiles, distances between proteins and 2D-gel data available for *L. lactis*:

The protein metabolic network was constructed from networks representing several metabolic pathways obtained from the database KEGG (Kanehisa *et al.*, 2002). We con-

structed a unique graph, composed of 333 proteins, our golden standard, with all available proteins in version KGML_v0.6.

The available microarray data belong to studies on *L. lactis* IL1403 comparing mutants and wild type strains or two growth conditions. They were obtained in a database harbored at the server of the MIG department at INRA: http://genome.jouy.inra.fr/efp/base/www and at the Gene Expression Omnibus at NCBI; we also used the data of Guedon *et al.* (2001). We treated 51 experiments possessing a different number of repetitions, making 115 hybridizations in total.

The genetic profiles (binary presence/absence vectors for the genes of *L. lactis*) were constructed for all genes from *L. lactis* IL1403. The presence of a gene was evaluated by similarity (BLAST) in 276 completely sequenced bacteria. The ARCT 0.9 program (http://genomics.senescence.info/software/) included in HAGR (de Magalhães *et al.*, 2005) was used to construct the profiles. If the sequence similarity had an E-value lower than 10^{-5} the gene was declared present, otherwise the gene was declared absent from the organism. These data inform about the co-evolution of genes, which is possibly related to a common function.

The position of the genes on the chromosome has been used to calculate the distance between them. We calculated the number of base pairs between the end of one gene and the beginning of the next one, so as to use this measure as distance. This data type has been included because generally neighboring genes participate to the same function in bacteria.

We had 2D gel data from 13 experiences (1 or 2 repetitions) at our disposal, revealing the protein quantity (expressed in volume percentage) of proteins with an isoelectric point between 4 and 7, on two different strains: *L. lactis* IL1403, as all other data used, and *L. lactis* NCDO763. All gels were run in the bacterial biochemistry laboratory (Unité de Biochimie Bactérienne) at INRA (France). Data on ten of these gels had already been published (Guillot *et al.*, 2003 and Gitton *et al.*, 2005). We conducted a test of maximum mean discrepancy described by Borgwardt *et al.* (2006), in order to determine if the data of the strain NCDO763 could be used together with that of IL1403. The conclusion was that data from both strains belong to the same distribution and therefore can be used together. This type of data contains many missing pieces of data compared to the transcriptomic data, phylogenetic profiles and distance on the chromosome. We used a completion strategy in order to deal with the missing data.

To apply a kernel method such as the KCCA, the first step is to define a valid kernel for each type of data. Herein kernels are, as described before, gene similarity matrices. We have at our disposal information on the genes on one hand and the protein metabolic network on the other hand. In order to represent the undirected graph of the protein metabolic network we used a Laplacian exponential diffusion kernel (Kondor and Lafferty, 2002). For the other data we used Gaussian and polynomial kernels. The most appropriate kernels for each type of data are listed in Table I. Before the KCCA was done, all kernels were normalized.

Parameters α_1, α_3, α_4, α_5 as well as component number and regularization parameter (δ) in KCCA were determined with a grid search leave-one-out cross validation (see supplementary data). In each of the 333 iterations, the set of 333 proteins in the golden standard was split into a training set and a test set composed of one protein. The feature space was trained on the training set. The graph was built progressively and compared with the original protein metabolic network, the golden standard. The parameters retained were those that made it possible to find known relationships with the lowest error. The minimal error in regard of the test proteins was calculated using the false positives (f) and the true positives (h), $\bar{e}=\sum_{p=1}^{333} f^p/f^p+h^p \approx \hat{E(e)}$. The most adequate values were chosen to minimize this error.

The 2D-gel data are the protein volumes of the observed protein spots. This quantity was normalized with the total protein volume on the gel to obtain a volume percentage for each protein. We transformed these quantities following the recommendations of Chich *et al.* (2007), transforming the volume percentage ($\%V$) into $T(\%V)$ as follows: $T(\%V)=(\%V)^{1/3}$. If we denote n the number of proteins for which the information is available in the three datasets used to construct K_2, K_3, and K_4, only $n_1 < n$ proteins are present in the dataset used to construct K_5.

Our strategy consists in completing kernel $K_{5(\mathrm{n1}\times\mathrm{n1})}$ to give kernel $K_{5(\mathrm{n}\times\mathrm{n})}$ of the same size of the other kernels. The most simple completion of K_5 would be to replace the missing data by zero ($K_{5\mathrm{zeros}}$). This means that a neutral value (the mean similarity) replaces the missing similarity values after centering of the kernel: $K_{5zeroes}=\begin{pmatrix} K^*_{5(n_1\times n_1)} & O_{n_1\times(n-n_1)} \\ O_{(n-n_1)\times n_1} & Id_{(n-n_1)\times(n-n_1)} \end{pmatrix}$ where n is the number of proteins present in all datasets and n_1 the proteins detected on the 2D gels.

Nevertheless, we have some information about the missing data, i.e., the proteins not detected on the gel, helpful in completing this kernel in a more "informative" way. We propose to create a kernel where missing data will be completed with qualitative data taking into account the information we have about the missing proteins (K_{5quali}). We know that some of the proteins are absent because it is not possible to detect them due to the experimental conditions, and that some others are absent but could have been detected. We supposed the proteins to belong to three object families: $X=X_1 \cup X_2 \cup X_3$. X_1 were the observed proteins, X_2 were the observable proteins (with a pH between 4 and 7 in the case of our dataset) which were undetected on the gels and X_3 were the proteins that were unobservable. We constructed a kernel for the objects belonging to X_1, kernel $K^*_{5(n_1\times n_1)}$. In order to complete the missing data we considered all possible interactions and

Table I. Kernels used for each type of data.

	Kernel			
Datatype	Kernel type	Kernel function	Explanations	Parameters
Metabolic network	Diffusion kernel	$K_1=e^{\alpha_1 L}$	L_{ij} $\begin{cases} 1 & for \quad i=j \\ -d_i & for \quad i=j \\ 0 & otherwise \end{cases}$ where $i\sim j$ means proteins i and j are joined in the graph and d_i is the number of proteins joined to gene i.	α_1 0.01–0.1
Phylogenetic profiles	Polynomial kernel	$K_2=\Phi_2^T\Phi_2$	where Φ_2 is a vector of features the size of which depends on the power of the polynomial in $K_2(x,y)=(\langle x,y\rangle)+\alpha_2)^d=\Phi_2(x)^T\Phi_2(y)$ This kernel construction has been chosen to give a higher weight to the interactions between two genes, than to the interactions of higher order.	$\alpha_2=40$ $d=5$
Distance between genes	Gaussian kernel	$K_3=e^{-\alpha_3 * dispos^2}$	where dispos is the distance in base pairs between the end and the beginning of two genes.	α_3 0.0005
Transcriptomic data	Gaussian kernel	$K_4=e^{-\alpha_4 * DD^2}$	where DD is the norm between the gene expression profiles.	α_4 0.001–0.011
2D-gel data	Completion of a Gaussian kernel	$K_5=\begin{pmatrix} K_5^* & K_{5,a}^T \\ K_{5,a} & K_{5,b} \end{pmatrix}$	$K_5^*=e^{-\alpha_5 * d^2}$ where d is the norm between the protein volume profiles. See the text for $K_{5,a}$ and $K_{5,b}$.	α_5 0.55

replaced the missing value by a value reflecting the interactions between each type of pair (Table II). The result is $K_{5quali(n\times n)}$. The similarity between observed and observable proteins as well as between observable and non-observable (ε) was chosen to be 0.01. The similarity between two observable proteins (θ) was chosen to be 0.02. This means that $K_{5quali(n\times n)}$ contains the values corresponding to the kernel $K^*_{5(n_1\times n_1)}$ for proteins detected on the gels and qualitative values or mean values for the proteins that were not detected (Table II). Other possibilities for the data completion could be considered, for example, different values can be used instead of uniform values.

At this point there was no guarantee that the K_{5quali} was a positive definite kernel (PDK). This was achieved minimizing the Frobenius distance between K_{5quali} and a PDK. The use of this distance has already been described in Yamanishi and Vert (2007). In the present case a PDK (K_5) based on the original K_5^* and the kernel K_{5quali}

Table II. Construction of $K_{5\text{quali}}$ with values for the protein similarities based on the information about missing and available data. Each cell of the table contains the similarity between two proteins x and x'. The values to complete missing data were $\varepsilon=0.01$, $\theta=0.02$, the mean similarity of the detected protein $x\in X_1$ with other proteins inside X_1 (m_x) and the overall mean similarity for the comparison of two nonobservable proteins (X_3). As the similarity between the protein and itself is maximal, the diagonal of $K_{5\text{quali}}$ is composed of ones.

	Observed X_1	Observable X_2	Nonobservable X_3
Observed X_1	$K_5^*(x,x')$	ε	$m_x=\sum_{\substack{x''\neq x \\ x''\in X_1}} \frac{K_5^*(x,x'')}{[\lvert X_1\rvert-1]}$
Observable X_2	ε	$\theta>\varepsilon$	ε
Nonobservable X_3	$m_x=\sum_{\substack{x''\neq x \\ x''\in X_1}} \frac{K_5^*(x,x'')}{[\lvert X_1\rvert-1]}$	ε	$\frac{1}{\lvert X_1\rvert}\sum_{x\in X_1} m_x=\overline{m}$

Table III. Bacterial trains used in this work.

Strain	Plasmid content	Resistance	Reference
E. coli TG1 repA+	pGhost9-pepF deleted (pTIL 120)	Ery	Nardi *et al.* (1997)
L. lactis IL1403	—	—	—
L. lactis IL1403 Δ pepF	—	—	This work
L. lactis IL1403 Δ pepF comp	pILN13-pepF low copy	Ery	This work
L. lactis NZ9000-pSEC1	pSEC1	Cm	de Ruyter *et al.* (1996), Chatel *et al.* (2001)
L. lactis NZ9000 Δ *PepF*	pSEC1	Cm	This work
L. lactis NZ9000 Δ pSEC1 comp	pSEC1 + pILN13-pepF low copy	Ery	This work
L. lactis NZ9000-pSEC1 + pepF	pSEC1 + pILN13-pepF high copy	Cm, Ery	This work
E. coli BL21 (DE3) Gold	—	—	Stratagen
E. coli BL21 (DE3) Gold-pET	pET28-pepF	Km	This work

results from minimizing the Frobenius distance $K_5 = \arg\min_{K\in\mathfrak{I}}\|K_{5quali} - K\|$, where $\mathfrak{I}$ is the set of positive semidefinite matrix of size $n \times n$. The resulting PDK is given by: $K_5 = \begin{pmatrix} K^*_{5(n_1\times n_1)} & K^T_{5quali(n_1\times(n-n_1))} \\ K_{5quali(n_1\times(n-n_1))} & K_{5quali(n_1\times(n-n_1))}(K^*_{5(n_1\times n_1)})^{-1}K^T_{5quali(n_1\times(n-n_1))} \end{pmatrix}$

The goal of KCCA is to find correlations between two datasets. One data set is the golden standard (K_I) and the second data set is composed of microarray data, phylogenetic profiles, etc., aggregated as $K_{II} = K_2 + K_3 + K_4$. A representation of all proteins is constructed in a way that both data sources are as closely correlated as possible. The proteins making up part of each dataset, the protein metabolic network in one hand and of the integrated dataset on the other hand, are represented in a way that reflects the distance between them. To construct this representation we used the method proposed by Yamanishi *et al.* (2003). Given kernels K_I and K_{II}, to each $x \in S$ corresponds a feature $\Phi_I(x)$ [respectively $\Phi_{II}(x)$] belonging to a functional space H_1 (respectively H_2). We searched for a direction f_1 in H_1 (resp. f_2 in H_2) such that the generalized canonical correlation $\rho(f_1, f_2)$ between $u_I(x) = \langle \Phi_I(x), f_1 \rangle$ and $u_{II}(x) = \langle \Phi_{II}(x), f_2 \rangle$ is maximized, where $\rho(f_1, f_2) = cov(u_1, u_2) / \sqrt{\mathrm{var}(u_1) + \delta\|f_1\|^2}\sqrt{\mathrm{var}(u_2) + \delta\|f_2\|^2}$. It is possible to show that $f_1 = \Sigma_{x'\in S}\alpha_{x'}\Phi_I(x')$, resp. $f_2 = \Sigma_{x'\in S}\beta_{x'}\Phi_{II}(x')$, and $u_I^{(x)} = \Sigma_{x'\in S}\alpha^{x'}K_I(x, x')$, resp. $u_{II}^{(x)} = \Sigma_{x'\in S}\beta^{x'}K_{II}(x, x')$. So $cov(u_I u_{II}) = \alpha^T K_I K_{II}\beta$, $\mathrm{var}(u_I) = 1/n\,\alpha^T K_I^2\alpha$, $\mathrm{var}(u_{II}) = 1/n\,\beta^T K_{II}^2\beta$ and $\|f_1\|^2 = \alpha^T K_I\alpha$, $\|f_2\|^2 = \beta^T K_{II}\beta$. Several orthogonal directions can be considered for summarizing the feature space. If we denote $(f_1^{(i)}, i = 1 \ldots m)$, resp. $(f_2^{(i)}, i = 1 \ldots m)$ the directions, then gene x is represented by $u_I(x) = u_I^{(i)}(x), \ldots, u_I^{(m)}(x)$, resp. $u_{II}(x) = u_{II}^{(i)}(x), \ldots, u_{II}^{(m)}(x)$. Therefore the distance between gene x and gene x' will be the Euclidian distance between $u_{II}(x)$ and $u_{II}(x')$.

We verified the validity of different kernel combinations using the parameters obtained by the leave-one-out validation reconstructing known relations for more than one protein at the same time testing different combinations of kernels. The combination of all kernels, which gave the least false positives, was used to make predictions on our protein of interest, PepF.

To define the possible partners of PepF, the threshold was chosen as follows: the highest distance found between proteins known to be neighbors in the protein metabolic network was calculated and chosen to be the maximum distance to accept a relationship between two proteins.

Experimental validation of possible relationships

The bacterial strains and plasmids used in this study are listed in Tables III and IV. *L. lactis* strains were grown at 30 °C in M17 (Difco) medium supplemented with 5% glucose. The chemical minimal medium contained only seven amino acids essential for all lactococci (Cocaign-Bousquet *et al.*, 1995) as well as arginine (1.2 g/l) and threonine (2.3 g/l) essential for *L. lactis* IL1403.

The following antibiotics were added as selective agents when appropriate: erythromycin (5 μg ml^{-1} for *L. lactis*, 150 μg ml^{-1} for *E. coli*), chloramphenicol (5 μg ml^{-1} for *L. lactis*; 20 μg ml^{-1} for *E. coli*) and ampicillin (100 μg ml^{-1} for *E. coli*), kanamycin (20 μg ml^{-1} for *E. coli*).

Table IV. Plasmids used in this work.

Plasmid	Characteristics	Resistance	Reference
pTIL 120	pGhost 9-pepF deleted: 162 b.p. deletion including the active site	Ery	Nardi *et al.* (1997)
pSEC1	expression under *PnisA* encodes SPUsp:NucB	Cm	Chatel *et al.* (2001)
pILN13	allows switch to low copy number	Ery	Renault *et al.* (1996)
pET28	conceived for heterologous protein production	Km	Novagen

Table V. Primers used in this work (the introduced restriction enzyme sites are underlined)

Primer name	Sequence
FpepF	GCGGATATTAAGTTACCTATGGT
RpepF	TTTGGCAATTACTTCTAAAGGAT
PetPepF-For	CATGCCATGGTTGCTAAGAATAGAAATGAAAT
PetPepF-Rev	GGAAGATCTAAGATGGACTCCTTTTTCAA
PilPepF-For	CTGCAGGCAGAAGAAGGATATGAATGAATG
PilPepF-Rev	GCGGCCGCATTTTTAAAGATGGACTCCTTTTTCAAC

Molecular cloning techniques were performed using standard procedures (Sambrook and Russel, 2001). Plasmids were extracted by using a QIAprep Spin miniprep kit (Qiagen). Total *L. lactis* and *E. coli* DNA was isolated as described previously (Hoffman and Winston, 1987). Restriction enzymes (New England Biolabs), T4 DNA ligase from the Fast-link ligation kit, (Epicentre), *Taq* DNA polymerase MP (Qbiogene) and the TripleMaster PCR system (Eppendorf) were used according to the suppliers' recommendations. PCRs were run using a Mastercycler gradient thermal cycler (Eppendorf). All constructions were verified by sequencing with an Applied Biosystems 310 automated DNA sequencer using the ABI PRISM Dye Terminator Cycle Sequencing Kit (Perkin Elmer). Primers for *L. lactis* were selected on IL1403 (Bolotin *et al.*, 2001) and for *E. coli* on K12 (Blattner *et al.*, 1997) genome sequences. The oligonucleotides were purchased from Invitrogen. Annealing was performed between 52 and 60 °C, depending on the primers used.

We constructed *pepF* mutants by replacement with a deleted *pepF* gene as explained in Nardi *et al.* (1997), electroporating pTIL120 into *L. lactis* IL1403 and NZ9000. Integration of pTIL120 into the chromosome and subsequent excision was achieved using the thermosensibility of the plasmid. Mutant strains were screened first on their resistance to erythromycin and second on the size of the fragment amplified by PCR with primers FpepF and RpepF (Table V). The presence of a correct insertional event and absence of the vector was further verified by Southern blotting using and ECL detection system (Amersham).

A sequence corresponding to *pepF* was amplified by PCR from *L. lactis* IL1403 total DNA with the primers PetPepF-For and PetPepF-Rev (Table V) containing *NcoI* and *BglII* sites. PCR fragments were digested and cloned in-frame upstream of the hexa-His pET28 vector (Novagen). *E. coli* BL21 (DE3) Gold competent cells were transformed with the resulting plasmids.

The whole *pepF* gene with its promoter region was introduced into the plasmid pILN13 (or pILNew) (Renault *et al.*, 1996). PILNew is a high copy number plasmid. It is possible to transform it into a low copy number plasmid restoring the replication repressor by a *Kpn*I restriction and further ligation. We used this construction for two purposes: (i) in high copy number to study the effects on protein secretion of the overproduction of PepF in the NZ9000-pSEC1 strain and (ii) in low copy number to complement our mutants. The primers used to amplify the gene with its promoter region and to introduce the needed restriction sites, *Pst*I and *Not*I, were PilPepF-For and PilPepF-Rev (Table V).

Cellular extracts were prepared to analyze PepF activity, presence of peptides, and metabolites of the pyruvate metabolism. Once the cultures had been harvested at $OD_{600\ nm}$=0.6 by centrifugation and washed with phosphate buffer 0.2 M, cell lysis was achieved with a cell disruptor (Constant System Ltd). The cytoplasmic fraction was obtained by ultracentrifugation at 4 °C and 50 000 rpm for 20 min (Centrikon T-1080, Kontron intruments).

PepF activity was measured by its hydrolytic activity with a fluorescent quenched substrate: Mc-Pro-Leu-Gly-Pro-Lys-(DNP)OH. The fluorescence is emitted when the peptide is cleaved (Tisljar *et al.*, 1990) and was followed over 100 s on a spectrofluorometer 25 (Kontron instruments).

Pyruvate, acetate, lactate and formate in cell extracts were quantified by HPLC on an Aminex-HPX-87H column (BioRad) with an isocratic elution with 5 mM H_2SO_4 at a flow rate of 0.35 ml/min at 35 °C. Proteins had previously been precipitated with H_2SO_4 (2% final concentration). The peak surfaces obtained were integrated and the quantity of acid was calculated by comparison with the calibration curve of each acid of interest.

Peptidoglycan from *L. lactis* IL1403 and its *pepF* mutant was prepared from an exponentially growing culture ($OD_{600\ nm}$=0.3) according to the protocol of Atrih *et al.* (1999). Briefly, cells were boiled in 4% (w/v) sodium dodecyl sulfate (SDS) for 30 min. The insoluble cell wall was washed six times with distilled water to wash out the SDS. To remove the proteins, the cell wall pellet was treated with Pronase (200 μg ml^{-1}) for 16 h at 37 °C, then with trypsin (200 μg ml^{-1}) for 16 h at 37 °C. The cell wall was then treated with fluorohydric acid to eliminate teichoic acids. Once a final digestion with muramidase had been completed, the muropeptides were reduced with borate and analyzed by HPLC on a C18 Hypersyl PEP100 column. Muropeptides were identified by comparison with a standard in which peptides had been identified by mass spectrometry. Once the number of the relative quantity (peak surface) of dimers (d), trimers (t), and tetramers (te) of muropeptides had been ob-

tained, a cross-linking index reflecting the peptidoglycan reticulation level was calculated as follows (Glauner, 1988): *Cross-linkage*=1/2Σ*d*+2/3Σ*t*+3/4Σ*te*/Σ*muropeptides*.

Secretion was studied exclusively in the strain NZ9000-pSEC1 which allows the induction of a secreted nuclease (NucB) of *Staphylococcus aureus* possessing the signal peptide of the lactococcal protein Usp45 and under control of the nisin promoter. As NZ9000 possesses the *nisRK* genes on its chromosome it is possible to induce the production and further secretion of NucB by addition of nisin to the growth medium. We tested four different nisin doses: 1, 2.5, 5, and 7.5 ng/ml. Protein cellular extracts were prepared from 200 ml culture after 3 h nisin induction, which we had started at $OD_{600\ nm}$=0.5. Cell lysis was achieved by disruption (Constant System Ltd.) in $NaP0_4$ 20 mmol buffer containing protease inhibitor cocktail P8465 (Sigma-Aldrich). After ultracentrifugation, the supernatant contained the intracellular extract and the pellets containing the envelope fraction were resuspended in the same $NaP0_4$ buffer. In order to compare the profiles between wild type and mutant, with induction of 5 ng/ml nisin, a constant number of cells was analyzed on 4–12% polyacrylamide gels (NuPAGE).

The secreted proteins were prepared from 5 ml supernatant (after centrifugation of 6 ml), precipitated with TCA (20% final concentration), incubated for 30 min at 4 °C and centrifuged for another 30 min at 4 °C. Once the supernatant had been eliminated, the pellet was washed with cold acetone, air dried and resuspended in 500 µl NaOH 50 mM before analysis on 4–12% polyacrylamide gels (NuPAGE).

This test is based on the measurement of mono- and dinucleotides released by the DNA hydrolysis activity of NucB. The incubation was carried out at 37 °C in 500 ml buffer (Tris 25 mM pH 8,8; $CaCl_2$ 10 mM; BSA 0,1 mg/ml) containing 1 mg/ml sonicated salmon sperm DNA (Sigma) with 10 µl supernatant. The reaction was stopped by the addition of perchloric acid, which precipitates non-hydrolyzed DNA. After incubation for 15 min and centrifugation for 7 min, the optical density corresponding to the liberated nucleotides was measured at 260 nm.

The first step in determining the accumulation of the signal peptide was a separation of intracellular and envelope proteins of wild type and mutant by one-dimensional sodium dodecyl sulfate-polyacrylamide (4–12%) gel electrophoresis (SDS-PAGE) using the NuPage system (Invitrogen). We then cut the gel in the molecular weight range between 0 and 6 kDa. Peptides were obtained from these fragments through three subsequent washes with ACN 50% followed by TFA 0.1% without digestion. The obtained peptides were pooled and dried in a SpeedVac concentrator for 1 h, and then resolubilized in 25 µl of HPLC loading buffer (0.08% TFA and 2% ACN) and then analyzed by LC-MS-MS. The peptide mixtures (4 µl) were injected onto the precolumn PepMap C18 (300 µm ID×5 mm, 100 Å) with a flow rate of 20 µl/min to remove salts. The peptides were analyzed in a 50 min gradient of 2–80% of acetonitrile in water containing 0.1% formic acid. A flow rate of 300 nl/min was used to elute peptides from the C-18 PepMap100 reversed-phase nanocolumn (75 µm ID×15 cm, 3 µm, 100 Å) (LC Packings, Amsterdam, The Netherlands) to a PicoTipTMEMITER nanospray needle (360 OD×20 µm, 10 µm ID) (New Objective, USA) for ionization and peptide fragmentation on an ion trap mass spectrometer. MS/MS spectra were acquired for the 200–2000 m/z range and batch processed by using Bioworks 3.2 software packages and searched against the *L. lactis* MG1363 (NZ9000 being a derivate of this strain) protein database using SEQUEST software.

A culture of *E. coli* containing the plasmid to induce heterologous production of PepF-6histidines under the T7 promoter was done in LB medium. The production was induced by adding IPTG at a final concentration of 1 mM to the culture at an $OD_{650\ nm}$ of 0.5. Bacteria were grown at 37 °C until IPTG addition and were then transferred at 30 °C during the expression time (4 h) to avoid the formation of inclusion bodies. The cells were harvested by centrifugation and broken by one passage at a pressure of 1600 bar with a Constant Cell Disruption System. The soluble fraction containing the recombinant protein was collected by centrifugation at 15 000 g for 15 min at 4 °C. The hexa-His-tagged proteins were purified by affinity chromatography on Ni^{2+}-nitrilotriacetic acid spin columns (Qiagen) according to the manufacturer's instructions.

For the localization of PepF in the cell, polyclonal antibodies raised against PepF were produced by PARIS (Production d'Anticorps, Réactifs Immunologiques & Services, Compiegne, France). Once the antibodies had been tested by a Western Blot, cell cultures were fixed in 4% paraformaldehyde, then dehydrated for 1 h in ethanol 30% (at 4 °C), 50%, 70%, (at −20 °C), 90% and 100% (each at −35 °C) and immersed at −35 °C for 3 h in three baths of Lowicryl K4M (Delta microscopies-Labège-France)/ethanol 100% (1v/2v, 1v/1v, and 2v/1v, respectively), followed by two baths in Lowicryl K4M (16 h and 2 h). Polymerization was done at 320 nm for 48 h at −35 °C, and increased temperature, for three days, up to 20 °C following Leica AFS procedure (Leica-Microsystems Rueil-Malmaison—France). Thin sections (90 nm) were mounted on nickel grids. After blocking reaction in buffer containing polybutene sulfone (PBS)-1% BSA-0.1% cold water fish skin gelatin, thin sections were incubated for 2 h at room temperature with the PARIS antibody raised against PepF, diluted 1:100 from solution at 4.45 mg/ml, in buffer containing 0.1% PBS and 0.1% BSAc (Aurion-BioValley-France). The grids were rinsed in the same buffer and incubated for 30 mn in protein A (1:20) conjugated with gold particles of 10 nm (Aurion-BioValley-France). Once PepF molecules were observable, 50 independent cells of each culture (*L. lactis* NZ9000 nisin induced

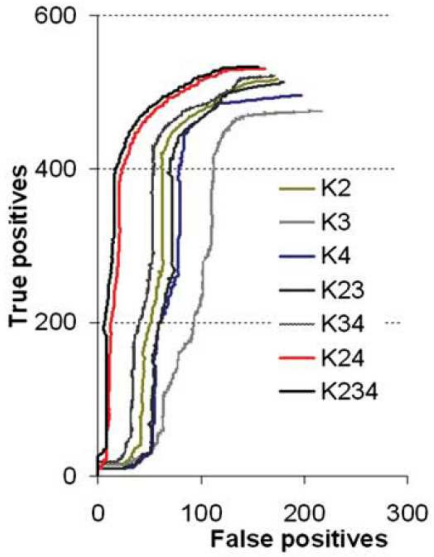


Figure 1. Kernel performance represented by their capacity to find known relationships between proteins (true positives) in comparison with the wrongly detected relationships (false positives). K_2: polynomial kernel constructed on the phylogenetic profiles; K_3: Gaussian kernel constructed on the distance between genes on the chromosome; K_4: Gaussian kernel constructed on the gene expression profiles from microarray data. The combination of kernels (i.e., *K*23) was done by the addition of each kernel: K_2+K_3.

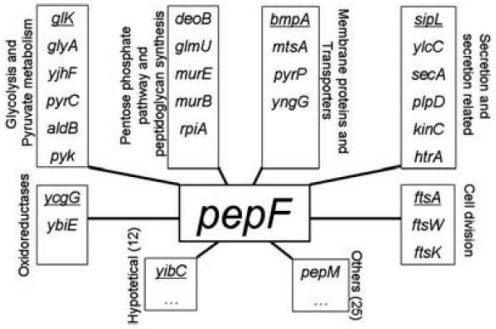


Figure 2. Graphical representation of the predicted relationships found for PepF by KCCA organized by functional categories. Experimental validation was done for the most represented functional groups. Proteins belonging to the first ten relationships of PepF and represented here are underlined. For the exact distances please refer to the supplementary data.

and not induced) were observed and the quantity of marked cells and of molecules per cell was counted.

To study cell division, cultures of IL1403, NZ9000-pSEC1 and their *pepF* mutants were harvested and fixed chemically with a solution of 2% glutaraldehyde and 0.1M sodium cacodylate, included in resin Epon, and then cut at room temperature. The cultures of NZ9000-pSEC1 were induced for 2 h with 5 ng/ml nisin, which we started at OD_{600}=0.3. All sections were examined with a Zeiss EM902 electron microscope operated at 80 kV and images were acquired with a charge-coupled device camera (Megaview III) and analyzed with ITEM Software (Eloïse, France; MIMA2 Platform, INRA-CRJ).

RESULTS

Our results include (1) the inference: validation of the KCCA on known relationships and (2) application of this method to infer new relationships and the experimental validation of the predicted relationships.

Inference

The protein metabolic network constructed from the metabolic pathways existing in KEGG for *L. lactis* contains 333 proteins. In order to test the performance of each kernel we evaluated their ability to reconstruct the known protein metabolic. In Fig. 1 we plotted the mean false positives against the mean true positives until the expected number of arrows (present in the known network) was placed. When kernels were used alone, their performances were not good, for example, when K_3 was used alone the highest number of false positives was found. The combination of the kernels K_2, K_3, and K_4 turned out to be the best one (Fig. 1) and was used later on to make the predictions for PepF.

In order to evaluate the performance of *K*5 (proteomic kernel), we worked with a smaller group of 104 proteins for which 2D-gel data were available. We constructed the proteomic kernel K5ori for this group of proteins and tested its performance to reconstruct the network obtaining an error (percentage of false positives) of 0.385. We then split this dataset into two parts and used only 54 proteins to construct the kernel. Using the strategy of kernel completion, we constructed *K5Q* and obtained an error of 0.394. The completed kernel K_5 has the highest error in comparison with the kernels constructed for the other data: K_2, K_3, and K_4 (error =0.294 for K_2; 0.356 for K_3; and 0.256 for K_4, Fig. 1). It should be noticed that all kernels, if used alone, have a low performance. The best results were obtained when data were fused by summation. Using all kernels together, we decreased the error from 0.18 (obtained for K_{234}, Fig. 1) to 0.17 (obtained for K_{2345}).

Using three kernels (K_2, K_3, and K_4) (Table I) 63 proteins were found to be potentially related to PepF (Fig. 2); see supplementary data for complete list of predicted relationships and exact distances. This means that the distance to these 63 proteins was below the chosen threshold (maximal obtained distance between two proteins known to be related on the protein metabolic network). Using the four kernels (K_2, K_3, K_4, and K_5) we found very similar results: 65 proteins potentially related to PepF; 61 proteins belong to the 63 found using K_2, K_3, and K_4, and four were proteins which had not been identified before: MalE, DfP, ArsC, and FtsQ (see supplementary data).

The proteins which were found to be related to PepF can be divided into two main categories: (i) proteins belonging to

known metabolic pathways (represented on the network) and (ii) proteins not belonging to known metabolic pathways (not represented on the network). The relationships of PepF to proteins of the first category belong principally to two metabolic pathways: pyruvate metabolism and peptidoglycan synthesis. In the second category, we found enzymes responsible for cell division and protein secretion that are strongly represented. Moreover, the strongest relationship (the shortest distance) was found to the signal peptidase SipL. This relationship was reinforced by several common relations of these two enzymes with proteins belonging to both categories mentioned above (Fig. 2).

Experimental validation of predicted relationships for PepF

In order to determine possible relationships of PepF with the predicted enzymes, we constructed two deletion mutants of *pepF* in the *L. lactis* IL1403 and NZ9000 strains and compared the phenotypes of wild type and mutant. The *L. lactis* IL1403 strain is the strain for which the data for the predictions was used; *L. lactis* NZ9000 contains the pSEC1plasmid (Chatel *et al.*, 2001) that carries the gene coding for an exported nuclease. We used this strain because current lactococci export only few proteins and we wanted to test our hypothesis in a strain where secretion could be boosted and regulated. This construction allowed us to overproduce a secreted nuclease (NucB) which had the signal peptide of Usp45 (sp_{45}), the naturally most secreted protein in lactococci (Chatel *et al.*, 2001). As *nucB* expression depended on a nisin inducible promoter, we were able to determine the effect of the absence of PepF in high secretion conditions.

The approach of Tisljar *et al.* (1990) using a quenched fluorescent substrate was used to assess the absence of PepF activity in the *pepF* mutant and *pepF* overproducing strain, as well as to check that activity was restored in complemented mutants (see supplementary data).

We analyzed the acids present in the supernatant by HPLC (Fig. 3). We observed that the PepF mutant strains produce less lactate. In *L. lactis* NZ9000-PSEC1 the pyruvate quantity of the mutant increases in comparison with the wild type. Modifications of acetate quantities were also observed but the two strains behave in different ways. These results confirmed an alteration of the pyruvate metabolism in the absence of PepF.

For *L. lactis* IL1403 we found that, in rich media, the mutant showed a more reticulated peptidoglycan. The relative measurement of dimers, trimers, and tetramers of muropeptides of peptidoglycan makes it possible to calculate the cross-linking index which was higher in the *pepF* mutant (45.26) than in the wild type (35.36), indicating a higher reticulation level of the cell wall in the mutant. In regards to the differences in peptidoglycan composition, we studied the resistance of mutant and wild type to 1 mg/ml lysozyme. As expected, the mutant, with a more reticulated peptidoglycan,

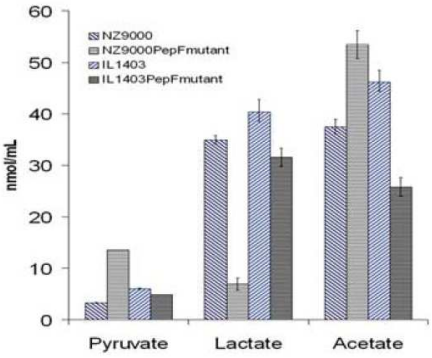


Figure 3. Quantity (in nm/ml) of acids of the pyruvate metabolism in the culture supernantant determined by HPLC comparing wild type (blue) and mutant strains (black) of *L. lactis* IL1403 and NZ9000-PSEC1.

was more resistant than the wild type to lysozyme (Fig. 4).

We measured the activity of the nisin induced secreted nuclease NucB in the supernatant and we did SDS-page gels of secreted proteins. We observed that the export of proteins was negatively affected in the *pepF* mutant when a high quantity of nuclease was produced (induction with 5 and 7.5 ng/ml of nisin) (Fig. 5) and, correlated with this result, we observed a rundown in the nuclease activity in the supernatant (data not shown). We confirmed that this phenomenon was only provoked by the absence of PepF, since in the complemented mutant, secretion was restored. We also tested the effect of an overproduction of *pepF* by transforming *L. lactis* NZ9000-pSEC1 with a high copy number plasmid carrying the *pepF* gene. No obvious difference was observed in the overproducing strain, suggesting that the quantity of PepF present in the wild type was not limiting. Additionally, the observation of strain NZ9000-PSEC1 by electronic microscopy allowed us to observe a detachment of the cell wall from the cytoplasm (Fig. 6), which we attributed

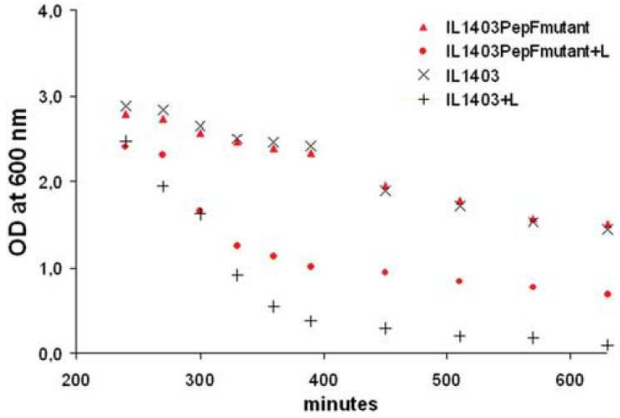


Figure 4. Response of *L. lactis* IL1403 wild type and *pepF* mutant cultures at stationary phase to 1 mg/ml lysozyme (+L). The cell densities were measured at an optical density of 600 nm.

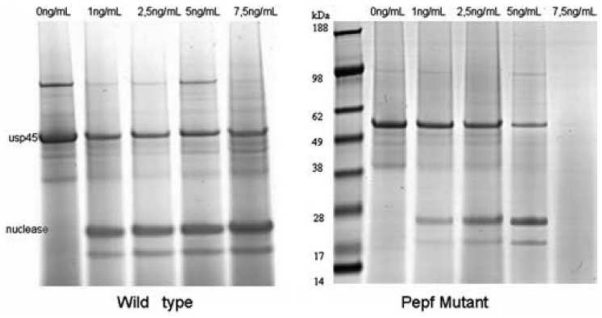


Figure 5. SDS-PAGE electrophoresis of secreted proteins at different nisin concentrations comparing the wild type NZ9000-pSEC1 nuclease overproducing strain with its *pepF* mutant.

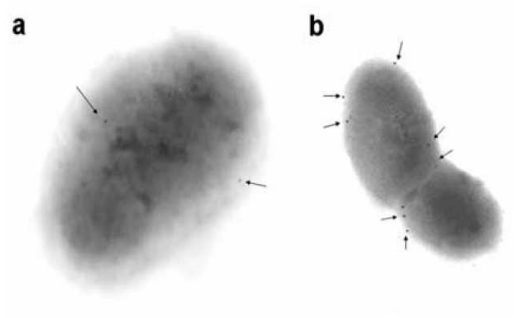


Figure 7. Electronic microscopy observations of the immunogold-labeled peptidase PepF in *L. lactis* NZ9000-PSEC1. When secretion is not induced PepF (a) then after induction with 5 ng/mL nisin (b). The arrows indicate gold-labeled PepF molecules.

to the strongly induced secretion of the nuclease. It was possible to observe that this detachment did not occur in the mutant strain, in which, as we know, protein secretion was diminished or even absent.

The analysis of wild type and mutant strains allowed us to detect a peptide in the low molecular weight fraction of the cell cytoplasm of the *pepF* mutant corresponding to the first ten amino acids of the signal peptide of Usp45. This signal peptide is the most abundant signal peptide present in the cell because of the induction of the expression of the nuclease having the signal peptide of Usp45. The sequence of this peptide is: MKKIISAILMSTVVLSAAAPLSGVYA. The detected peptide was: MKKIISAILM with two variants corresponding to different oxidation states of methionin: MKKIISAILMox (646.6837 kDa) and MoxKKIISAILMox (654.2787 kDa). These peptides were undetected in the wild type's same fraction.

We were able to observe PepF in the periphery of the cell, in *L. lactis* NZ9000-pSEC1 (Fig. 7). Furthermore, in the nisin induced culture of NZ9000-pSEC1 we observed a higher percentage of bacterial sections with at least one PepF labeled molecule: approximately 70% in comparison to 30% of the non-induced ones. The number of PepF molecules detected in the not induced culture is also lower (Fig. 7).

As far as cell division is concerned (i.e., morphology of the septum) no differences were observed in the electronic microscopy observations.

DISCUSSION

This research sought to assess the role of the bacterial peptidase PepF using a global statistical approach to guide experimental studies. Our approach took advantage of existing knowledge since available heterogeneous data had been analyzed before experimental work was started. The approach allowed us (i) to discover a global role of PepF and decipher consequences of the principal function of this protein and (ii) guide laboratory work in order to avoid useless and time consuming experiments.

The results obtained during the validation of the KCCA with known proteins making up part of the metabolic network proved the power of the method. The error rates on regaining the known protein metabolic networks are similar to the ones obtained by Yamanishi *et al.* (2003) and Yamanishi and Kanehisa (2004). The good quality of predictions is certainly due to both the possibility of integrating more than one data type and the training with a known network by correla-

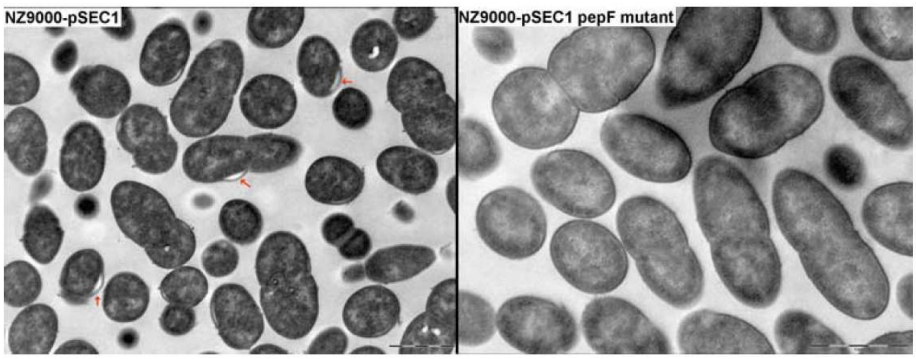


Figure 6. Electronic microscopy observations of *L. lactis* NZ9000-PSEC1 showing the detachment of the cell wall (gray arrows) in the nuclease overproducing strain at 5 mg/ml nisin induction compared to its *pepF* mutant.

tion to the metabolic network. The use of different data types and the fact that we were able to introduce even data with missing values, allowed us to improve predictions and added even more flexibility to the KCCA than before. The strategy we introduce can be a starting point for any type of missing data. The kernel completion we propose for 2D gel data can be improved by the use of different similarity values instead of uniform values, for example.

The KCCA predictions allowed us to find a strong relationship with secretion proteins that had not been evoked in previous studies on PepF. SipL, the protein that was predicted to be the closest to PepF (see supplementary data) is the signal peptidase I that cleaves the signal peptide from secreted proteins other than lipoproteins. In the model for *B. subtilis* proposed by Tjalsma *et al.*, (2000) the signal peptidase SipL cleaves the signal peptide of secreted proteins at the moment of translocation. SipL is thought to cut this peptide into two parts again separating both parts and allowing the hydrophilic fragment to reach the cytoplasm (Tjalsma *et al.*, 2000), where the degradation activity by signal peptide peptidases (sppases) takes place (Novak *et al.*, 1982). Using a *pepF* mutant we determined that PepF is needed to achieve the secretion of proteins. The blockage of protein secretion in the mutant and the presence of a hydrophilic fragment of the signal peptide in the *pepF* mutant support the hypothesis that PepF is a signal peptide peptidase (sppase). Its principal function would consist in hydrolyzing and recycling the liberated signal peptide. As PepF is an endopeptidase that hydrolyses peptides between 7 and 17 amino acids (Nardi *et al.*, 1997) the hydrophilic fragment of the signal peptide is in its range of action. Furthermore, an inhibition of secretion by signal peptides has been observed in *E. coli* (Chen *et al.*, 1987; Wicker *et al.*, 1987). We think that the blockage of protein secretion observed in our *pepF* mutant is due to an accumulation of fragments of the signal peptide. As growth is not affected in the absence of PepF and the effects on protein secretion are only observed during strong induction, it seems that PepF is not the sole sppase in *L. lactis*. Similarly, two sppases exist in *B. subtilis*: the membrane bound SppA (*yteI*) and the cytoplasmic TepA (Tjalsma *et al.*, 2000). The fact that we observed PepF localized in the periphery of the cell reinforces its participation in protein secretion, which occurs in the cell membrane. The fact that PepF is more abundant when secretion is induced also confirms its role in this process. In light of our results, the presence of a second copy of *pepF* on the lactose plasmids of *L. lactis* NCDO763 strains (Monnet *et al.*, 1994; Siezen *et al.*, 2005) together with proteins needed for growth in milk, as, for example, the cell-envelope- protease that has to be secreted and the peptide transport system OppCBFD (Siezen *et al.*, 2005) becomes clearer. In a general manner, it is not surprising that an additional copy of *pepF* is required for the proper localization of membrane and surface proteins implicated in casein processing and peptide transport. Among the membrane proteins with possible relationships to PepF we found, in one of the first positions, BmpA, a basic membrane lipoprotein of unknown function that is in fact an outer membrane in *Borrelia burgdorferi* (Shin *et al.*, 2004) and thus has to undergo secretion in this organism. In *L. lactis* it possesses a signal peptide of 27–30 amino acids as predicted by SignalP 3.0 (Bendtsen, 2004). It is therefore not surprising that this possibly secreted protein has a strong relationship with PepF.

OpdA from *E. coli* is a protein similar to PepF showing 53% similarity around the active site (positions in the amino acid sequence 378–439 and 457–525 of PepF and OpdA, respectively). OpdA has been described as being a possible sppase (Dev and Ray, 1990; Novak *et al.*, 1982; Ichihara *et al.*, 1984). *OpdA* mutants affect the secretion of several proteins (Emr and Bassford, 1982; Emr and Silhavy, 1980; Conlin *et al.* 1992). We tried to complement our *pepF* mutants with *opdA* from *E. coli* to prove that both proteins have the same function but *opdA* seems to be toxic in *L. lactis*, because we were not able to obtain cells containing the pILN13 plasmid containing the *opdA* gene (data not shown).

When studied in several bacteria, both PepF and OpdA were implicated in several functions. We have shown that peptidoglycan structure was modified in the *pepF* mutant in rich medium. We attribute this change to a collateral response to the absence of PepF that causes a stress. A change in the peptidoglycan structure has been documented in response to osmotic and nutritional stresses in *Lactobacillus* (Piuri *et al.*, 2005) or in *E. coli* (Gyaneshwar *et al.*, 2005). In regards to the relationships of PepF with enzymes of the pyruvate cycle, heterofermentation seems to be preferred in the deletion mutants. In the NZ9000-PSEC1 strain, acetate is produced at the expense of lactate. In the IL1403 strain it seems to be the production of acetolactate that is preferred, whether any more acetate nor formate is detected. The alteration of the pyruvate metabolism can explain the presence of oxidoreductases in the possible partners of PepF that could be responsible for rebalancing the redox potential due to a fermentation modification. We did not observe differences in cell division between wild type and *pepF* mutant observed by electronic microscopy. It is possible that the morphology of the mutant is not affected and that a technique other than the observation of dividing cells would reveal differences between wild type and *pepF* mutant. It is not surprising to find some of the proteins participating in cell division among our possible relationships of PepF. The cell division pro teins (FtsW, FtsK, FtsA) of lactococci are homologous to *B. subtilis* sporulation proteins and in *B. subtilis* the overproduction of PepF inhibits sporulation apparently due to its possible participation in the maturation of a signaling peptide (Kanamaru *et al.*, 2002). Recently Kavanaugh *et al.* (2007) demonstrated the involvement of an analogue of SipL in *Staphylococcus aureus* in the maturation of an autoinducing peptide implicated in quorum sensing the precursor of which is a signal peptide. Taking into account the strong re-

lationship predicted between PepF and SipL, the involvement of SipL in quorum sensing in *S. aureus* and the fact that sporulation is affected by PepF (Kanamaru *et al.*, 2002), we cannot exclude that the peptides maturated by PepF serve as extracellular signals (quorum sensing) or as intracellular signals (gene regulation) which could explain its indirect participation in different cellular functions.

We have shown that a global statistical analysis, with the flexibility of the KCCA, can be used to predict the role of a protein by inferring possible relationships with other proteins. This approach allows one to pose a hypothesis about the role of a single protein using existing data in order to guide the laboratory work. At the same time it enables one to find global relationships and pleiotropic roles that would not have been detected with other approaches. The experimental validations allowed us to confirm predicted roles and give biological sense to the predicted relationships. Nevertheless, even if a global approach was used, the complexity of biological systems impedes the complete elucidation of the role of a single protein. We increased our knowledge of PepF and confirmed its participation in protein secretion, but the precise mechanisms by which it interferes and other cellular functions we studied remains unclear. This situation encourages the development of complex and at the same time precise biological models that take into account pleiotropy and connectedness of all cellular functions.

Supporting information is available in an EPAPS document.

ACKNOWLEDGMENTS

This work is the result of collaboration between the Microbiology and the Applied Mathematics divisions of INRA and was financed by this institution. It was also financially supported by the Ile de France regional council, especially for the LC-MS-MS experiment. We thank Jean-Philippe Vert for providing the programs which made possible the combination of all our KEGG data into a sole graph. At INRA in Jouy en Josas we would like to thank: Gaëlle Bergot at the Unité de Biochimie Bactérienne for the measurements of metabolites of the pyruvate cycle and the group of Marie Pierre Chapot-Chartier for sharing with us their protocol of peptidoglycan analysis; at the Unité de Génetique Microbienne, we thank Marion Velten, Sophie Cheruel, and Patrice Polard for their assistance in the genetic construction for the heterologous production of PepF and the material they provided us with and Eric Guedon for sharing with us his microarray data; at the Unité d'Ecologie et de Physiologie du Système Digestif, we thank Philippe Langella for his advice and for providing us with the nuclease overproducing strain. We would also like to thank Alain Guillot at the Proteomic Platform (PAPSS) and Sophie Chat at the Microscopic Platform (MIMA2). Finally, we would like to thank Mireille Yvon, Kiên Kiêu, and Donald White for the revision of the manuscript and their helpful comments.

REFERENCES

Aerts, S, Lambrechts, D, Maity, S, Van, Loo P, Coessens, B, De Smet, F, Tranchevent, LC, De Moor, B, Marynen, P, Hassan, B, Carmeliet, P, and Moreau, Y (2006). "Gene prioritization through genomic data fusion." *Nat. Biotechnol.* **34**, 537–544.

Akesson, M, Förster, J, and Nielsen, J (2004). "Integration of gene expression data into genome-scale metabolic models." *Metab. Eng.* **6**, 285–293.

Atrih, A, Bacher, G, Allmaier, G, Williamson, MP, and Foster, SJ (1999). "Analysis of peptidoglycan structure from vegetative cells of *Bacillis subtilis* 168 and Role of PBR 5 in Peptidoglycan Maturation." *J. Bacteriol.* **181**, 3956–3966.

Bendtsen, JD, Nielsen, H, von Heijne, G, and Brunak, S (2004) "Improved prediction of signal peptides: SignalP 3.0." *J. Mol. Biol.* **340**, 783–795.

Blattner, FR, Plunkett, G, Bloch, CA, Perna, NT, Burland, V., Riley, M, Collado-Vides, J, Glasner, JD, Rode, CK, Mayhew, GF, Gregor, J, Davis, NW, Kirkpatrick, HA, Goeden, MA, Rose, DJ, Mau, B, and Shao, Y (1997) "The complete genome sequence of *Escherichia coli* K-12." *Science* **277**, 1453–1474.

Bolotin, A, Wincker, P, Mauger, S, Jaillon, O, Malarme, K, Weissenbach, J, Ehrlich, SD, and Sorokin, A (2001). "The complete genome sequence of the lactic acid bacterium *Lactococcus lactis*." *Genome Res.* **11**, 731–753.

Borgwardt, KM, Gretton, A, Rasch, MJ, Kriegel, HP, Schölkopf, B, and Smola, AJ. (2006). "Integrating structured biological data by Kernel Maximum Mean Discrepancy." *Bioinformatics* **22**, 49–57.

Chatel, JM, Langella, P, Adel-Patient, K, Commissaire, J, Wal, JM, and Corthier, G (2001). "Induction of mucosal immune response after intranasal or oral inoculation of mice with *Lactococcus lactis* producing bovine beta-lactoglobulin." *Clin. Diagn. Lab Immunol.* **8**, 545–551.

Chen, L, Tai, PC, Briggs, MS, and Gierasch, LM (1987). "Protein translocation into *Escherichia coli* membrane vesicles is inhibited by functional synthetic signal peptides." *Biol. Chem.* **262**, 1427–1429.

Chich, JF, David, O, Villers, F, Schaeffer, B, Lutomsli, D, and Huet, S (2007). "Statistics for proteomics: experimental design and 2-DE differential analysis." *J. Chromatogr., B: Biomed. Appl.* **849**, 261–272.

Cocaign-Bousquet, M, Garrigues, C, Novak, L, Lindley, ND, and Loubiere, P (1995). "Rational development of a simple synthetic medium for the sustained growth of *Lactococcus lactis*." *J. Appl. Bacteriol.* **79**, 108–116.

Conlin, CA, Trun, NJ, Silhavy, TJ, and Miller, CG (1992). "*Escherichia coli* prlC Encodes an Endopeptidase and is homologous to the *Salmonella typhimurium* opdA gene." *J. Bacteriol.* **174**, 5881–5887.

Covert, MW, Knight, EM, Reed, JL, Herrgard, MJ, and Palsson, BO (2004). "Integrating high-throughput and computational data elucidates bacterial networks." *Nature (London)* **429**, 92–96.

Dev, IK, and Ray, PH (1990). "Signal peptidases and signal peptide hydrolases." *J. Bioenerg. Biomembr.* **22**, 271–290.

Emr, SD, and Bassford, PJ (1982). "Localization and processing of outer membrane and periplasmic proteins in *Escherichia coli* strains harboring export-specific suppressor mutations." *Biol. Chem.* **257**, 5852–5860.

Emr, SD, and Silhavy, TJ (1980). "Mutations affecting localization of *Escherichia coli* outer membrane protein, the bacteriophage λ receptor." *J. Mol. Biol.* **141**, 63–90.

EPAPS Document No. E-HJFOA5-2-002801 for supplemental material. This document can be reached through a direct link in the online article's HTML reference section via the EPAPS homepage (http://www.aip.org/pubservs/epaps.html).

Gitton, C, Meyrand, M, Wang, J, Caron, C, Trubuil, A, Guillot, A, and Mistou, MY (2005). "Proteomic signature of *Lactococcus lactis* NCDO763 cultivated in milk." *Appl. Environ. Microbiol.* **71**, 7152–7163.

Glauner, B (1988). "Separation and quantification of muropeptides with high-performance liquid chromatography." *Anal. Biochem.* **172**, 451–464.

Guedon, E, Serror, P, Ehrlich, SD, Renault, P, and Delorme, C (2001). "Pleiotropic transcriptional repressor CodY senses the intracellular pool of branched-chain amino acids in *Lactococcus*

lactis." *Med. Mundi* **40**, 1227–1239.
Guillot, A, Gitton, C, Anglade, P, and Mistou, MY (2003). "Proteomic analysis of *Lactococcus lactis*, a lactic acid bacterium." *Proteomics* **3**, 337–354.
Gyaneshwar, P, Paliy, O, McAuliffe, J, Popham, DL, Jordan, MI, and Kustu, S (2005). "Sulphur and nitrogen limitation in *Escherichia coli* K12, specific homeostatic responses." *J. Bacteriol.* **187**, 1074–1090.
Ho, Y, Gruhler, A, Heilbut, A, Bader, GD, Moore, L, Adams, SL, Millar, A, Taylor, P, Bennett, K, and Boutilier, K (2002). "Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry." *Nature (London)* **415**, 180–183.
Hoffman, CS, and Winston, F (1987). "A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of *Escherichia coli*." *Gene* **57**, 267–272.
Ichihara, S, Beppu, N, and Mizushima, S (1984). "Protease IV, a cytoplasmic membrane protein of *Escherichia coli*, has signal peptide peptidase activity." *Biol. Chem.* **259**, 9853–9857.
Ito, T, Chiba, T, Ozawa, R, Yoshida, M, Hattori, M, and Sakaki, Y (2001). "A comprehensive two-hybrid analysis to explore the yeast protein interactome." *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4569–4574.
Kanamaru, K, Stephenson, S, and Perego, M (2002). "Overexpression of the PepF oligopeptidase inhibits sporulation initiation in *Bacillus subtilis*." *J. Bacteriol.* **184**, 43–50.
Kanehisa, M, Goto, S, Kawashima, S, and Nakaya, A (2002). "The KEGG databases at GenomeNet." *Nucleic Acids Res.* **30**, 42–46.
Kondor, RI, and Lafferty, J (2002). "Diffusion kernels on graphs and other discrete input spaces." International Conference for Machine Learning pp. 315–322.
Kavanaugh, JS, Thoendel, M, and Horswill, AR (2007). "A role for type I signal peptidase in *Staphylococcus aureus* quorum sensing." *Mol. Microbiol.* **65**, 780–798.
de Magalhães, JP, Costa, J, and Toussaint, O (2005). "HAGR: the Human Aging Genomic Resources." *Nucleic Acids Res.* **33**, Database issue D537-D543.
Monnet, V, Nardi, M, Chopin, A, Chopin, MC, and Gripon, JC (1994). "Biochemical and genetic characterization of PepF an oligoendopeptidase from *Lactococcus lactis*." *J. Biol. Chem.* **269**, 32070–32076.
Nardi, M, Renault, P, and Monnet, V (1997). "Duplication of the PepF gene and shuffling of DNA fragments on the lactose plasmid of *Lactococcus lactis*." *J. Bacteriol.* **179**, 4164–4171.
Noirot, P, and Noirot-Gros, MF (2004). "Protein interaction networks in bacteria." *Curr. Opin. Microbiol.* **7**, 505–512.
Novak, P, Ray, PH, and Dev, IK (1982). "Localization and purification of two enzymes from *Escherichia coli* capable of hydrolyzing a signal peptide." *Biol. Chem.* **261**, 420–427.
Piuri, M, Sanchez-Rivas, C, and Ruzal, SM (2005). "Cell wall modifications during osmotic stress in *Lactobacillus casei*." *J. Appl. Microbiol.* **98**, 84–95.
Qi, Y, Klein-Seetharam, J, and Bar-Joseph, Z (2005). "Random forest similarity for protein-protein interaction prediction from multiple sources." Biocomputing: Proc. Pacific Symposium 10, Hawaii.
Renault, P, Corthier, G, Goupil, N, Delorme, C, and Ehrlich, SD (1996). "Plasmid vectors for Gram-positive bacteria switching from high to low copy number." *Gene* **183**, 175–182.
de Ruyter, PG. G. A, Kuipers, OP, Beerthuyzen, MM., van Alen-Boerrigter, I, and de Vos, WM (1996). "Functional analysis of promoters in the nisin gene cluster of *Lactococcus lactis*." *J. Bacteriol.* **178**, 3434–3439.
Sambrook, J, and Russel, DW (2001). Molecular cloning: a laboratory manual, Cold Spring Harbor Laboratory Press.
Schölkopf, B, Tsuda, K, and Vert, JP (2004). Kernel Methods in Computational Biology, MIT Press, Cambridge, MA.
Shin, JJ, Bryksin, AV, Godfrey, HP, and Cabello, FC (2004). "Localization of BmpA on the exposed outer membrane of *Borrelia burgdorferi* by monospecific anti-recombinant BmpA rabbit antibodies." *Infect. Immun.* **72**, 2280–2287.
Siezen, RJ, Renckens, B, van Swam, I, Peters, S., van Kranenburg, R, and de Vos, WM (2005). "Complete sequence of four plasmids of *Lactococcus lactis* subsp cremoris SK11 reveal extensive adaptation to the dairy environment." *Appl. Environ. Microbiol.* **71**, 8371–8382.
Tisljar, U, Knight, CG, and Barrett, AJ (1990). "An alternative quenched flourescence substrate for Pz-peptidase." *Anal. Biochem.* **186**, 112–115.
Tjalsma, H, Bolhius, A, Jongbloed, JDH, Bron, S, and van Dijl, JM (2000). "Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the secretome." *Microbiol. Mol. Biol. Rev.* **64**, 515–547.
Werhli, A, and Husmeier, D (2007). "Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge." *Stat. Appl. Genetics Mol. Biol.* 6, Issue 1, Article 15.
Wicker, W, Moore, K, Dibb, N, Geissert, D, and Rice, M (1987). "Inhibition of purified *Escherichia coli* leader peptidase by the leader (signal) peptide of bacteriophage M13 procoat." *J. Bacteriol.* **169**, 3821–3822.
Yamanishi, Y, Vert, JP, Nakaya, A, and Kaneisha, M (2003). "Extraction of correlated clusters from multiple genomic data by generalized kernel canonical correlation analysis." *Bioinformatics* **19** Suppl. 1, i323–i330.
Yamanishi, Y, Vert, JP, and Kanehisa, M (2004). "Protein network inference from multiple genomic data: a supervised approach." *Bioinformatics* **20**, Suppl. 1, i363–i370.
Yamanishi, Y, and Vert, JP (2007). "Kernel Matrix Regression, *Cornell University Library*." (submitted: http://arxivorg/abs/q-bio/0702054v1).

4.2 L'export de protéines est affecté lors d'une surproduction des protéines exportées

Dans l'article I, nous avons observé que, dans des conditions de surproduction d'une protéine sécrétée, en l'occurrence NucB, l'absence de PepF limitait la sécrétion de la protéine majoritaire du sécrétome de *L. lactis*, Usp45 et de NucB elle-même. Nous avons complété ces données en évaluant l'effet sur la sécrétion de Usp45 lors de la surproduction d'une protéine exportée substrat de la sortase, ainsi que l'effet sur protéine exportée et ancré à la membrane, une lipoprotéine (OptA lors de la surproduction d'une protéine sécrétée.

4.2.1 Analyse d'une lipoprotéine lors de la surproduction d'une protéine sécrétée (NucB)

Nous résultats jusqu'a présent indiquent que la quantité des protéines exportées est affectée dans un mutant PepF lors de la surproduction de NucB. Nous avons également étudié la présence d'une lipoprotéine, exporté et attaché à la membrane cellulaire. Pour cela nous avons réalisé un Western Blot avec des anticorps contre la lipoprotéine OptA (Figure 4-1).

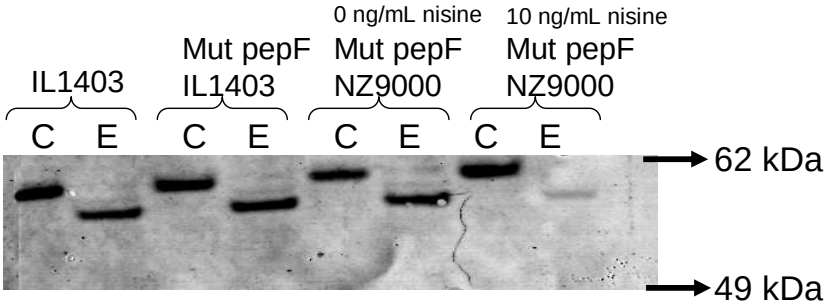


Figure 4-1 : Détection de OptA par Western Blot chez les souches *L. lactis* IL1403 et NZ9000.

4.2.2 Analyse des protéines sécrétées (Usp45) lors de la surproduction d'un substrat de la sortase

La protéine surproduite est codé par un gène contentant le peptide signal de la protéine Usp45 de *L . lactis*, la protéine NucA de *S. aureus* et l'ancre pour être attacher à la paroi de la protéine M de *S. pyogenes.* Le gène est sous contrôle du promoteur nisine et peut être induit en présence de nisine, compte tenu que les gènes *nisRK* sont présents dans le chromosome de la souche NZ9000. Il s'agit du même système d'induction utilisé pour la surproduction de NucB décrit dans l'article I. La seule différence est qu'il s'agit d'une protéine ancrée dans la membrane et non pas libérée dans le milieu extracellulaire (Figure 4-3).

Les trois fractions cellulaires, cytoplasmique (C), d'enveloppe (M) et le surnageant contenant les protéines sécrétées (S) ont été analysées par SDS-Page après une culture en présence de 5ng/mL de nisine dans le but de comparer les profils entre le mutant *pepF* et le sauvage (Figure 4-2).

Le gel montre qu'il y a une différence entre le sauvage et le mutant quand à la quantité de Usp45, la protéine majoritairement sécrété par le lactocoque. En effet, chez le mutant, moins de Usp45 est détecté dans le surnageant.

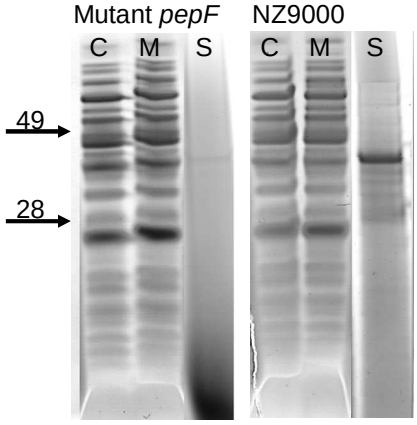


Figure 4-2 : Gel SDS-Page des trois fractions cellulaires cytoplasmique (C), d'enveloppe (M) et surnageant (S) d'une culture de NZ9000 et son mutant *pepF* contenant pVE5547.

Nous avions montré l'accumulation de morceaux du peptide signal surproduit dans un mutant *pepF* (NZ9000-pSec1), où la protéine sécrétée NucB est surproduite, il serait intéressant de détecter le peptide signal et le peptide C-terminal contenant l'ancre dans les souches NZ9000-pVE5547 sauvage et mutant *pepF*. Les peptides recherchés seraient le peptide signal provenant de Usp45 : MKKIISAILMSTVVLSAAAPLSGVYA et le peptide C-terminal contenant l'ancre de la protéine M : GETANPFFTAAALTVMATAGVAAVVKRKEEN (Figure 4-3). Nous avons rencontré des difficultés techniques au cours de cette expérience. L'échantillon préparé comme pour la détection du peptide signal chez le mutant PepF ne contenait pas assez de matériel. Au contraire, un échantillon préparé à partir du cytoplasme sans passer par le gel, même ultrafiltré en ne gardant que la fraction inférieur à de 3kDA, s'est avéré très complexe et difficile à analyser. Il serait alors nécessaire d'inclure d'autres étapes de purification.

Avec ces résultats nous pouvons dire que l'export des protéines ancrées sur la membrane sont affectées dans un mutant PepF de la même manière que les protéines sécrétées comme NucB. Ce que nous ne savons pas est si uniquement le peptide signal ou aussi le peptide correspondant à l'ancre sont hydrolysés par PepF.

En conclusione, ces résultats montrent que les trois types de protéines (sécrétées, ancrées sur la paroi et lipoprotéines) sont exportées par la même voie et que son fonctionnement est perturbé en absence de PepF.

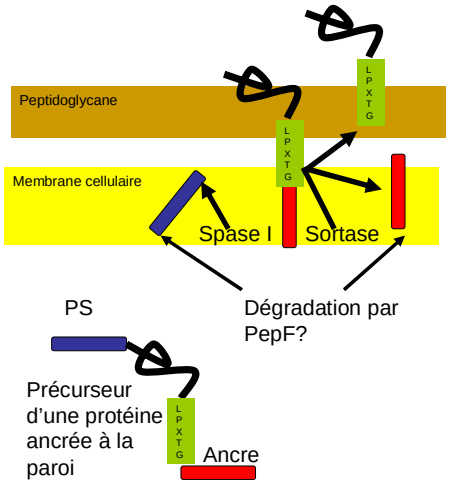


Figure 4-3 : Schéma des peptides libérés lors de la maturation et l'ancrage des protéines substrats de la sortase qui illustre les peptides possiblement dégradés par PepF

Avec ces résultats nous pouvons dire que l'export des protéines ancrées à la paroi par la sortase est affecté dans un mutant PepF de la même manière que les protéines sécrétées comme NucB. Ce que nous ne savons pas est si uniquement le peptide signal ou aussi le peptide correspondant à l'ancre, qui possède de caractéristiques similaires aux peptides signaux, sont hydrolysés par PepF.

Nous avons observé qu'en absence de PepF et lors de la surproduction d'une protéine sécrétée ou ancrée à la paroi deux types de protéines n'arrivent pas à leur localisation habituelle à l'extérieur de la membrane, une protéine sécrétée possédant le peptide signal de Usp45 et une lipoprotéine, OptA. Compte tenu de la présence des morceaux du peptide signal de Usp45 chez le mutant *pepF* (et leur absence chez la souche sauvage), ainsi que la localisation de PepF en périphérie de la cellule, nous postulons que PepF est nécessaire pour un fonctionnement correct de la machinerie de sécrétion qui, en son absence, pourrait être encombrer par le peptide signal non dégradé.

D'autres explications alternatives existent. Il est possible que en absence de PepF les deux types de protéines exportées soient dégradés, possiblement par la dérégulation d'une protéase membranaire (inconnue) causée par l'absence d'un peptide produit par PepF. Il est d'ailleurs possible que la modification du cycle du pyruvate, notamment, la diminution en production de lactate, provoque indirectement un disfonctionnement de la machinerie de sécrétion par un possible manque d'ATP.

5 Résultats des tests expérimentaux des liens prédits pour YvjB (Eep)

Les résultats des tests expérimentaux des liens prédits de YvjB indiquent sa participation dans la synthèse et l'export des protéines prises en charge par la particule de reconnaissance du peptide signal SRP. Le lien le plus fort est prédit avec Ffh la protéine qui compose la SRP. De plus, la grande quantité de protéines ribosomales et de synthèse de protéines retrouvés parmi ces liens renforce le lien avec SRP qui est impliquée dans sécrétion co-traductionnelle.

Parmi les protéines qui peuvent être prise en charge par SRP, nous avons favorisé les lipoprotéines comme cible de YvjB à cause de la présence de la signal peptidase II et de la prolipoprotéine diacylglycerol transférase, ainsi que les connaissance sur Eep chez *E. faecalis*, qui suggèrent que les précurseurs des peptides produites par Eep sont les peptides signaux des lipoprotéines.

Pour tester l'hypothèse que YvjB participe dans la maturation et sécrétion des lipoprotéines nous avons construit un mutant de délétion du gène *yvjB*. Nous avons étudié la fonctionnalité, localisation et maturation des lipoprotéines. Les résultats de ces expériences sont le sujet de l'article II.

5.1 Article II : YvjB, an Eep-like Protein, is an essential partner of the signal peptidase II in the maturation of certain lipoproteins in *Lactococcus lactis*

Article soumis à la revue *Journal of Bacteriology*

YVJB, AN EEP-LIKE PROTEIN, IS AN ESSENTIAL PARTNER OF THE SIGNAL PEPTIDASE II IN THE MATURATION OF SOME LIPOPROTEINS IN *Lactococcus lactis*

Running title: YvjB, implicated in *L. lactis* lipoprotein maturation

Liliana Lopez Kleine[1,2] , Alain Guillot[1§], Alain Trubuil[2], Véronique Monnet[1*]

[1] INRA, Unité de Biochimie Bactérienne, [§]Plate-Forme d'Analyse Protéomique de Paris Sud-Ouest, UR477,
F-78350 Jouy-en-Josas, France.
[2] INRA, Unité de Mathématiques et Informatique Appliquées, UR341. F-78350 Jouy-en-Josas, France.

*Correspondent footnote: Unité de Biochimie Bactérienne – Domaine de Vilvert - F-78350 Jouy-en-Josas, France, Telephone: 33 (0)134652149, Fax: 33 (0)134652163,
veronique.monnet@jouy.inra.fr

Although proteolysis is involved in numerous biological processes, the precise role of different proteolytic enzymes often remains unknown. Proteolysis has been studied in *Lactococcus lactis* because of its crucial role during growth in milk. In addition to this role, and despite a significant number of proteins possessing conserved proteolytic domains, little is known about other roles for proteolysis in this bacterium. Using a statistical approach, coupled to an experimental validation, we predicted relationships for the putative membrane-embedded protease YvjB and validated them using wet-lab experiments. The predicted relationships for YvjB concerned proteins involved in protein export and protein synthesis, as well as lipoprotein maturation. In the light of these predictions, we hypothesized its participation in the process of lipoprotein co-translational secretion and maturation. By analyzing an YvjB knock-out mutant, we confirmed its participation in this process through the recognition of some lipoproteins mediated by the hydrophobicity of the signal peptide. YvjB is also necessary for the cleavage of the signal peptide of at least one lipoprotein.

Introduction

Proteolysis is of major importance to various bacterial functions. Among the proteases known to play key roles are those secreted by pathogenic bacteria which interact with the defense system of the host or directly attack host tissues and constitute important virulence factors (30). Another important role of proteolysis is in protein turnover, which has been very well characterized in *Escherichia coli* (18). In *Lactococcus lactis*, proteolysis has been studied extensively because of its crucial role during growth in milk, a very important characteristic of this bacterium used as starter dairy industry (15, 29). Several research groups have participated in characterizing the complex proteolytic system involved in nitrogen nutrition in *L. lactis* (24, 32, 39). Nevertheless, little is known about other roles of proteolysis in this bacterium. Apart from the HtrA (33) or Clp (42) proteases which degrade unfolded or damaged proteins, or DacA, DacB (13) and YjgB (35) which participate in peptidoglycan maturation and ComC and PepF potentially involved in competence (44) and protein secretion (28), respectively, little is known of the precise role of other proteolytic enzymes of *L. lactis*. The role of these proteins is still difficult to investigate via targeted experiments without any clue. We applied an approach in which the significant amount of post-genomic data available for this bacterium is used in order to infer the role of proteins. We recently applied this approach to predict potential partners of the oligoendopeptidase F (PepF)

(45). The predictions guided laboratory experiments, and allowed us to successfully determine the implication of PepF in protein secretion and other pleiotropic roles (28). We here applied this approach to another putative proteolytic enzyme, YvjB. Taking into account the presence of two conserved proteolytic domains, HEXXH and NPDG, and their localization within the first and third predicted transmembrane domains, YvjB likely belongs to a known family of membrane embedded proteases, the S2P family (37). Some members of this family also have a PDZ domain (23), as it is the case for YvjB. This family includes other members: YaeL (1, 5) in *Escherichia coli*, YluC (7) and SpoIVFB (37) in *Bacillus subtilis* and Eep in *Enterococcus faecalis* (3, 4). Considering these similarities, we were interested in YvjB because of its possible role in regulatory processes and cell-cell communication.

Statistical inference predicted a strong relationship between YvjB and proteins participating in protein secretion, protein synthesis and lipoprotein maturation. These predictions allowed us to hypothesize the participation of YvjB in the co-translation secretion of lipoproteins and their maturation. By studying the oligopeptide binding protein OptA with respect to its localization, maturation and functionality, and the cellular localization of all lipoproteins, we were able to reveal the participation of YvjB in the recognition and maturation of a group of lipoproteins mediated by the hydrophobicity of the signal peptide. Furthermore, we observed that YvjB is necessary to the cleavage of the signal peptide of at least one lipoprotein.

Materials and Methods

Predictions

Predictions concerning the roles of proteins were obtained using a Kernel Canonical Correlation Analysis (KCCA) (45), as described previously (28). This method enables the integration of different data sources: microarrays, phylogenetic profiles, chromosome gene distances and 2D-gels, as well as the KEGG protein metabolic network (22). It is a supervised method that uses the KEGG protein metabolic network as golden standard (representing the known relationships between proteins of *L. lactis*). The aim of our work was to expand this network adding proteins of unknown function to it, predicting relationships between proteins of unknown function and proteins already on the metabolic network, but also between the new added proteins themselves. These predicted relationships are distances between all genes, calculated on the basis of four data sources (microarrays, phylogenetic profiles, chromosome gene distances and 2D-gels), the knowledge on the protein metabolic network is included by correlation. A threshold is chosen (the highest distance between two proteins, known to be neighbors on the protein metabolic network) to define possible partners of the protein of interest. In conclusion, the role predictions are given by the distances between proteins of interest and other proteins in the organism, based on all data sources at the same time (28, 45). This list of proteins was then used to define functional categories. These categories are based on known and potential functions attributed to the potentially related proteins. We then used this classification to pose a hypothesis concerning the role of YvjB, based on the premise that YvjB and its predicted partners participate in the same functions.

Experimental protocols

Strains and plasmids

The bacterial strains and plasmids used during this study are listed in Table 1 and Table 2. *L. lactis* strains were grown at 30°C in M17 (Difco) medium supplemented with 5% glucose. The chemically-defined medium (CM) (31) was prepared containing all amino acids or omitting the essential amino acid leucine which was added in the form of a peptide (Leu-enkephalin: Tyr-Gly-Gly-Phr-Leu, BACHEM) to obtain the same quantity of leucine as in the CM. When necessary,

erythromycin was added to the medium at a concentration of 5 µg ml^{-1} for *L. lactis*.

Molecular cloning techniques
Molecular cloning techniques were performed using standard procedures (38). Plasmids were extracted using a QIAprep Spin miniprep kit (Qiagen). Total *L. lactis* DNA was isolated as described previously (20). Restriction enzymes (New England Biolabs), T4 DNA ligase from the Fast-link ligation kit (Epicentre), *Taq* DNA polymerase MP (Qbiogene) and the TripleMaster PCR system (Eppendorf) were used according to the suppliers' recommendations. PCRs were run using a Mastercycler gradient thermal cycler (Eppendorf). Annealing was performed between 52°C and 67°C, depending on the primers used. All constructions were verified by sequencing with an Applied Biosystems 310 automated DNA sequencer, using the ABI PRISM Dye Terminator Cycle Sequencing Kit (Perkin Elmer). Primers for *L. lactis* were selected on IL1403 (6) genome sequences. The oligonucleotides were purchased from Eurogentec and are listed in Table 3.

Construction of negative mutants
The pGhost9-Δ*yvjB* and pGhost9-Δ*lspA* plasmids were constructed by ligating PCR fragments of around 1000 b.p. located upstream and downstream of the genes to be deleted. The PCR fragments were obtained using the eepUF-eepUR or lspAUF-lspAUR primers, respectively, and eepDF-eepDR or lspALF-lspALR, respectively, in a pGhost9 plasmid after digestion of the plasmid by the restriction enzymes *Xho*I and *Spe*I for YvjB and *Xho*I and *Sma*I for LspA. PCR fragments were digested prior to ligation by *Xho*I and *Eco*RI for eepUF-eepUR; *Xho*I and *Bam*HI for lspAUF-lspAUR; *Spe*I and *Eco*RI for eepDF-eepDR; *Bam*HI and *Sma*I for lspALF-lspALR. The PCRs used to screen clones and for sequencing were performed with eepF and eepR or lspAF and lspAR that had been designed to amplify approximately 300 b.p. and 500 b.p. when the gene was deleted. The *yvjB* gene was completely deleted, including the intergenic region between *cdsA* and *yvjB* preceding the *yvjB* gene. In the *lspA* knock-out strain, 105 b.p. at the beginning of the sequence were present and a stop codon was added to be sure that transcription stopped after this 105 b.p. For the complementation of *yvjB,* three genes of the operon (*upps-cdsA-yvjB*) and the intergenic sequence before *upps* were amplified, and the PCR product ligated into pILN13 (36) low copy number after *Pst*I and *Not*I digestion.

Construction of the strain with replacement of the OptA signal peptide with that of YvdF
In order to construct pGhost9-SP_{YvdF}OptA, the sequence coding for the signal peptide of *optA* was replaced by successive PCR reactions in which fragments of the sequence to be replaced were contained in the primers. These overhanging parts of the primers were aligned at the same temperature as the primers used for the PCR that followed. In total, six PCR reactions were performed: i) PCR1 and PCR2, followed by PCR3 to insert the upper sequence of the *yvdF* signal peptide (UyvdF) and ii) PCR4 and PCR5, followed by PCR6 to insert the lower sequence (LyvdF) (Figure A1, A for appendix). The final PCR was cloned into the pGEMT plasmid. After digestion with *Spe*I and *Apa*I, the fragment was liberated, purified and ligated into the pGhost9 plasmid.

For both types of construction described above, we obtained the mutants electroporating the plasmids into *L. lactis* IL1403 competent cells and replacing the wild type with a deleted gene (knock-out mutants *yvjB* and *lspA*) or with the *optA* gene containing the signal peptide of *yvdF* (SP_{YvdF}OptA mutant),. Mutant strains were selected first of all with respect to their resistance to erythromycin and secondly on the size of the fragment amplified by PCR. Integration into the chromosome and subsequent excision were achieved using the thermosensitivity of the plasmid.

For the SP_{YvdF}OptA mutant, the PCR screening of clones could not be done with respect of the length of the amplified fragment, as the wild type and the mutant have genes of the same length. The screening was performed with the LyvdFi-UyvdFi and UoptSF primers using as criteria the absence of amplification in the wild-type strain (Figure A1).

Cell envelope and supernatant fraction preparation

To prepare the protein extract, cultures were harvested at OD_{600nm}= 0.6 by centrifugation and washed with phosphate buffer 0.2 M at pH 6.4. Cell lysis was achieved with a cell disruptor (Constant System Ltd.). The cytoplasm and cell envelope fractions were separated by ultracentrifugation at 4°C and 50,000 rpm for 20 min (Centrikon T-1080, Kontron intruments). The proteins secreted were prepared from 5 ml supernatant (after the centrifugation of 6 mL), precipitated with TCA (20% final concentration), incubated for 30 minutes at 4°C and centrifuged for a further 30 minutes at 4°C. Once the supernatants had been eliminated, the pellet was washed with cold acetone, air-dried and resuspended in 500µL NaOH 50 mM.

1D-SDS Electrophoresis – LC-MS-MS

Protein fractions were separated using one-dimension sodium dodecyl sulfate-polyacrylamide (4-12%) gel electrophoresis (SDS-PAGE) using the NuPage system (Invitrogen). For the MS-MS analysis of lipoproteins in the envelope fraction (carried out by the Plateforme d'Analyse Protéomique de Paris Sud-Ouest), three independent cultures and extractions were processed for each strain. The gel was cut in the mass range where lactococcal lipoproteins were predicted to be, based on the analysis of the IL1403 genome (43), i. e. between 14 and 62 kDa, resulting in nine samples per lane for analysis. After trypsin digestion (overnight at 37°C) with 100 ng per sample of the gel fraction, peptides were obtained from these fragments through three subsequent washes with ACN 50% followed by one wash with TFA 0.1%. The peptides thus obtained were pooled and dried in a Speed-Vac concentrator for 1h, then resolubilized in 25 µl of HPLC loading buffer (0.08% TFA and 2% ACN) and analyzed by LC-MS-MS on LTQ-Orbitrap Discovery (Thermo Fisher, San Jose, USA). The peptide mixtures (4 µl) were injected onto the PepMap C18 pre-column (300 µm ID x 5 mm, 100 Å, LC Packings, Amsterdam, Netherlands)) at a flow rate of 20 µl/min to remove salts. The peptides were analyzed in a 50 min gradient of 2-80% acetonitrile in water containing 0.1% formic acid. A flow rate of 300nL/min was used to elute peptides from the C-18 PepMap100 reversed-phase nanocolumn (75 µm ID x 15cm, 3 µm, 100 Å) to a picoTip nanospray needle (360 OD x 20 µm, 10 µm ID) (New Objective, USA) for ionization and peptide fragmentation on an ion trap mass spectrometer.

The LTQ-Orbitrap mass spectrometer was operated using Xcalibur 2.07 with the Nth order double play parallel method: one MS scan event on the Orbitrap analyser for a maximum of four MS/MS fragmentation events on the LTQ analyser only on peptides charged two or three times. Intact peptides were detected in the Orbitrap at 7500 resolution on centroid mode on the 300-1600 m/z range. Internal calibration was performed using the ion signals of 391.2843 and 536.1656 as lock mass. Maximal ion accumulation time allowed on the Orbitrap was 100 msec. 3.10^4 ions were accumulated in the ion trap for generation of CID spectra. CID normalized collision energy was set to 40%.

Database searching and quantization

All protein identifications were performed with Bioworks 3.3.1 SP1 (Thermo Fisher, USA). The raw data were converted and filtered to create a peak list using the default data generation parameter for LTQ-Orbitrap mass spectrometer. All peak lists of precursors and ion fragments were matched automatically against 32 lipoproteins of the *L. lactis* IL1403 genome. The Bioworks search parameters were: trypsin specificity with one missed cleavage, methionine

oxidation; mass tolerance fixed to 20 ppm for the precursor ion and to 0.5 Da for fragment ions. The search result was filtered with a multiple threshold filter applied at the peptide and protein levels consisting of the following criteria: Xcorr magnitude up to 2.2 and 3.0 for di- and tri-charged peptides, respectively; peptide probability lower than 0.01; ΔCn greater than 0.1; peptide mass tolerance lower than 10 ppm; only the first match result for each identified peptide. We retained only proteins identified by two different peptides in two out of three independent repetitions. The results of the nine analyses for each sample were merged with the multiconsensus function of Bioworks 3.3.1SP1. The false discovery rate for protein identification after merging was estimated to be <0.7% with a complete reverse IL1403 database.
The quantification was carried out by spectral counting approach. We used the average number of scan events that identified a protein as quantitative parameter. We classified proteins into two different groups when the difference in quantity was higher than 25% and concerned more than three spectra.

Western blots
Western blots were performed to detect OptA with antibodies (25). OptA antibodies (dilution 1/1000, 2 hours) were incubated after the electrophoretic transfer of proteins from the polyacrylamide gel to a nitrocellulose membrane (Trans-Blot, Biorad, 0.45µm). Horseradish peroxidase-coupled anti-rabbit antibodies were used as a second antibody (dilution 1/3000, 1 hour) and revealed with an ECL kit (Plus Western Blotting Detection System, Amersham Biosciences).

Results
Predicted links for YvjB and its possible role in Lactococcus lactis
The KCCA predicted 78 proteins to be in close relationship with YvjB (Figure 1 and Table A1 in the appendix for the complete list of predicted relationships). Most of these proteins are involved in protein synthesis. Eight of them are ribosomal proteins and eighteen are proteins participating in functions such as elongation (FusA), termination (NusG) or peptide release (PrfA) and which are all necessary to synthesis. The strongest relationship for YvjB was predicted to be Ffh, a protein which forms part of the signal recognition particle (SRP) responsible for conducting proteins endowed with a signal peptide to the Sec translocon. The receptor for SRP in the membrane, FtsY, as well as a member of the Sec translocon, SecY, are amongst the predicted partners of YvjB. Both ClpX and the transporter EcsAB, already described as regulators of protein secretion (including lipoproteins) in *Bacillus subtilis* (27, 34) were also found to be linked to YvjB. All these predicted relationships suggest a participation of YvjB in protein secretion. A link between lipoproteins and YvjB appeared through its relationship with signal peptidase II (LspA) which cleaves lipoprotein signal peptides and with the prolipoprotein diacylglycerol transferase (Lgt) which participates in the lipoprotein modification. Taking into account the most closely represented categories of proteins linked to YvjB and the strong relationship with Ffh, we hypothesized that YvjB participates in the co-translational secretion of lipoproteins which regroups the three principal categories in which the predicted partners of YvjB can be classified: protein secretion, protein synthesis and lipoprotein maturation (Figure 3). Our prediction indicates a possible participation of YvjB in translocation, localization and/or signal peptide cleavage (maturation).

Twenty-two of the predicted relationships had an unknown function. These included a cluster of proteins (YhhG, YlxQ, YhhC) encoded by neighboring genes. The first two are probably involved in transcription and translation, while the third is a putative integral membrane protein homolog of the LrgB of *Staphylococcus aureus*, implicated in the regulation of murein hydrolase

(21). More generally, the category of membrane proteins was well represented, with five proteins predicted to be linked with YvjB.

Experimental validation of predicted roles

We chose to work with the *yvjB* mutant in order to determine the effect of the absence of YvjB. The growth of this mutant was lower in M17-glu compared to the wild-type strain (Figure A2). We started analyzing the secretion of the naturally most abundant secreted protein in *L. lactis*, Usp45, and did not detect any difference in the amount of secreted Usp45 in the wild-type and the *yvjB* mutant (data not shown). Then, we focused on one lipoprotein, the oligopeptide binding protein OptA, whose functionality, can easily be demonstrated in experiments where an essential amino acid is only available in form of oligopeptide. If this protein is not functional, the strain is not able to transport oligopeptides (25). We observed that the *yvjB* mutant was not able to grow when leucine was only present in form of the pentapeptide leu-enkephalin and had the same phenotype as an *optA* mutant (Figure A3a). We checked that the complemented *yvjB* mutant retrieved the wild-type phenotype, i.e. its ability to transport leu-enkephalin (Figure A3b).

Once we knew that OptA was not functional in an *yvjB* mutant, we turned to its correct production and localization. We prepared supernatant, cytoplasmic and envelope protein extracts and detected OptA by Western blotting with the antibodies produced by Lamarque and colleagues (25). OptA was not detected in the envelope fraction of the *yvjB* mutant. We knew that *optA* was expressed in the *yvjB* mutant because we observed a normal, even increased quantity of OptA in the cytoplasm (Figure 3). The absence of OptA from the envelope explains why it was not functional in this strain (Figure 3). Consequently, two explanations remained possible to account for the absence of OptA in the envelope fraction of the *yvjB* mutant: either OptA was not exported or it was exported but rapidly degraded. Nevertheless, we did not detect any OptA degradation product by LC-MS-MS in the SDS PAGE. In contrast, we detected another lipoprotein, BmpA, in both, the mutant and the wild-type strain, in several bands between 14 and 39 kDa. We did not detect OptA or OptA fragments by Western blot in the supernatant either.

We decided to perform a global experiment to investigate the presence of all *L. lactis* lipoproteins in the cell envelope of the *yvjB* mutant using LC-MS-MS. We identified 15 lipoproteins on which the absence of YvjB exerted different effects. Three different behaviors were observed (Table 4):

I) A first group of lipoproteins whose abundance diminished in the envelope (or disappeared completely, e.g. OptA and YpcG).

II) A second group of lipoproteins whose abundance increased in the cell envelope.

III) A third group of lipoproteins which could not be classified in the other two groups because no significant difference in quantity was observed between the wild-type and the *yvjB* mutant.

Different characteristics, such as the mass, codon adaptation index, length and hydrophobicity (gravy index) of the signal peptide were assessed in order to discriminate between these two groups. Generally speaking, proteins with a mass higher than 35 kDa and a less hydrophobic signal peptide belonged to the first group (group I), and tended to disappear from the cell envelope in an *yvjB* mutant. The FrdC, OptA, and YpcG proteins, which undoubtedly form part of group I, had two common characteristics: a KXXL motif in position 6 of the signal peptide, and a hydrophobic amino acid in position -2 of the lipobox, contrasting with group II lipoproteins (YvdF, YjgC, FhuD, BmpA) which presented a hydrophilic amino acid in this position (for more details see table A2). At this point, we knew that YvjB participated in the recognition and correct localization of OptA and probably of other lipoproteins of group I.

To test the importance of the peptide signal, we decided to investigate the effect of replacing the OptA signal peptide by the YvdF signal peptide, the protein whose abundance increased most in

an *yvjB* mutant and had one of the most hydrophobic signal peptides. In the presence of YvjB (in the wild-type strain), we observed that replacement of the OptA signal peptide had no effect on its localization and maturation which remained correct. In the *yvjB* mutant, replacing the OptA signal peptide correctly restored the localization of OptA in the cell envelope (Figure 3B). Nevertheless, it was present in its precursor form, similar to that found in the cytoplasm. Interestingly, OptA was shown to be functional in this strain, as it was able to transport the pentapeptide leu-enkephalin. This observation indicates that the correct localization of OptA in the envelope in the *yvjB* mutant depends on the nature of the signal peptide. Furthermore, it showed that the correct maturation of the protein also depends on the presence of YvjB (Figure 3B).

Consequently, we wondered whether YvjB itself could release the signal peptide from OptA and act as a lipoprotein signal peptidase. We constructed a signal peptidase II (LspA) mutant and checked whether, in an *lspA* mutant, the OptA and SP_{YvdF}OptA signal peptides were cleaved. As this was not the case, we concluded that LspA and not YvjB cleaved the signal peptide of these proteins. We also observed that OptA was functional in both the *lspA* and SP_{YvdF}OptA mutants (data not shown). Construction of a double mutant (*lspA-yvjB*) allowed us to confirm that OptA was correctly localized in the membrane, but only in the presence of YvjB (Figure 4). These results confirmed that LspA cleaves the signal peptides of lipoproteins and that YvjB is needed for the correct maturation of SP_{YvdF}OptA.

Discussion

Predictions helped to pose a biological hypothesis

Predictions concerning the role of YvjB we made using the KCCA method led us to hypothesize the participation of YvjB in lipoprotein export. The strongest relationship, predicted to be with Ffh, gave some indications about its crucial role in the maturation of proteins carrying a signal peptide and secreted via the so called co-translational secretion pathway (40). Export occurs at the same time as proteins are synthesized, and is ensured by SRP, composed of Ffh and the small RNA subunit. The fact that several proteins participating in protein synthesis were predicted to be close to YvjB reinforces the prediction of its participation in this process and not another possible role for Ffh, such as its implication in the acid stress response (19). Furthermore, relationships with the signal peptidase II (LspA) and the prolipoprotein diacylglycerol transferase (Lgt) indicated that YvjB might be involved in the processing of lipoproteins. We tested the hypothesis of the participation of YvjB in the maturation of lipoproteins and observed that it participates in both, the correct localization and maturation of at least one lipoprotein, OptA, the oligopeptide binding protein, and its derivative SP_{YvdF}OptA.

We did not check on all of the predicted links, although most of them seemed pertinent. As an example, we can report the link between YvjB and the Ecs transport system. This relationship seemed reasonable because the Ecs system has been shown to act as a regulator of lipoprotein secretion in *Bacillus subtilis* (27, 34). Consequently, our prediction method produced a large number of specific new clues that need to be explored.

Global analysis of lipoproteins reveals that YvjB is necessary for correct localization of OptA

The results of our global analysis showed that not all lipoproteins were affected in the same way by the absence of *yvjB.* Three types of behavior were observed. In the first group, we found lipoproteins including OptA, whose abundance decreased in the absence of YvjB, while in the second group were lipoproteins whose abundance increased (such as YvdF) and in the third group were lipoproteins that were not affected by the absence of YvjB. Thus, an initial conclusion can

be drawn from these results: YvjB participated in the correct localization of group I lipoproteins. In its absence, these proteins did not reach, or only poorly attained, the cell envelope. Restoration of the correct localization of the OptA protein containing the YvdF signal peptide instead of its original peptide showed that the nature of the signal peptide mediated the recognition of lipoproteins by YvjB. Interestingly, OptA with the YvdF signal peptide, i.e. SP_{YvdF}OptA, was correctly localized in the cell envelope; not in its mature form, but nevertheless functional. At least one example of an immature but functional lipoprotein has already been reported for *L. lactis* (43). The effect of the absence of YvjB on several lipoproteins which generally ensure the substrate binding of transporters might explain why growth of the *yvjB* mutant was affected.

We wondered weather OptA could be liberated into the extra-cellular media in the *yvjB* mutant as had already been observed for other lipoproteins in *B. cereus* (17), which could explain its absence from the cell envelope in this strain. For this reason, we screened the presence of OptA in the culture supernatants during exponential phase and at the beginning of the stationary phase. We were not able to detect OptA in the supernatant of cells in the exponential phase and detected a weak signal (in both, the mutant and the wild-type) at the beginning of the stationary phase, probably due to cell lysis or to a release of OptA in late phase of growth as a part of a regulatory event. As the same observation was made in the mutant and the wild type strain only in stationary phase, we conclude that OptA was not specifically liberated into the extracellular medium, in a form detectable by the antibodies used, during the exponential phase of growth in the mutant strain.

As already mentioned, it is possible to imagine that OptA could be hydrolyzed while its SP_{YvdF}OptA derivative would remain more stable. Nevertheless, we did not detect any OptA fragments neither with our MS-MS analysis nor with antibodies. So, we concluded that *optA* was expressed but the corresponding protein OptA was not able to reach the cell envelope in an *yvjB* mutant. Even though, it is possible the OptA is completely degraded into oligopeptides or free amino acids, undetectable by the techniques we used.

YvjB is implicated in recognition and signal peptide cleavage

By combining these results, it is possible to say that YvjB is necessary at two steps of lipoprotein maturation of group I:

i) Correct localization of group I lipoproteins on the basis of the hydrophobicity of the signal peptide and
ii) Cleavage of the signal peptide of at least SP_{YvdF}OptA.

In *B. subtilis* and *E.coli*, the hydrophobicity of the signal peptides drives proteins through the SRP pathway: the most hydrophobic signal peptides tend to be recruited by SRP (26, 46). As the lipoproteins recognized by YvjB seem to be the ones with a less hydrophobic signal peptide, we propose that the recognition of group I lipoproteins by YvjB occurs in a later step of secretion, once the pre-secretory protein has been delivered to the SRP-receptor FtsY or even to the Sec machinery.

Taking account of the presence of PDZ and the results obtained by Fuh and colleagues that demonstrate its involvement in polypeptide recognition (16). Possibly, the consensus motif KXXL, found in the signal peptide proteins of group I (Table A2) could be the target of this recognition. The most likely hypothesis would be that the PDZ motif is implicated in the recognition of lipoproteins. Nevertheless, the PDZ motif has been predicted to be in the outer part of the cell in homologues of YvjB (11, 23). It thus remains unclear whether and how this motif might participate in recognizing the lipoprotein precursors.

The proteolytic domain is probably implicated in the second function of YvjB. Our results indicate that the presence of YvjB is necessary for signal peptidase II to cleave the signal peptide. With respect to Eep in *E. faecalis* (of which YvjB is the homolog) it has been determined that the N-terminal part of the peptide is recognized by this enzyme (10). The authors propose that the signal peptidase II (LspA) cleaves the signal peptide and that Eep acts on the peptide thus released. Nevertheless, An and Clewell (2) also showed that cleavage of the signal peptide by Eep could occur independently of cleavage by LspA. Our findings confirm that both YvjB and LspA are necessary for signal peptide cleavage. Other cleavages of lipoprotein precursors have already been observed, for example, on PrsA in *B. subtilis* (41) and on MtuA in *Streptococcus uberis* , in the absence of LspA (14). In *S. uberis,* because the alternatively cleaved form of MtuA disappears in a double mutant called *lspA-eep* (*eep* being the homologue of *yvjB*), the authors conclude that the alternative cleavage is achieved by Eep. These results show that the action of Yvjb could occur before cleavage of the signal peptide by LspA.

It has been postulated by (2) that LspA and Eep work in very close association, placing the lipoprotein to be matured between them like a sandwich. On this basis, and from our results, we can hypothesize that either cleavage by YvjB must occur first to allow further cleavage by LspA, or that YvjB needs to be present to maintain the lipoprotein in close contact with LspA, thus allowing cleavage to occur. A punctual mutation in the active site of YvjB could help to answer to this question.

Are group II lipoproteins recognized by another enzyme?
Our results also indicate that there may be a second enzyme with functions similar to YvjB that processes group II lipoproteins, as in the absence of YvjB and when OptA carries a group I signal peptide this protein overcomes the first step in maturation. This protein might possibly be YueF, which is endowed with a proteolytic domain (COG0612, PqqL, Predicted Zn-dependent peptidase) and two (by Toppred) predicted membrane spanning region (http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py?form=toppred). Our statistical analysis predicted it to be linked to YvjB.

Does the function of YvjB form part of RIP?
As already mentioned, YvjB is most likely a member of the S2P family. These proteins are all involved in regulating a cellular process. The generic mechanism by which these proteins participate in protein maturation has been called Regulated Intermembrane Proteolysis (RIP) (8). This refers to a mechanism in which a membrane embedded protease produces signals during the degradation of a lipoprotein or membrane protein. The RIP acts on a portion of the substrate protein that is smaller than 30 amino acids, in *E. faecalis*, for instance, it is the signal peptide of lipoproteins (8). We have demonstrated the participation of YvjB in the processing of a particular type of lipoproteins. Its implication in this process had been suggested by statistical predictions. We have also demonstrated that YvjB participates in the maturation of at least one lipoprotein, i.e. SP_{YvdF}OptA. Comparison of the peptide sequences known to be released by Eep (LFSLVLAG, (2) and LVTLVFV, (9)) in *E. faecalis* and SP_{YvdF}OptA (...TLGTVALGSAALLAAC) in *L. lactis* indicates a certain similarity and the presence of a leucine between positions -7 to -9 of LspA cleavage. It is possible that this amino acid plays a crucial role in the production of a peptide resulting from lipoprotein maturation. We have not yet been able to detect the peptide that may be released and could not confirm our cleavage hypothesis It is not clear either whether, in *L. lactis*, the released peptide is simply recycled or acts as a signaling molecule, in a way similar to what happens in Enterococcus.

In conclusion, we have shown that YvjB participates in the correct localization of lipoproteins with low hydrophobicity signal peptides. Its implication involves the recognition of these types of proteins. YvjB is also necessary for the cleavage of lipoproteins with high hydrophobicity signal peptides. In this way, YvjB can cleave the lipoprotein signal peptide firstly, to allow complete cleavage by LspA and/or to retain the lipoprotein to be matured within the membrane so as to allow its cleavage by LspA. Further studies are needed to determine precisely at which moment YvjB participates in the recognition and maturation of lipoproteins and what is its exact action.

Acknowledgements
This work was entirely financed by the "Institut National de la Recherche Agronomique" INRA in France. In the "Unité de bactéries lactiques et pathogènes opportunistes" at INRA in Jouy-en-Josas we thank Vincent Juillard and his team for providing us with the OptA knock-out mutant and OptA antibodies. We thank Rozenn Gardan, Vincent Juillard and Isabelle Poquet for their critic reading of the manuscript and their advice.

Figures
Figure 1 : Graphical representation of the predicted relationships found for YvjB by KCCA, organized by functional categories. Predicted distances are in brackets. The complete list and precise distances can be found in the appendix (Table A1). Proteins belonging to the first ten relationships of YvjB are underlined. Their complete names are RplE: 50S ribosomal protein L5, PrfA : peptide chain release factor RF-1, GidA: glucose inhibited division protein, NusG: transcription antitermination protein, Ffh: signal recognition particle protein, YuaA: putative CMP-binding factor, YhhG: hypothetical protein, YlxQ: probable ribosomal protein, YccI: contains N-acetyltransferase motif, YhhC: homolog of LrgB of *Staphylococcus aureus*.

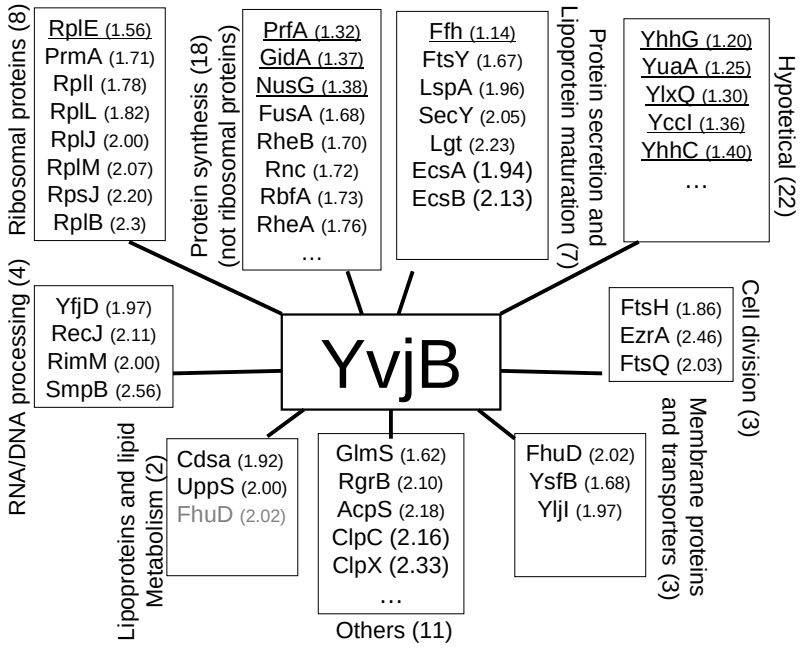


Figure 2 : Representation of the proteins predicted to be related to YvjB and participating in co-translational secretion mediated by SRP and lipoprotein maturation. Arrows with dashed lines indicate possible relationships between YvjB and other proteins. SP: signal peptide.

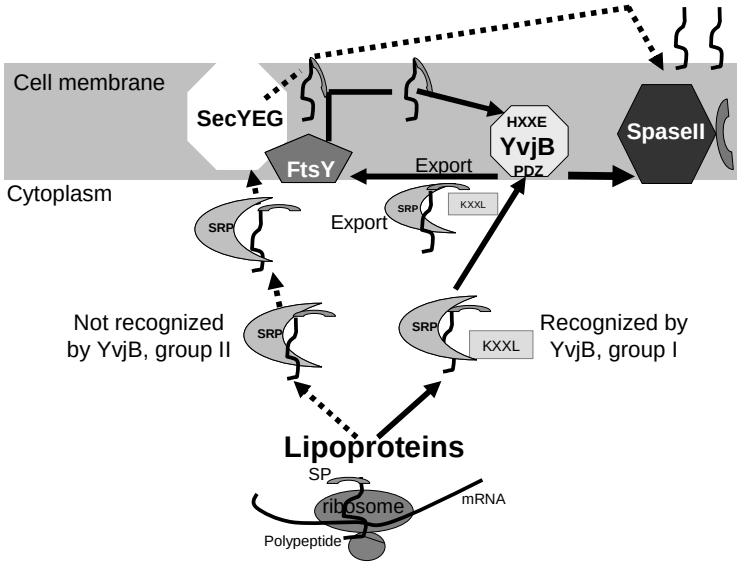


Figure 3 : A) Detection of OptA by Western blot in the wild-type strain (IL1403) and the *yvjB* mutant in three cell fractions at DO_{600}=0.6 and in the extracellular fraction at DO_{600}=2.. B) Detection of OptA in the strain containing an YvdF signal peptide instead of the original OptA signal peptide (SP_{YvdF}OptA). C: cytoplasmic fraction, E: envelope fraction, S: extracellular fraction. Masses were determined by comparison with a known marker (not shown). The detection of OptA in the wild-type strain (IL1403) and the *yvjB* mutant in the cytoplasm and envelope fraction at DO_{600}=0.6 has been done 5 times and the same results were obtained.

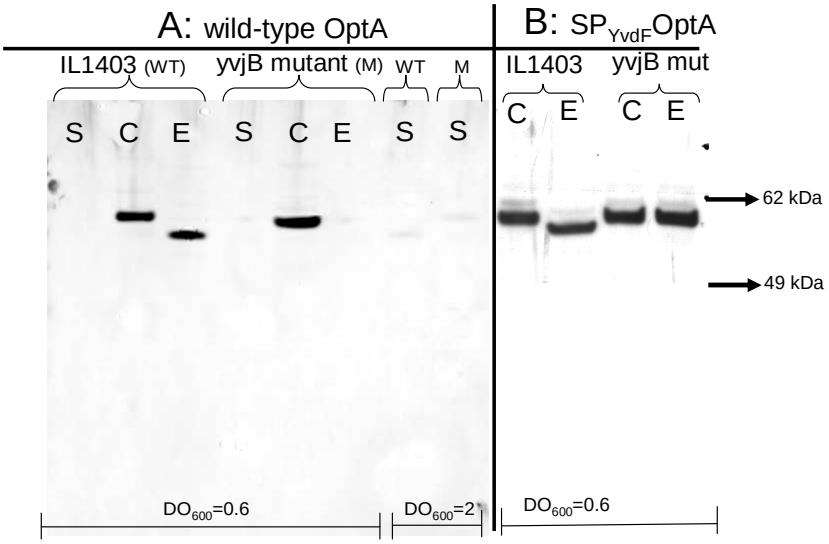


Figure 4: Detection of OptA by Western blot in the wild-type strain (IL1403), the *lspA* mutant, the double mutant *lspA-yvjB* and the *lspA* mutant containing an YvdF signal peptide instead of the original OptA signal peptide (SP_{YvdF}OptA). C: cytoplasmic fraction, E: envelope fraction. Masses were determined by comparison with a known marker (not shown).

Tables

Table 1 : Bacterial strains used in this work

Strain	Plasmid content	Resistance	Source
L. lactis IL1403	-	-	(12)
L. lactis IL1403 Δ*yvjB*	-	-	this work
L. lactis IL1403 Δ*yvjB* compl.	pILN13-*yvjB* operon	Ery	this work
L. lactis IL1403- SP$_{YvdF}$OptA	-	-	this work
L. lactis IL1403 Δ*yvjB* - SP$_{YvdF}$OptA	-	-	this work
L. lactis IL1403 Δ*lspA*	-	-	this work

Table 2 : Plasmids used in this work

Plasmid	Characteristics	Resistance	Source
pGhost9-Δ*yvjB*	-	Ery	this work
pILN13- *yvjB* operon	low copy number	Ery	this work
pGhost9- SP$_{YvdF}$OptA	-	Ery	this work
pGhost9-Δ*lspA*	-	Ery	this work

Table 3 : List of primers used in this work. Restriction sites are underlined.

Primer name	Primer sequence	Restriction sites
eepUF	CTCCTCGAGAGAACATCTGGTGAACAAAG	*Xho*I
eepUR	GAAGAATTCTTAGAATAGTCCAAGTAGATGC	*Eco*RI
eepDF	GAAGAATTCATCTATTCGTCTTTCTATTTC	*Eco*RI
eepDR	AACTACTAGTTTCAGTATAAGCTACATTTG	*Spe*I
eepF	GTAAAAGATTCAGGTAAAAT	
eepR	GAAATAGAAGACGAATAGAT	
CeepF	TGCACTGCAGTAATAACCAAAGAGCATAAT	*Pst*I
CeepR	GGCCGCGGCCGCATTTCCTTTATGAGTTTGGT	*Not*I
UoptSF	CTTTACACCAGTGGGAATGAG	
UyvdFi-UoptSR	TTGCAAGAGTCAACAATTTTATTTTCTTCATGTAATTTCCCTC	
LoptAR	CATACCCGCTGGTGTTAAGC	
UyvdF-LopAF	AAATAAAATTGTTGACTCTTGCAA TTGCTTGCAGCATGCGG	
LyvdFi-UyvdFi	GCCAATTGAGGCAATTCCAACTGTTGCAAGAGTCAACAATTTTATTT	
LyvdF-LoptAF	CAGTTGGAATTGCCTCAATTGGC TTGCTTGCAGCATGCGG	
opta-seqF	GTGAAAACTTTCCGCTATTT	
opta-seqR	AAGTATGGTGTTGGTTGTGC	
lspAUF	CTCCTCGAGATAAACTTCGTGCCTACTCA	*Xho*I

lspAUR	GATCGGATCC*TTA*TTTTTCTGTATCTCCAAGCT	*Bam*HI *stop*
lspALF	GATCGGATCCAATTGATTAAAAAATCACTG	*Bam*HI
lspALR	CCCGGGTAAATACAAAAAAAACAAGT	*Sma*I
lspAF	TATATGTTTAGGGACTGTCA	
lspAR	TACTTTGCTTTTTTTGCTTC	

Table 4 : Results of MS-MS analysis of the cell envelope of the wild-type strain (IL1403) and the *yvjB* mutant in terms of the number of peptide spectra (SC) per protein detected. Da: dalton. Gravy: hydrophobicity index SP: signal peptide, NS: statistically not significant.

		Protein characteristics			**LC-MSMS results**			
Group	**Protein**	**Function**	**Mass in Da**	**Gravy SP**	**Average SC IL1403**	**Average SC *yvjB***	**SC difference**	**% SC difference**
I	**FrdC**	Fumarate reductase flavoprotein subunit	52852	1.135	32.0±6.9	23.3±1.2	-8.7	-27
	OptA	Oligopeptide ABC transporter substrate binding protein	59697	0.9	18.0±4.6	0.0±0.0	-18.0	-100
	YpcG	Sugar ABC transporter substrate binding protein	52938	1.039	3.0±1.0	0.0±0.0	-3.0	-100
II	**YvdF**	Amino acid ABC transporter substrate binding protein	31023	1.461	8.3±0.6	18.0±6.2	9.7	+117
	YjgC	Amino acid ABC transporter substrate binding protein	30609	1.383	12.3±2.9	21.0±6.1	8.7	+70
III	**MtsA**	Manganese ABC transporter substrate binding protein	35499	1.688	3.7±2.3	2.7±1.2	-1.0	NS
	YtcC	Hypothetical protein	39930	1.5	11.3±4.0	8.7±1.5	-2.7	NS
	ApbE	Thiamine biosynthesis lipoprotein	39676	1.805	7.7±2.3	6.7±1.5	-1.0	NS
	PmpA	Maturation protein	33813	1.172	7.3±1.5	7.7±3.1	0.3	NS

YwaI	Unknown protein	36700	1.127	3.7±3.5	5.3±0.6	1.7	NS
PstE	Phosphate ABC transporter substrate binding protein	30560	1.533	8.0±1.7	9.3±3.8	1.3	NS
ZitS	Zinc ABC transporter substrate binding protein	30725	1.533	3.3±1.2	2.0±1.0	-1.3	NS
YfcG	Peptide-binding protein	59922	1.143	7.3±2.3	4.3±2.1	-3.0	NS
FhuD	Ferrichrome ABC transporter substrate binding protein	34337	1.54	10.7±1.5	13.3±0.6	2.7	NS
BmpA	Basic membrane protein A	36652	1.527	50.7±15.0	54.7±6.1	4.0	NS

References

1. **Akiyama, Y., K. Kanehara, and K. Ito.** 2004. RseP (YaeL), an *Escherichia coli* RIP protease, cleaves transmembrane sequences. The EMBO Journal **23:**4434-4442.
2. **An, F. Y., and D. B. Clewell.** 2002. Identification of the cAD1 sex pheromone precursor in *Enterococcus faecalis*. Journal of Bacteriology **184:**1880-1887.
3. **An, F. Y., M. C. Sulavik, and D. B. Clewell.** 1999. Identification and characterization of a determinant (*eep*) on the *Enterococcus faecalis* chromosome that is involved in production of the peptide sex pheromone cAD1. Journal of Bacteriology **181:**5915-5921.
4. **Antiporta, M. H., and G. M. Dunny.** 2002. *ccfA*, the genetic determinant for the cCF10 peptide pheromone in *Enterococcus faecalis* OG1RF. Journal of Bacteriology **184:**1155-1162.
5. **Bohn, C., J. Collier, and P. Bouloc.** 2004. Dispensable PDZ domain of *Escherichia coli* YaeL essential protease. Molecular Microbiology **52:**427-435.
6. **Bolotin, A., P. Wincker, S. Mauger, O. Jaillon, K. Malarme, J. Weissenbach, S. D. Ehrlich, and A. Sorokin.** 2001. The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *lactis* IL1403. Genome Research **11:**731-753.
7. **Bramkamp, M., L. Weston, D. Richard, and J. Errington.** 2006. Regulated intramembrane proteolysis of FtsL protein and the control of cell division in *Bacillus subtilis*. Molecular Microbiology **62:**580-591.
8. **Brown, M. S., Y. Jin, R. B. Rawson, and J. L. Goldstein.** 2000. Regulated Intramembrane Proteolysis: A Control Mechanism Conserved from Bacteria to Humans.

Cell **100:**391-398.
9. **Buttaro, B. A., M. H. Antiporta, and G. M. Dunny.** 2000. Cell-associated pheromone peptide (cCF10) production and pheromone inhibition in *Enterococcus faecalis*. Journal of Bacteriology **182:**4926-4933.
10. **Chandler, J. R., and G. Dunny.** 2008. Characterization of the sequence specificity determinants required for processing and control of sex pheromone by the intramembrane protease Eep and the plasmid-encoded protein PrgY. Journal of Bacteriology **190:**1172-1183.
11. **Chen, J. C., V. P.H., and L. Shapiro.** 2005. A membrane metalloprotease participates in the sequential degradation of a *Caulobacter* polarity determinant. Molecular Biology **55:**1085-1103.
12. **Chopin, A., M.-C. Chopin, A. Moillo-Batt, and P. Langella.** 1984. Two plasmid-determined restriction and modification systems in *Streptococcus lactis*. Plasmid **11:**260-263.
13. **Courtin, P., G. Miranda, A. Guillot, F. Wessner, and C. Mézange.** 2006. Peptidoglycan Structure Analysis of *Lactococcus lactis* Reveals the presence of an L,D-Carboxypeptidase Involved in Peptidoglycan Maturation. Journal of Bacteriology **188:**5293-5298.
14. **Denham, E. L., P. N. Ward, and J. A. Leigh.** 2008. Lipoprotein Signal Peptides Are Processed by Lsp and Eep of *Strepotococcus uberis*. Journal of Bacteriology **190:**4641-4647.
15. **Exterkate, F. A.** 1976. The proteolytic system of a slow lactic-acid-producing variant of *Streptococcus cremoris* HP. Netherlands Milk and Dairy Journal **30:**3-8.
16. **Fuh, G., M. T. Pisabarro, Y. Li, C. Quan, L. A. Lasky, and S. S. Sidhu.** 2000. Analysis of PDZ domain-ligand interactions using carboxy-terminal phage display. Journal of Biological Chemistry **275:**21486-21491.
17. **Gilois, N., N. Ramarao, L. Bouillaut, S. Perchat, S. Aymerich, C. Nielsen-LeRoux, D. Lereclus, and M. Gohar.** 2007. Growth-related variations in the *Bacillus cereus* secretome. Proteomics **7:**1719-1728.
18. **Gottesman, S.** 1996. Proteases and their targets in *Escherichia coli*. Annual Review of Genetics **30:**465-506.
19. **Gutierrez, J. A., P. J. Crowley, D. G. Cvitkovitch, L. J. Brady, I. R. Hamilton, J. D. Hillman, and A. S. Bleiweis.** 1999. *Streptococcus mutans ffh*, a gene encoding a homologue of the 54 kDa subunit of the signal recognition particle, is involved in resistance to acid stress. Microbiology **145:**357-366.
20. **Hoffman, C. S., and F. Winston.** 1987. A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of *Escherichia coli*. Gene **57:**267-272.
21. **Kajetan, H., B. A. Firek, D. F. Fujimoto, and K. W. Bayles.** 2000. The *Staphylococcus aureus lrgAB* operon modulates murein hydrolase activity and penicillin tolerance. Journal of Bacteriology **182:**1794-1801.
22. **Kanehisa, M., S. Goto, S. Kawashima, and A. Nakaya.** 2002. The KEGG databases at GenomeNet. Nucleic Acids Res **30:**42-46.
23. **Kinch, L. N., K. Ginalski, and N. V. Grishin.** 2006. Site-2 protease regulated intramembrane proteolysis: Sequence homologs suggest an ancient signaling cascade. Protein Science **15:**84-93.
24. **Kunji, E. R., G. Fang, P. Margot, A. P. Bruins, B. Poolman, and W. N. Konings.** 1998. Reconstruction of the proteolytic pathway for use of beta-casein by *Lactococcus lactis*. Molecular Microbiology **27:**1107-1118.

25. **Lamarque, M., P. Charbonnel, D. Aubel, J. C. Piard, D. Atlan, and V. Juillard.** 2004. A Multifunction ABC Transporter (Opt) Contributes to Diversity of Peptide Uptake Specificity within the genus *Lactococcus*. Journal of Bacteriology **186:**6492-6500.
26. **Lee, H. C., and H. D. Bernstein.** 2001. The targeting pathway of *Escherichia coli* presecretory and integral membrane proteins is specified by the hydrophobicity of the targeting signal. Proc Natl Acad Sci **98:**3471-3476.
27. **Leskelä, S., E. Wahlström, H. L. Hyyrlyläinen, M. Jacobs, A. Palva, M. Sarvas, and V. P. Kontinen.** 1999. Ecs, an ABC transporter of *Bacillus subtilis*: dual signal transduction as well as their secretion. Molecular Microbiology **31:**533-543.
28. **Lopez Kleine, L., V. Monnet, C. Pechoux, and A. Trubuil.** 2008. Role of bacterial peptidase F inferred by statistical analysis and further experimental validation. HFSP Journal **2:**29-41.
29. **Mierau, I., E. R. S. Kunji, K. J. Leenhouts, M. A. Hellendorn, A. J. Haandrikman, B. Poolman, W. N. Konings, G. Venema, and J. Kok.** 1996. Multiple-peptidase mutants of *Lactococcus lactis* are severely impaired in their ability to grow in milk. Journal of Bacteriology **178:**2794-2803.
30. **Miyoshi, S., and S. Shinoda.** 2000. Microbial metalloproteases and pathogenesis. Microbes and Infection **2:**91-98.
31. **Otto, R., B. ten Brink, H. Veldkamp, and W. N. Konings.** 1983. The relation between growth rate and the electrochemical proton gradient of *Streptococcus cremoris*. FEMS Microbiology Letters **16:**69-74.
32. **Poolman, B., E. R. Kunji, A. Hagting, V. Juillard, and W. N. Konings.** 1995. The proteolytic pathway of *Lactococcus lactis*. Journal of Applied Bacteriology, Symp. Suppl. **79:**65S-75S.
33. **Poquet, I., V. Saint, E. Seznec, N. Simoes, A. Bolotin, and A. Gruss.** 2000. HtrA is the unique surface housekeeping protease in *Lactococcus lactis* and is required for natural protein processing. Microbiology **35:**1042-1051.
34. **Pummi, T., S. Leskelä, E. Wahlström, U. Gerth, H. Tjalsma, M. Hecker, M. Sarvas, and V. P. Kontinen.** 2002. ClpXP Protease regulates the Signal peptide Cleavage of Secretory preproteins in *Bacillus subtilis* with a mechanism Distinct from That of the Ecs ABC Transporter. Journal of Bacteriology **184:**1010-1018.
35. **Redko, Y., P. Courtin, C. Mézange, C. Huard, and M. P. Chapot-Chartier.** 2007. *Lactococcus lactis* Gene yjgB Encodes a gamma-D-Glutaminyl-L-Lysyl-Endopeptidase Which Hydrolyzes Peptidoglycan. Applied and Environmental Microbiology **73:**5825-5831.
36. **Renault, P., G. Corthier, N. Goupil, C. Delorme, and S. D. Ehrlich.** 1996. Plasmid vectors for Gram-positive bacteria switching from high to low copy number. Gene **183:**175-182.
37. **Rudner, D. Z., P. Fawcette, and R. Losick.** 1999. A family of membrane-embedded metalloproteases involved in regulated proteolysis of membrane-associated transcription factors. Proc Natl Acad Sci **96:**1007-1018.
38. **Sambrook, J., and D. W. Russel.** 2001. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press.
39. **Savijoki, K., H. Ingmer, and P. Varmanen.** 2006. Applied Microbiology and Biotechnology. Appl Microbiol Biotechnol **71:**394-406.
40. **Scott, J. R., and T. C. Barnett.** 2006. Surface Proteins of Gram-Positive Bacteria and How They Get There. Annual Review of Microbiology **60:**397-423.
41. **Tjalsma, H., V. P. Kontinen, Z. Pragai, H. Wu, R. Meima, G. Venema, S. Bron, M. Sarvas, and J. M. van Dijl.** 1999. The role of lipoprotein processing by signal peptidase

II in the Gram-positive Eubacterium *Bacillus subtilis*. The Journal of Biological Chemistry **274:**1689-1707.
42. **Varmanen, P., H. Ingmer, and F. K. Vogensen.** 2000. *CtsR* of *Lactococcus lactis* encodes a negative regulator of *clp* gene expression. Microbiology **146:**1447-1455.
43. **Venema, R., H. Tjalsma, J. M. van Dijl, A. de Jong, K. Leenhoust, G. Buist, and G. Venema.** 2003. Active Lipoprotein Precursors in the Gram-positive. The Journal of Biological Chemistry **278:**14739-14746.
44. **Wydau, S., E. Dervyn, J. Anba, D. Ehrlich, and E. Maguin.** 2006. Conservation of key elements of natural competence in *Lactococcus lactis* ssp. FEMS Mcrobiology Letters **263:**223-228.
45. **Yamanishi, Y., J. P. Vert, A. Nakaya, and M. Kanehisa.** 2003. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. Bioinformatics **19:**i323-i330.
46. **Zanen, G., E. N. G. Houben, R. Meima, H. Tjalsma, J. D. Jongbloed, H. Westers, B. Oudega, J. Luirink, J. M. van Dijl, and W. J. Quax.** 2005. Signal peptide hydrophobicity is critical for early stages in protein export by *Bacillus subtilis*. FEBS Journal **272:**4617-4630.

Appendixes

Table A1: summarizing the results of predictions for proteins related to YvjB by kernel canonical correlation analysis:

position	distance	gene	protein
1	1.1438	ffh	signal recognition particle protein Ffh
2	1.1984	yhhG	hypothetical protein, predicted nucleic acid binding protein implicated in transcription termination
3	1.2474	yuaA	putative CMP-binding factor
4	1.2947	ylxQ	hypothetical protein, putative ribosomal protein
5	1.3231	prfA	peptide chain release factor RF-1
6	1.3584	yccI	hypothetical protein, putative N-actetyltransferase
7	1.3663	gidA	glucose inhibited division protein GidA
8	1.3778	nusG	transcription antitermination protein
9	1.4027	yhhC	hypothetical protein, putative effector of murein hydrolase
10	1.5610	rplE	50S ribosomal protein L5
11	1.6201	glmS	glucosamine--fructose-6-phosphate aminotransferase
12	1.6762	fusA	elongation factor EF-G
13	1.6968	rheB	ATP-dependent RNA helicase
14	1.6970	ftsY	cell division protein FtsY
15	1.7141	prmA	ribosomal protein L11 methyltransferase
16	1.7232	rnc	ribonuclease III
17	1.7236	rbfA	ribosome-binding factor A
18	1.7553	rheA	ATP-dependent RNA helicase
19	1.7758	rplI	50S ribosomal protein L9
20	1.7865	yueF	putative protease
21	1.8190	rplL	50S ribosomal protein L7/L12
22	1.8576	ftsH	cell division protein FtsH
23	1.8801	gidB	glucose-inhibited division protein GidB
24	1.8889	sunL	rRNA methylase; K03500 Sun protein
25	1.9179	cdsa	phosphatidate cytidylyltransferase; Glycerophospholipid metabolism
26	1.9419	ecsA	ABC transporter ATP binding protein

27	1.9446	rnhB	ribonuclease HII
28	1.9488	yedA	hypothetical protein
29	1.9608	lspA	lipoprotein signal peptidase
30	1.9687	yljI	permease
31	1.9718	yfjD	tRNA/rRNA methyltransferase
32	1.9956	rplJ	50S ribosomal protein L10
33	1.9960	uppS	undecaprenyl pyrophosphate synthetase
34	2.0012	rimM	16S rRNA processing protein
35	2.0062	yjaI	unknown protein
36	2.0211	fhuD	ferrichrome ABC transporter substrate binding protein
37	2.0320	ftsQ	cell division protein FtsQ
38	2.0541	secY	preprotein translocase SecY subunit
39	2.0691	ysfB	ABC transporter ATP-binding protein
40	2.0711	rplM	50S ribosomal protein L13
41	2.0983	rgrB	GntR family transcriptional regulator
42	2.1076	recJ	single-stranded DNA specific exonuclease
43	2.1347	ecsB	ABC transporter permease protein
44	2.1382	yhbE	unknown protein
45	2.1521	obgL	GTP-binding protein Obg
46	2.1554	infC	translation initiation factor IF-3
47	2.1581	clpC	ATP-dependent protease ATP-binding subunit
48	2.1645	priA	primosomal protein N'(replication factor Y) (superfamily II helicase)
49	2.1684	prfB	peptide chain release factor RF-2
50	2.1739	yjaD	transcription regulator
51	2.1808	acpS	holo-[acyl-carrier protein] synthase
52	2.1942	yccE	unknown protein
53	2.1982	rpsJ	30S ribosomal protein S10
54	2.2000	yfdB	hypothetical protein
55	2.2071	yacB	hypothetical protein
56	2.2327	lgt	prolipoprotein diacylglycerol transferase
57	2.2361	pheS	phenylalanyl-tRNA synthetase alpha chain
58	2.2916	tig	trigger factor
59	2.3031	apt	adenine phosphoribosyltransferase; Purine metabolism
60	2.3040	rplB	50S ribosomal protein L2;
61	2.3052	yccL	hypothetical protein
62	2.3097	gatB	Glu-tRNA amidotransferase subunit B
63	2.3139	dnaA	replication initiation protein DnaA
64	2.3337	clpX	ATP dependent Clp protease
65	2.3410	xpt	xanthine phosphoribosyltransferase
66	2.3439	ywdG	hypothetical protein
67	2.3593	yebE	hypothetical protein
68	2.3626	ydgI	hypothetical protein
69	2.3830	ptpL	protein-tyrosine phosphatase
70	2.4024	yhfD	hypothetical protein
71	2.4035	yqdA	hypothetical protein
72	2.4313	ygcC	hypothetical protein
73	2.4409	yofM	hypothetical protein
74	2.4453	ksgA	dimethyladenosine transferase
75	2.4502	ybaF	hypothetical protein
76	2.4577	lysS	lysyl-tRNA synthetase
77	2.4626	ezrA	cell division regulator
78	2.5570	smpB	tmRNA-binding protein SmpB

Table A2: Peptide signal of lipoproteins recognized by YvjB (group I) and not recognized by YvjB (group II). Common characteristics: lysine (K) in position 6, leucine (L) in position 9 and hydrophobic amino acid in position -2 of the lipobox, are highlighted in dark grey.

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Group I																							
FrdC	M	K	I	W	T	K	L	G	L	L	T	L	V	G	L	S	L	T	S	C			
OptA	M	K	T	W	K	K	V	T	L	G	T	V	A	L	G	S	A	A	L	L	A	A	C
Group II																							
YvdF	M	K	K	I	K	L	L	T	L	A	T	V	G	I	A	S	I	G	L	L	A	A	C
YjgC	M	R	T	S	L	K	V	T	F	A	A	L	S	L	I	A	A	G	T	L	V	A	C

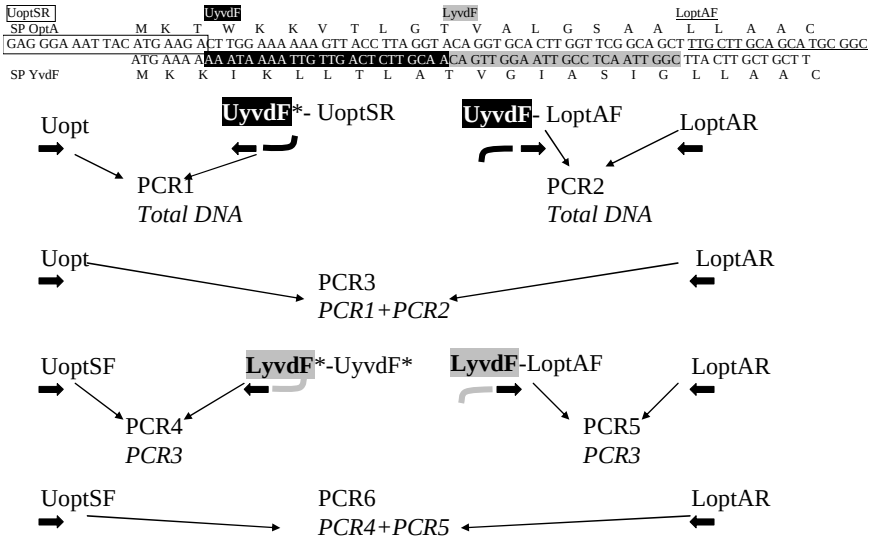


Figure A1: Procedure for the replacement of the signal peptide nucleotide sequence of *optA* by the signal peptide of *yvdF*. The regions to be replaced were added to the primers during two subsequent PCRs. PCR substrates are in italics. The first region to have been replaced is highlighted in black in the text and represented in black on the primer. The second region to be replaced is represented in grey.

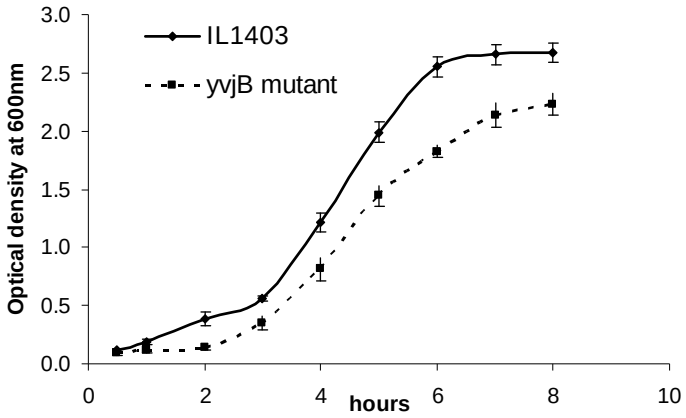


Figure A2: Growth of the *yvjB* mutant compared to the wild-type strain (IL1403) in M17-glu medium.

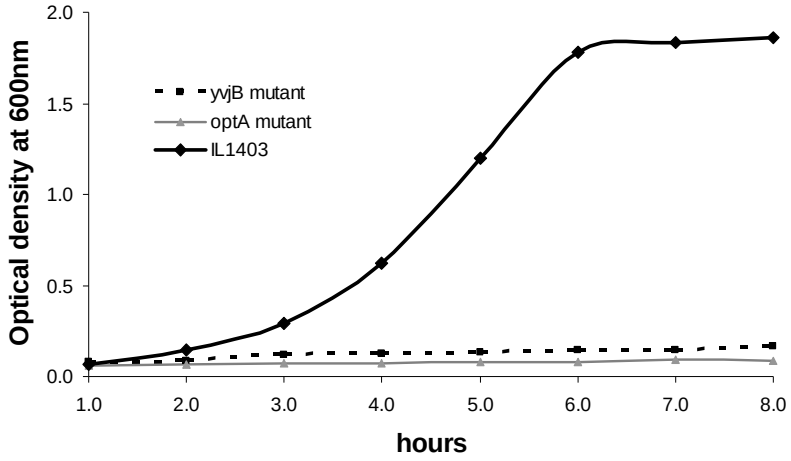


Figure A3a: Growth of the wild type strain (IL1403), *yvjB* mutant and *optA* mutant in a medium containing leucine only in the form of leu-enkephalin:

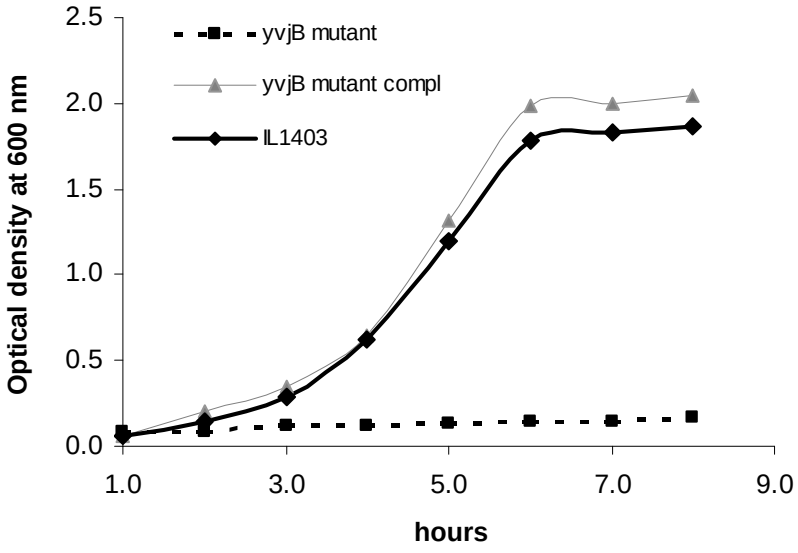


Figure A3b: Growth of the wild-type strain (IL1403), the *yvjB* mutant and the complemented mutant in a medium containing leucine only in the form of leu-enkephalin:

6 Résultats de l'analyse de sensibilité

La démarche que nous avons utilisée dans cette étude pour déterminer le rôle des protéines PepF et YvjB comporte deux grandes étapes :

1. La prédiction de rôles par une analyse statistique (KCCA).
2. La validation expérimentale des liens prédits.

Nous nous retrouvons à ce stade du travail avec un certain nombre de liens validés, parce que nous avons confirmé que PepF et YvjB interviennent dans les fonctions prédites par la KCCA. La démarche est satisfaisante, puisqu'elle nous a permis d'avancer dans la connaissance sur ces deux protéines. Néanmoins, plusieurs questions en relation avec les données les plus importantes pour la prédiction des liens peuvent être posées : Est-ce que il y a des données cruciales pour arriver aux prédictions ? Est-ce que, au contraire, il y a des données qui n'apportent rien à la prédiction ? Serait il possible de trouver un sous-ensemble minimal stable de données conduisent aux même prédictions que la totalité de données ?

Pour répondre à ces questions nous nous sommes inspirés de l'analyse de sensibilité qui consiste, comme expliqué dans la synthèse bibliographique, à étudier comment varie la réponse d'un algorithme en fonction des variations des données d'entrée.

Dans le cadre de l'analyse réalisée nous avons plusieurs types de données ou paramètres d'entrée :

1. Le graphe des voies métaboliques
2. Les données transcriptomiques
3. Les profils phylogénétiques
4. Les données protéomiques
5. Les distances sur le chromosome
6. Les paramètres (des noyaux et de la KCCA)

Chaque source de données est constituée d'un ensemble d'éléments portant de l'information (333 gènes sur le graphe des voies métaboliques, 115 hybridations pour les données transcriptomiques, 276 organismes avec lesquels les profils phylogénétiques ont été construits, etc.). Nous cherchons à identifier les éléments importants dans ces ensembles. Il n'est cependant pas possible de tester toutes les combinaisons de sous-ensembles correspondants à toutes ces données. Afin d'identifier les combinaison d'éléments de données importante nous avons associé à chaque élément des données testé une variable binaire ou facteur indiquant son utilisation (ou non utilisation) dans la prédiction.

Nous avons décidé de ne pas faire varier le graphe des voies métaboliques, qui est notre source de donnée standard pour l'apprentissage, et de nous concentrer uniquement sur les deux sources de données les plus informatives : les données transcriptomiques et les profils phylogénétiques. Pour éviter de faire varier les paramètres des noyaux et de KCCA nous avons décidé d'inclure dans l'analyse la validation croisée, qui permet de déterminer les valeurs de paramètres optimaux pour chaque jeu de données, c'est à dire pour chaque simulation réalisée au cours de l'analyse de sensibilité. L'algorithme que nous voulons analyser se compose alors d'une KCCA (avec détermination des paramètres) faite sur les données transcriptomiques et les profils phylogénétiques

en utilisant le graphe complet (sans réduire) des voies métaboliques comme référence (Figure 6-1).

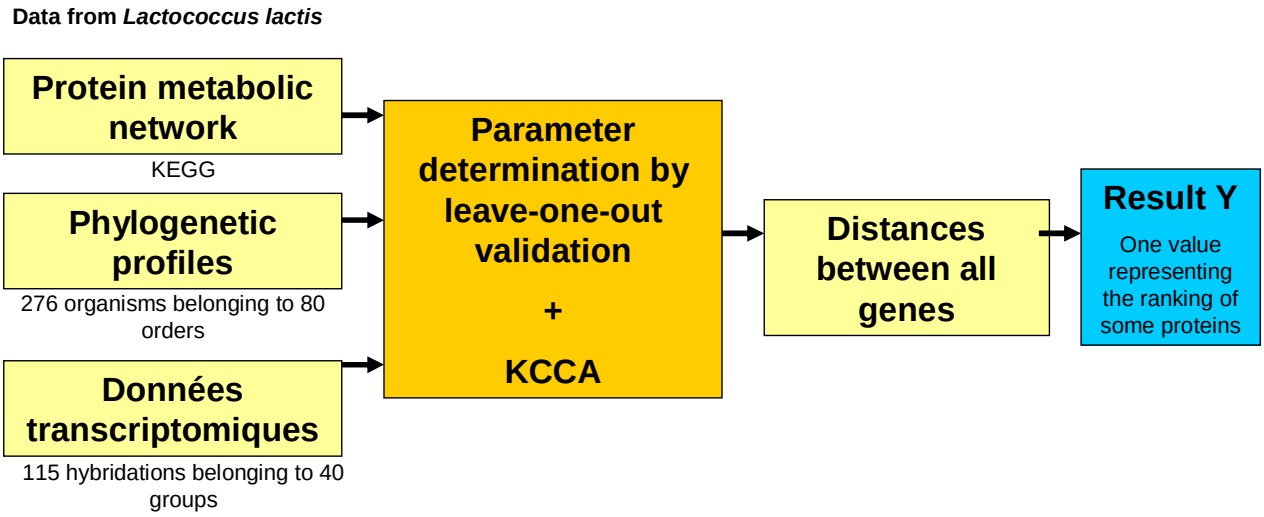


Figure 6-1 : Modèle analysé par analyse de sensibilité qui montre les données disponibles pour *L. lactis* qui ont été utilisées pour réalisée les simulations (détermination de paramètres, puis KCCA, puis obtention d'une valeur de sortie *y* pour chaque simulation).

Le nombre de combinaisons demeure encore très grand ($2^P \times 2^E$, pour P expériences transcriptomiques et E organismes). Nous avons réduit le nombre de facteurs en regroupant les données transcriptomiques par répétitions et par expériences similaires et les organismes avec lesquels les profils phylogénétiques ont été construits par la catégorie taxonomique qui regroupe les familles « ordre ». Cela permet de passer de 115 à 40 facteurs les données transcriptomiques et de 276 à 80 facteurs pour les profils phylogénétiques. Néanmoins, les possibilités à tester restent grandes, 2^{40} et 2^{80} si un type de données est testé et l'autre fixé. Il est donc nécessaire d'utiliser un plan d'expériences pour réaliser les simulations. Ce plan d'expériences doit comprendre un nombre réalisable de simulations de manière à construire un modèle qui permette d'attribuer un effet à chaque facteur. Il est aussi important que les facteurs principaux ne soient pas confondus avec les facteurs d'interactions. Nous avons alors construit un plan d'expériences qui possède ces caractéristiques, et qui nous a permis de déterminer les effets principaux sans confusion avec les effets d'interactions de deuxième ordre. Certains effets d'interaction sont confondus « par paquets », ce qui conduit à une structure d'alias. Ainsi par exemple l'effet d'interaction 1×2 peut être confondu avec 3×7 et 4×15. Ce plan nous a fourni les simulations à réaliser en incluant (1) ou n'incluant pas (-1) les expériences transcriptomiques, ou les organismes des profils phylogénétiques. Nous avons ainsi réalisé deux analyses de sensibilité. La première considère 128 simulations pour déterminer les effets des données transcriptomiques et toutes les données de profils phylogénétiques. Ces dernières ne sont donc pas perturbées dans cette première analyse. La deuxième a comporté 256 simulations pour étudier les effets des organismes des profils phylogénétiques, laissant fixes les données transcriptomiques.

L'analyse de sensibilité est conçue de manière à ce que le résultat des simulations soit représenté par un scalaire. Un modèle linéaire pour déterminer les effets des facteurs est associé et estimé par moindres carrés à partir de la matrice d'expériences et du vecteur contenant la sortie des

simulations. La sortie de l'algorithme est un vecteur qui décrit la distance de PepF ou YvjB avec les autres protéines du lactocoque et non pas un scalaire. Pour résumer cette sortie en une seule valeur *y* nous nous sommes concentrés sur certaines protéines et spécifiquement sur leur changement de rang dans la liste de protéines prédites comme étant liées à PepF ou YvjB. La valeur de référence a été fixée à *y=0*. Si les protéines analysées changent de rang, cette valeur de référence est augmentée d'une valeur qui dépend de l'ordre de grandeur du changement et du poids donné (arbitrairement) à chacune des protéines liées et sont :

$$y = \frac{1}{2}\sum_{i=1}^{L} w_i erfc(h(S_i - V_i))$$

où w_i désigne le poids d'une protéines liée, S_i : la tolérance d'écart au classement de référence, V_i : l'écart de classement, h : écart type de la gaussienne. Nous avons régularisé cette fonction d'écart à la prédiction de référence (avec toutes les données) en convoluant avec une gaussienne de variance proportionnelle à *h*.

Les résultats de l'analyse de sensibilité pour PepF sont décrits dans l'article III. Cette analyse nous a permis de mieux comprendre l'importance de certaines données et surtout de trouver un ensemble de données suffisantes et nécessaires pour arriver aux mêmes prédictions qu'en utilisant toutes les données de départ. L'algorithme pour déterminer ce sous-ensemble est expliqué brièvement dans l'article III. Une explication plus précise est donnée dans la section Matériels et Méthodes. Cependant, l'interprétation des résultats pour le cas de PepF s'avère difficile, en raison de son rôle pleïotropique, qui fait intervenir des expériences et des organismes très hétérogènes. Par ailleurs, l'analyse nous a permis de confirmer que les prédictions les plus répétables lors des différentes simulations sont en effet les liens qui ont été confirmés comme indiquant la fonction principale des protéines cibles, à savoir le lien entre PepF et SipL et le lien entre YvjB et Ffh.

6.1 Article III: Finding important data subsets for kernel-based bacterial protein role predictions by performing a sensitivity analysis

Article en préparation pour être soumis au Journal *Bioinformatics.*

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–5

Finding important data subsets for kernel-based bacterial protein role predictions by performing a sensitivity analysis

Alain Trubuil [1]*, Véronique Monnet [2] and Liliana López Kleine [1,2]

[1] Unité de Mathématiques et Informatique Appliquées, UR341 INRA, Domaine de Vilvert, F-78350 Jouy en Josas,France
[2]Unite de Biochimie Bactérienne, UR477 INRA, Domaine de Vilvert, F-78350 Jouy en Josas, France

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Genome sequencing has allowed the production of high-throughput genomic data, which in turn has provoked the development of protein role inference methods. Some of these methods allow the use of different kinds of heterogeneous data of different quality which are more or less informative. However, the sensitivity of the data used and the research of important subsets of these data in the inference of protein roles has not been investigated. In this study we aimed to define important subsets of data for the prediction of protein roles in the framework of a kernel method, used to predict the role of bacterial proteins integrating three different types of data: the protein metabolic network, microarray experiments and phylogenetic profiles. This analysis helps also to investigate the quality of the data used and to interprete biological predictions.

Method: We carried out a sensitivity analysis based on a fractionary experimental design to study the influence of different microarray experiments as well as of bacterial orders (groups of famililies) used to construct the phylogenetic profiles.

Results: We did a model selection in order to identify a subset of experiments and bacterial orders which result in the same predictions as when the whole data set is used. The results of this analysis allowed us to understand better which subsets of the original data were important for the predictions and which predictions were more stable. We realized that the subsets of important data were more heterogeneous than expected in our case study. We attributed this result to the fact that the protein we were interested in participated in more than one cellular function.

Availability: Codes and data are accessible upon request.

Contact: alain.trubuil@jouy.inra.fr

1 INTRODUCTION

The sensitivity analysis studies how perturbations on the inputs X of a model M generates perturbations in the output Y. It allows to spread the incertitude on the output between the input variables (factors) when regression and correlation measures are computed on the output (Saltelli *et al.*, 2000).

Two types of data can be considered, continuous and discrete data. The sampling procedures are different in both cases, here we worked with discrete data. The input data can be very numerous and complex. The output data were very simple. The sampling method used to generate the different simulations is the clue of the sensitivity analysis. The best solution would be to test all the possible combinations of input variables and analyse the output. Nevertheless, taking into account that the number of input variables can be very high, the number of possibilities to test increases exponentially and such screening becomes impossible. Methods for sensitivity analysis based on Monte Carlo Sampling have been for example used for the investigation of complex models. The model was treated as a black-box and the distribution of inputs and outputs is analyzed in a global basis (Helton and Breeding., 1993). Other sampling methods can be used which allow a better resolution and assure the determination of factor importance without interactions as for example the sampling method of (Plackett and Burman, 1946). We were concerned with discrete data and chose fractional factorial design (Box *et al.*, 1978) for sampling the space of factors.

The program we analysed is a system that predicts the role of proteins from the lactic acid bacterium *Lactococcus lactis* IL1403 by obtaining distances between all proteins of this organism. We used a Kernel Canonical Correlation Analysis (KCCA) (Yamanishi *et al.*, 2003) for this task including different data types. Nevertheless, we would like to point out that this is a general approach applicable to all entirely sequenced organisms for which enough post-genomic data are available. The predictions of the proteins with the shortest distance to our target protein (the peptidase PepF), were validated experimentally. Most of the predictions were confirmed making of the KCCA a valuable tool for the prediction of protein roles (Lopez Kleine *et al.*, 2008). However, we did not know if a subset of the data (some of the microarray experiments, some of the organisms used to construct the phylogenetic profiles) was important for the predictions and if some of the data were absolutely not informative for our predictions. The approach we propose here is applicable for other kernel methods as for example the em projection method (Tsuda *et al.*, 2003) or the kernel metric learning (Vert and Yamanishi, 2005) as well as for other protein prediction methods.

*to whom correspondence should be addressed

© Oxford University Press 2005.

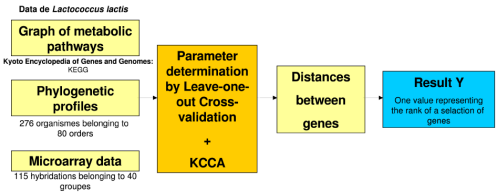


Fig. 1. Representation of the KCCA model used for the prediction of protein roles in this study.

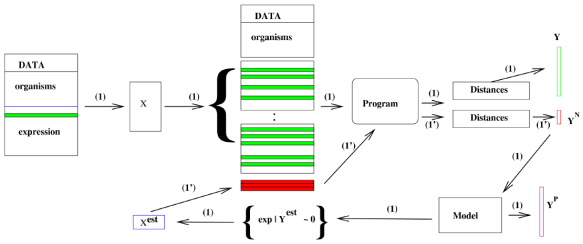


Fig. 2. Framework for the identification of important data subsets. The obtained Model is the emulation of the Program.

We interested ourselves in two data-sets, the microarray experiments and the phylogenetic profiles. Varying the inputs of these two data-sets we were able to better understand the obtained predictions and to put the accent on which data seemed to be useful for understanding the target protein roles. The results of the sensitivity analysis and the search of a simplified model (an emulation of the initial program) allowed us i) to construct acceptable linear models with low residual error, ii) to identify the most important variables in terms of importance in the model and in terms of predictability (i.e. identify a subset of variables that allows us to obtain the same results), iii) to understand better the role prediction we had obtained for the target protein, assessing the robustness of our predictions.

2 METHODS

2.1 Important Subset Identification Framework

Let us consider a collection of objects and a collection of data (that is partitioned in data subsets) with respect to these objects. Let us assume that relationships may exist between these objects but only few of them are known and the relationships of one chosen object (the target protein) has to be deciphered. Hereafter we are interested in highlighting which parts of the data were important to obtain the predicted relationships. Moreover we also consider the robustness of important subsets, this means a stable prediction of relationships even when other data from the data-set are added. The general setting is illustrated in (Figure 1) for biological network reconstruction. Objects are proteins, data are expression and phylogenetic profiles, known metabolic networks, relationships consists in the participation to a same process. Target objects are in the case of our algorithm some proteins for which roles are not known, i.e. PepF (Lopez Kleine *et al.*, 2008).

The space of possible subsets of important data is very huge, 2^p subsets for a collection of p pieces of data! It is not possible to explore all these possibilities. The general framework we propose represented in (Figure 2). It consists in sampling the data space according to an experimental design, constructing a response variable according to inferred roles, emulating the program with the help of a linear model associated to the experimental design and predicting (using a simpler model) the subsets of data that could lead to the reference results (the same inferred roles). Finally, a validation of these subsets is done completing the experimental design and iterating the process until a satisfactory subsets of data (factors) is identified.

2.2 Experimental design and Program Emulation

Let us consider the case of an experiment depending on p factors. We assumed that each factor can take only two values, which was the case in this study. This means some data were chosen within the p microarray experiments (resp organisms) or not. To identify which subsets of factors seem to be important we needed to sample the set ($N = 2^p$ elements) of values for the p factors. Fractional factorial design (Box *et al.*, 1978) allowed an efficient sampling of the space and analysis of the program. By choosing an appropriate design matrix X of size $n \times p$ where n is the number of simulations one can estimate main and interaction effects between factors and model the system with the help of a linear model: $Y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \sum_{i=1}^{p} \sum_{j=1}^{p} \beta_{i,j} x_i x_j + \epsilon$, where ϵ is an error random variable. The coefficients β_0 , $\beta_1, \ldots, \beta_p, \beta_{i,j}$, $i = 1, \ldots, p$ and $j = 1, \ldots, p$ were respectively the mean of the response variable, the main effects and the interactions. A least square estimator of these coefficients was straightforward and given by $\hat{\beta} = (X^T X)^{-1} X^T Y$. In the case of a resolution 4 design, the matrix X is orthogonal and $\hat{\beta} = D^{-1} X^T Y$ where D was diagonal, the main effects were not confounded with interactions of order 2 but there are confusion of several interaction effects or aliasing. We used the FACTEX procedure in SAS to construct our simulation matrix X.

2.3 The program or inference algorithm

The model we analysed is a system that predicts the role of proteins from the lactic acid bacterium *Lactococcus lactis* IL1403 by obtaining distances between the genes of this organism. The smaller the distance between two genes were, higher the chances that the two genes participate in a common metabolic function were. We used a Kernel Canonical Correlation Analysis (KCCA) (Yamanishi *et al.*, 2003) for this task including three different data-types. For the predictions of PepF (Lopez Kleine *et al.*, 2008), five data types were used. Here we left out two of them to simplify the sensitivity analysis and concentrated on metabolic pathways, microarray data and phylogenetic profiles. The three data-types are integrated to finally obtain distances between genes. We did not used this analysis to reconstruct a network containing all the proteins of *L. lactis*, but we interested ourselves in one particular protein. The distances obtained by the KCCA provide valuable information about the predicted role of the chosen protein. Taking into account these distances, the predictions have been validated experimentally for the bacterial peptidase PepF. Most of the proteins predicted to be close to PepF showed to participate in functions in which PepF participates as well. So, its participation in protein secretion, peptidoglycan synthesis and pyruvate metabolism could be confirmed. Only its predicted participation in cell division was not confirmed experimentally. The KCCA showed to be a valuable tool for the prediction of protein roles (Lopez Kleine *et al.*, 2008).

There is a set of parameters associated to the kernels and the KCCA method. Cross-validation was used to estimate those parameters and makes also part of the program. Actually, the algorithm or program we analyzed comprises a KCCA (Yamanishi *et al.*, 2003) used to predict distances between all proteins of an organism. The input variables were the protein metabolic network, the phylogenetic profiles constructed for 276 bacteria belonging to 80 different bacterial orders and microarray data representing the expression of genes in 115 hybridisations belonging to 40 groups of conditions (Figure 1). The groups of conditions were constructed based on repetitions of the same hybridisations and comparisons of the same mutants

obtained under different techniques. Each type of data is represented with a kernel.

2.4 Construction of a Response Variable

Depending on applications the construction of the response variable is more or less straightforward. Here we present the procedure for the predictions of KCCA on the target protein PepF. The value Y was calculated taking into account the position of six proteins SipL, SecA, FtsA, MurB, Glk and AldB. These proteins were chosen because they represent the four of the cellular fonctions in which PepF was predicted to be involved: protein secretion (SipL and SecA), peptidoglycan synthesis (MurB), pyruvate metabolism (Glk and AldB) and cell division (FtsA). Only the implication of PepF in cell division was not confirmed by wet-lab experiments (Lopez Kleine *et al.*, 2008). Using the complete set of microarray data and phylogenetic profiles (reference KCCA), the position of these proteins with respect to PepF were: 1 for SipL, 24 for SecA, 8 for FtsA, 59 for MurB, 9 for Glk, 62 for AldB. For the reference KCCA the Y value was set to zero. For each simulation, Y was calculated as a measure of change in the ranking of these six proteins. It depends on:$Y = \frac{1}{2}\sum_{i=1}^{N} w_i erfc(\frac{S_i - A_i}{\sqrt{2}\sigma})$, where σ is the standard deviation of Gaussian, $N = 6$ is the number of proteins of interest, $A_i, i = 1, \ldots, N$ is the absolute value of the differences in positions between the reference and the program predictions for the N genes and $S_i, i = 1, \ldots, N$ are the tolerances with respect to classification and $w_i, i = 1, \ldots, N$ are the weights.

2.5 Analysis of the output

2.5.1 Linear Regression The sensitivity analysis itself was the analysis of the different Y values obtained by the simulations given in matrix X. Due to the orthogonality of X a handful of models corresponding to combinations of main effects and interaction effects can be estimated simultaneously by solving the equation $\hat{\beta} = D^{-1}X^T Y$ where X is augmented with some cross-products of columns corresponding to the main effects.

As the interactions were confounded, only one interaction representing a group of confounded ones was calculated. For example, for the interaction between factors x1 and x2, X was incremented by a column multiplying the first and second columns of X. In that way, for the microarray data 40 principal factors and 62 factors of interactions were calculated. For the phylogenetic profiles 80 principal factors and 175 factors of interactions were calculated.

We also used the simulations to determine which of the investigated partners changed their ranking in each one of the simulations.

2.5.2 Model Selection At this stage our aim consists in the selection of models having good predictive properties especially in the neighbourhood of $Y = 0$. Moreover, we also look for important subsets of factors or a minimal size sets of factors. With respect to these two goals our strategy consists in first choosing a parsimonious model fitting well the response variable and then looking for the smaller values taken by this model and the corresponding factor values.

Herein first of all we ranked the coefficients according to their module and consider models with increasing number of terms or coefficients. Computing standardized error and adjusted R^2, a few numbers of models may be selected for further investigation. A more powerful model selection can also be employed when response variable distribution is Gaussian (Zahn, 1975) or mixture of Gaussian (Box *et al.*, 1993; Adamczyk, 1997). The next stage of our procedure consists in still trying to reduce the number of factors involved in the model for further evaluation and identification of minimal size sets. With respect to this goal we use the aliasing structure for selecting inside the family of models the main and interaction factors that involve least factors. This selection turns out to be an NP-hard problem and we proposed an algorithm of low complexity providing an approximate solution, explained in the following section. At this point we got a family M of models containing few members.

2.5.3 Minimal or Important subsets construction Depending on the number of factors of the models in M, either an exhaustive or a sampled evaluation of the models can be made. As the exploration of the model implicated the test of a huge number of combinations we sampled the space M of possible models in several steps: i) we observed that a given number of $f < F$ terms (F being the original number of factors), principal and of interactions, were necessary to obtained a good model, ii) we explored these terms keeping for the model all terms corresponding to principal factors, iii) then we explored the interaction terms eliminating the terms containing factors already kept as principal factors in step ii), iv) we explored the space of interaction factors left out until now, giving them tag with the goal to find the most parsimonious model, i.e. containing the minimal number of factors. Then we made a second selection on the basis of the output value predicted by the obtained subset, keeping only the subsets that gave a predicted value of Y near to zero.

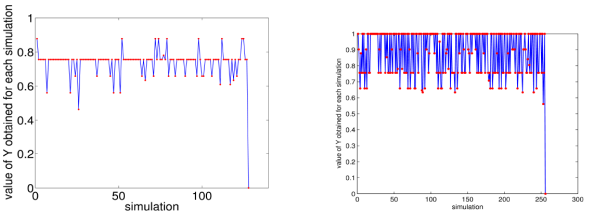
Each of these minimal important subsets of factors has now to be confirmed. The subset of factors is now considered as an input to the program and variable response is computed. This includes the determination of parameters by cross-validation and the predictions. Either the response variable is significantly non zero, and the minimal subset is rejected, either the program and model prediction were close and the minimal important subset is retained.

2.5.4 Robustness of the Minimal subset Once an important subset retained, it is useful to know if the addition of variables (experiments or organisms in our case), not making part of the subset, changes the prediction result. To investigate this point we added at random 3, 6 and 9 variables to the minimal subset 15 times and repeated this procedure 5 times, which gave us 225 simulations to test.

3 RESULTS

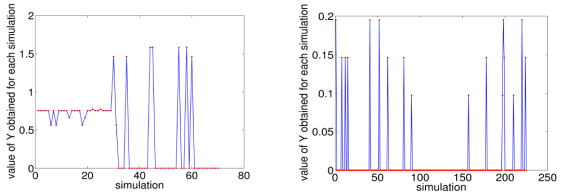
The framework has been applied to deciphering which subsets of data seem to be important for PepF role predictions. Being sure that our multiple simulations present a low error and that the parameters calculated for each one were optimal we calculated the response value Y for each simulation (Figure 3.a), Y=0, being the reference value for the last simulation in which the complete data-set was used. One simulation corresponds to one line of X (Figure 2). None of the other simulations lead to $Y = 0$. We obtained models with a low global residual error (0.0116 for the microarray data and 0.0078 for the phylogenetic profiles). Interactions seem very important and the residual errors for a model without considering them were higher but still acceptable (0.0938 and 0.1083 respectively). Several interaction effects were of the same order as main effects for both the model on microarray data and the model on phylogenetic profiles. In the left column of tables 1 and 2 the three more important principal factors of the linear models constructed for both types of data are shown.

3.0.5 Research of a simplified model Ranking of model coefficients and selecting factors considering the highest coefficients (crude method) even with the complete aliasing structure gave very poor results when used for prediction (e.g. output Y-value of 0.7073 with 33 factors of the microarray data!). For this reason we used the framework presented before to find a minimal subset of data. We obtained subsets of factors predicted to give an output Y as low as possible. We tested the ten first subsets kept after the selection procedure for 10, 15, 20, 25, 30, 35 and 40 members to see if the predicted output of Y was in fact low. The values of Y are plotted in (Figure4.a). This figure shows that several subsets containing around 30 factors give a result of Y=0 for the microarray data (20



(a) (b)

Fig. 3. Output Y of the simulations for microarray data (a) and phylogenetic profiles (b) indicated by red points. Blue lines between points added for visibility.



(a) (b)

Fig. 4. Output Y of the simulations for microarray data (a) and phylogenetic profiles (b) during the simulations done for the research of minimal subsets indicated by red points. Blue lines between points added for visibility.

factors for the phylogenetic profiles. Keeping only the factors that contribute to an output of Y=0 allowed us to reduce the factors to 18 (15 for the phylogenic profiles. The solution is not unique, but several factors are present in the majority of the subsets. Three of the factors (microarray experiments or orders of organisms) present in most of the minimal subsets are listed in table 1 for the microarray data and table 2 for the phylogenetic profiles. It is worth to notice that both types of data behave in a different way, as for the microarray data more factors are necessary to obtain an output value near to Y=0 than for the phylogenetic profiles.

Factor LM	Experiments	Factors Subset	Experiments
25	sup pyrF	25	sup pyrF
22	natural strain	3	codY
36	pip mutant	7	pip mutant

Table 1. Microarray data: Factors in order of importance (ordered coefficients of the linear model) and five factors making part of the important subsets (not ordered)

3.0.6 Robustness of the relationships of PepF For the six predicted partners of PepF we chose to evaluate their ranking across the different simulations to determine the robustness of the predictions. First we interested ourselves in the changes of ranking during the 128 simulations for microarray data (256 for phylogenetic profiles). In the first two columns of table 3 the number of simulations in which the ranking changes is shown. These results indicate that some predictions (i.e. SipL) are verified in more simulations than others (i.e. FtsA).

Factors LM	Order	Factors Subset	Order
16	E	16	E
41	H	41	H
31	L	62	S

Table 2. Phylogenetic profiles: Factors in order of importance (ordered coefficients of the linear model; LM) and five factors making part of the important subsets (not ordered). E: Enterobacteriales, H: Halobacteriales, L: Lactobacillales, S: Sphingomonadales

3.0.7 Perturbation of the important subsets of factors Once we identified 15 subsets of factors which gave an Y value of zero, we wondered if there are microarray experiments that could destabilize these results. We added 3, 6 and 9 microarray experiments chosen at random to the 15 simulations identified to give Y=0. This was done 5 times for each factor giving us 225 new simulations to test. The change in the ranking of the six partners of PepF investigated are shown in the last two columns of table 3. They show that the addition of new experiences does not destabilize the results and thus does not add noise to the predictions.

	SimMD	SimPP	T225MD	T225PP
SipL	47/128	68/256	0	7
SecA	63/128	138/256	4	16
FtsA	120/128	240/256	8	20
MurB	62/128	184/256	0	11
AldB	64/128	190/256	4	12
Glk	68/128	97/256	10	29

Table 3. Number of original simulations planned by (X) in which the ranking of the protein changed (Sim) and ranking changes in the Tolerance test (T). MD: microarray data, PP: Phylogenetic profiles

3.0.8 Finding important subsets for one predicted partner of PepF To construct the linear model presented before as well as to find the subset of important factors we investigated the result in rank changes of six proteins (SipL, SecA, FtsA, MurB, AldB, Glk). We wanted to know also if the model obtained was simpler if only the changes in rank of one protein (this means also one function) were investigated and if we could detect which microarray experiment and which bacterial order gave us the information about the link to the investigated protein. For this purpose we concentrated on the change in rank of SipL, the partner predicted to be the closest to PepF. We obtained a simpler model, in which the coefficient of factors decreases faster than in the other first model showing that some factors are very important (results not shown). We found again the microarray experiments 25 (sup pyrF) and 22 (natural strain) as being very informative for the relationship with SipL. The most important order

turned out to be again Enterococcales, which is not the bacterial order *L. lactis* belongs to.

4 DISCUSSION

The analyzed model is very complex because many variables and parameters are brought into play and the output is a strong simplification of the results. First of all, we simplify by investigating only the output concenring one protein, PepF, which is represented by one vector of the global output matrix of the KCCA. Then we simplified by taking into account the ranking of only some proteins in this output vector. A further simplification is done by the construction of one scalar that represents the ranking of the partners of PepF in the output vector in respect to the reference ranking. The choices to do this ranking, i.e. the thresholds to declare a ranking change as being important and the weight we give to the change are arbitrary and, if chosen in a different manner, can change the obtained model. Actually, the results obtained, this means, the identification of important microarray experiments and of orders of bacteria are only valid for one protein and some partners.

Taking into account this simplification it is not surprising that the obtained models are very complex and contain many interactions and that it is very difficult to assign one or two principal factors that explain the results and discard the rest. As microarray experiments are done in different conditions that affect only a small group of genes it is not surprising to find a high importance of interactions.

Regarding principal factors without interactions in the case of the microarray data, the experiences showing the highest coefficients come mainly from the microarray data of the European project Express Fingerprints. The most important experience group turned out to be the comparison of a pyrF mutant to a genetically modified organism (GMO) that contains a suppressor of this mutation on a plasmid (Ueda *et al*., 2006). For this and the other important microarray experiments it is very difficult to interpret their meaning in the prediction of PepF roles. As we are analyzing a pleiotropic enzyme it is not surprising to find that many different conditions are needed to explain is relationships with proteins belonging to different experiments. Nevertheless, knowing which experiments are important can help in other cases to point out interesting experiments. If only one partner and thus one function is investigated, the model becomes simpler and smaller subsets of factors are necessary to obtain reference predictions. To analyze a group of partners and then one by one can also help to point out which factor is responsible for each prediction. We know, for example, that the microarray data making part of factors 25 (sup pyrF) and 22 (natural strain) is responsible for the prediction of SipL as a partner of PepF.

In the case of the phylogenetic profiles, the most important orders belong in fact to the evolutionary more similar orders of the Lactobacillales. Nevertheless, the first order is not the order of *L. lactis* but of an close order: Enterococcales.

It is not astonishing to observe that the subset of factors can be more easily reduced for the orders of the phylogenetic profiles than for the microarray data, because these experiments are more heterogeneous than the co-evolution of proteins among the different orders of bacteria.

The fact that in most of the simulations the chosen partners for PepF did not change drastically their ranking reinforces the obtained predictions. The protein that changes its ranking most of the time is FtsA, for which the predicted relationship was not confirmed, which is not surprising. On the other hand the proteins that changes their ranking only in some simulations are the proteins implicated in secretion which was determined to be the main function of PepF. It was also interesting to observe that the stability of the prediction is not related to the position as SipL, FtsA and Glk, all belonging to the first 10 postions have a completely different behaviour.

We confirmed also the fact of the robustness of the model by the fact that adding experiments to a group of experiments that gives a good value of Y does not affect the prediction. This means that a subgroup of experiments is enough to obtain the results, but that the use of more experiments does not change the predictions.

5 CONCLUSION

The sensitivity analysis of a complex approach as the kernel method we used to predict protein roles is a useful approach to determine a minimal subset of data that still allows to obtain the reference results, which simplifies the model and the interpretation of the results. In our case the sensitivity analysis was done after the experimental validation of predicted results, nevertheless, this analysis could have been useful at the moment were biological hypothesis were posed, because the importance of certain relationships and input data is highlighted.

FUNDING

This work has been financed by the Institut National de la Recherche Agronomique (INRA).

ACKNOWLEDGEMENT

The authors would like to thank Jean-Pierre Gauchi and Hervé Monod for the construction of the simulation matrix and helpful discussions on sensitivity analysis.

REFERENCES

Adamczyk, K. (1997) Analyse des Expériences Factorielles sans Répétition, *PhD Thesis, Univ. Orsay.*

Box, G.E.P., Hunter, W.G., Hunter, J.S. (1978) 'Statistics for Experimenters, an Introduction to Design, Data Analysis and Model Building' *Wiley*

Box, G.E.P., Meyer, R.D.(1993) Finding the Active Factors in Fractionated Screening Experiments, *Journal of Quality Technology* **25**, 94-105.

Ueda, K., Yamamoto, Y., Ogawa, K. (2006) Bacterial SsrA system plays a role in coping with unwanted translational readthrough caused by suppressor tRNAs . *Genes to Cells* **7**, 509-519.

Zahn, D.A. (1975) Modifications of and Revised Critical Values for the Half-normal Plot, *Technometrics* **17**, 189-200.

Yamanishi, Y., Vert J.P., Nakaya A. and Kaneisha M. (2003) Extraction of correlated clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics* 19: Suppl. 1, i323-i330

Lopez Kleine, L., Monnet, V., Pechoux, C., Trubuil, A. (2008) Role of the bacterial peptidase F inferred by statistical analysis and further experimental validation. *HFSP Journal* 2: 29-41.

Saltelli, A., Chan, K., Scott E.M. (2000) 'Sensitivity Analysis' *Wiley*

Helton, J.C., Breeding, R.J. (1993) Calculation of reactor accident safety goals,*Reliab. Eng. System Safety*, **39**, 129-158.

Plackett R.L., Burman, J.P. (1946) The Design of Optimum Multifractional Experiments, *Biometrika* **33**, 305-325.

Tsuda K., Akaho S., Asai K. (2003) The em algorithm for kernel matrix completion with auxiliary data, *J. Mach. Learn. Res.* **4**, 67-81.

Vert J.P., Yamanishi Y.(2005) Supervised Graph inference. In Saul L. K., Weiss Y., Bottou L., editors, *Adv. Neural Inform. Process. Syst.* **17**, 1433-1440. MIT Press,

6.1.1 Données supplémentaires de l'article III

Crossvalidation and Predictions
INPUT:
Laplacian matrix of the protein metabolic network *L* phylogenetic profiles (trianing set), microarray data (training set), phylogenetic profiles (candidate proteins), microarray data (candidate proteins), groups of microarray data or orders of bacteria, simulation matrix *X*, the number of the simulation *n*

Crossvalidation to determine best parameter values

```
for t=t_min:t_max
    for dRBF=dRBF_min:dRBF_max
        for delta=delta_min:delta_max
            while countloo< size L
            Prediction of partners for one gene (predKCCA.m)
            Error of the prediction (erreurRecons.m)
            end
        end
    end
end
```

Predictions
Creates new datasets corresponding to the simulations.
Calculates distances between all genes (Distances.m).
Output: Distances between all genes

KCCA
Input: data for the training, data for candidate proteins, parameters
Computes diffusion kernel (diffusion.m)
Kernel for phylogenetic profiles (Kphylo.m)
Kernel for microarray data (Ktranscriptome.m)
KCCA (kcca.m)

OUTPUT:
Projections of proteins on the linear combinations found by KCCA

Construction of the Output (prepare_sortie.m)
INPUT:
index of protein to be analyzed, partners to investigate, thresholds for the rank change of partners, weights for each partner: $y = \frac{1}{2}\sum_{i=1}^{L} w_i erfc(h(S_i - V_i))$
OUTPUT: Y vector

Analysis of the Output (anal_effet_melanges.m)
INPUT: simulation matrix *X*, Output vector *Y*
linear model for principal factors: $Y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \sum_{i=1}^{p}\sum_{j=1}^{p}\beta_{i,j} x_i x_j + \in$
reads aliasing structure of interactions
linear model of interactions
R2
error
plots
OUTPUT:

effect of factors, R2, error

Data
laplacian matrix of the protein metabolic network: metatout.txt
phylogenetic profiles training set: proftout.txt
microarray data (training set): expretout.txt
phylogenetic profiles (candidate proteins): profcan.txt
microarray data (candidate proteins): exprecan.txt
groups of microarray data or orders of bacteria: groupes_hybridations.txt or ordres_organismes.txt
simulation matrix *X*: fraction40ready.txt or ready80.txt

6.2 Résultats de l'analyse de sensibilité sur YvjB

La démarche pour étudier les prédictions de YvjB par analyse de sensibilité a été la même que pour PepF. Les partenaires étudiés ont été Ffh, la protéine qui fait partie de la SRP, deux protéines participant dans la traduction, PrfA et RplE, deux protéines participant dans la maturation de protéines, Lgt et LspA, ainsi que YhhG, un homologue de LrgB, participant dans la régulation des muréine-hydrolases. Un rôle de YvjB dans la régulation des muréine-hydrolases n'a pas été confirmé expérimentalement.

Les facteurs avec des effets élevés et les facteurs faisant partie du sous-ensemble prédit ne se recouvrent pas dans le cas de YvjB. De plus les effets d'interactions semblent très importants (Tableau 6-1). Ainsi, parmi les cinq premiers facteurs nous en retrouvons trois correspondant aux interactions. Cela veut dire que plusieurs expériences pourraient être potentiellement impliquées, puisque les facteurs des interactions sont calculés par paquets (aliasing). Le sous-ensemble de facteurs qui nous a permis de retrouver des résultats semblables aux résultats de référence contiennent plusieurs expériences faites sur le mutant AldB, protéine impliquée dans le métabolisme du pyruvate, le mutant Pip et une souche naturelle. Il s'agit, comme pour le cas de PepF d'expériences très diverses.

Tableau 6-1 : Facteurs (correspondant aux expériences transcriptomiques) avec des effets importants dans le modèle linéaire (ML) et facteurs qui font partie du sous-ensemble essentiels qui permet de retrouver les résultats de référence.

# Facteur ML	Expérience transcriptomique	# Facteur sous-ensemble	Expérience transcr.
22	Souche naturelle	10	Souche naturelle
40	Trehalose vs. Trehalose-Fructose	11	Mutant Pip
89	Interactions	14	Mutant AldB
48	Interactions	12	Mutant AldB
49	Interactions	15	Mutant AldB

Nous avons également étudié les profils phylogénétiques pour YvjB. Les prédictions pour cette protéine s'expliquent plus par les organismes phylogénétiquement proches, comme les Lactobacillales (Tableau 6-2). Cela est vrai pour le choix du modèle linéaire mais aussi pour les facteurs qui font parti des sous-ensembles importants pour retrouver les résultats de référence.

Tableau 6-2 : Facteurs (correspondant aux ordres des profils phylogénétiques) avec des effets importants dans le modèle linéaire et facteurs qui font partie du sous-ensemble qui permet de retrouver les résultats de référence.

Facteur ML	Ordre	Facteur sous-ensemble	Ordre
31	Lactobacillales	31	Lactobacillales
16	Enterococcales	16	Enterococcales
10	Bacillales	10	Bacillales
69	Xanthomonadales	69	Xanthomonadales
14	Spirochaetales	13	Burkholderiales

Les partenaires qui changent au cours de différentes simulations ont aussi été étudiés pour le cas de YvjB. Les résultats sont moins parlants que pour le cas de PepF, néanmoins, Ffh, la protéine prédite comme étant la plus proche de YvjB est dans cette première position lors de la plupart de simulations (Tableau 6-3).

Tableau 6-3 : Nombre de simulations originales (Sim) dans lesquelles la position des protéines étudiées change de rang par rapport au classement de référence réalisé avec la totalité des données. Nombre de simulations du test de tolérance (T) dans lesquelles la position des protéines étudiées change. T : Expériences Transcriptomiques. PP : Profils phylogénétiques.

	SimMT	SimPP	T225T	T225PP
Ffh	56/128	78/256	5/225	7/225
PrfA	78/128	95/256	8/225	11/225
RplE	76/128	99/256	9/225	6/225
YhhG	64/128	121/256	15/225	17/225
Lgt	90/128	145/256	22/225	34/225
LspA	97/128	157/256	18/225	17/225

7 Matériels et Méthodes

Compte tenu de la grande quantité de protocoles utilisés et de la rédaction des articles, le but de cette partie n'est pas de donner des détails sur les protocoles mais le minimum d'information pour pouvoir comprendre comment les expériences ont été faites. La présentation des protocoles de microbiologie, biologie moléculaire et biochimie se fera dans les sections 7.1 à 7.4. Ensuite, seront présentés les différents logiciels utilisés (section 7.5), des détails sur la mise en forme des données (7.5.3), ainsi que la manière dont le KCCA et l'analyse de sensibilité ont été adaptés et mis en place lors de ce travail.

Les différentes souches, plasmides et amorces utilisés au cour de ce travail sont listés dans les tableaux 7-1, 7-2 et 7-3.

Tableau 7-1 : Souches utilisées au cours de ce travail. *Mutants complémentés avec le plasmide pILN13 en faible nombre de copies.

Souche	Plasmide	Résistance	Référence
L. lactis IL1403	-	-	(Chopin *et al*, 1984)
E. coli TG1 repA+	pGhost9-*pepF* inactivé (pTIL 120)	Ery	(Nardi *et al.*, 1997)
L. lactis IL1404 Δ*pepF**	pILN13-*pepF*	-	Ce travail
L. lactis NZ9000-pSEC1	pSEC1	Cm	(De Ruyter *et al.*, 1996) (Chatel *et al.*, 2001)
L. lactis NZ9000 Δ*pepF**	pSEC1, pILN13-*pepF*	Cm	Ce travail
E. coli BL21 (DE3) Gold	-	-	Stratagen
E. coli BL21 (DE3) Gold	pET28-*pepF*	Km	Ce travail
L. lactis NZ9000 pVE5547	-	-	UBLO
L. lactis IL1404 Δ*yvjB*(Eep)*	pILN13-*yvjB*	-	Ce travail
L. lactis IL1404 Δ*optA*	-	-	UBLO
L. lactis IL1404 OptA_ps YvdF	-	-	Ce travail
L. lactis NZ9000-pVE5547	pVE5547	-	Ce travail
L. lactis NZ9000 Δ*pepF**-pVE5547	pVE5547	-	Ce travail

Tableau 7-2: Plasmides utilisés

Plasmide	Caractéristiques	Résistance	Référence
pTIL 120	pGhost 9-pepF inactivé	Ery	(Nardi *et al.*, 1997)
pSEC1	expression sous P*nisA* codant SPUsp:NucB	Cm	(Chatel *et al.*, 2001)
piLN13	haut nombre de copies, permet de passer à un faible nombre de copies par restriction avec KpnI	Ery	(Renault *et al.*, 1996)
pet28	conceived for heterologous protein production	Km	Novagen
pVE5547	expression sous P*nisA* codant pour SPUsp:NucA:ancrage-LPTX-protM (substrat de la sortase)	Ery	UBLO, J.C. Piard
pYvjB	pGhost 9-pour déléter *yvjB*	Ery	ce travail
pPepF	pET28 contenant PepF-6His	Km	ce travail
pOptA-ps-YvdF	pGhost9 contenant *optA* avec le peptide signal de YvdF	Ery	ce travail

Tableau 7-3: Amorces utilisés dans ce travail

Primer name	Sequence
FpepF*	GCGGATATTAAGTTACCTATGGT
RpepF*	TTTGGCAATTACTTCTAAAGGAT
PetPepF-For	CATGCCATGGTTGCTAAGAATAGAAATGAAAT
PetPepF-Rev	GGAAGATCTAAGATGGACTCCTTTTTCAA
eepUF	CTCCTCGAGAGAACATCTGGTGAACAAAG
eepUR	GAAGAATTCTTAGAATAGTCCAAGTAGATGC
eepDF	GAAGAATTCATCTATTCGTCTTTCTATTTC
eepDR	AACTACTAGTTTCAGTATAAGCTACATTTG
Feep*	GTAAAAGATTCAGGTAAAAT
Reep*	GAAATAGAAGACGAATAGAT
UoptSF	CTTTACACCAGTGGGAATGAG
UyvdFi-UoptSR	TTGCAAGAGTCAACAATTTTATTTTCTTCATGTAATTTCCCTC
LoptAR	CATACCCGCTGGTGTTAAGC
UyvdF-LopAF	AAATAAAATTGTTGACTCTTGCAA TTGCTTGCAGCATGCGG
LyvdFi-UyvdFi	GCCAATTGAGGCAATTCCAACTGTTGCAAGAGTCAACAATTTTATTT
LyvdF-LoptAF	CAGTTGGAATTGCCTCAATTGGC TTGCTTGCAGCATGCGG
opta-seqF	GTGAAAACTTTCCGCTATTT
opta-seqR	AAGTATGGTGTTGGTTGTGC
lspAUF	CTCCTCGAGATAAACTTCGTGCCTACTCA
lspAUR	GATCGGATCCTTATTTTTCTGTATCTCCAAGCT
lspALF	GATCGGATCCAATTGATTAAAAAATCACTG
lspALR	CCCGGGTAAATACAAAAAAAACAAGT
lspAF*	TATATGTTTAGGGACTGTCA
lspAR*	TACTTTGCTTTTTTTGCTTC

*vérification de la délétion et séquençage de la région contentant la délétion

7.1 Constructions génétiques

7.1.1 Construction du mutant *pepF* chez *L. lactis* IL1403 et *L. lactis* NZ9000

Le mutant *pepF* a été construit utilisant la construction réalisée par (Nardi *et al.*, 1997), pour la souche *L. lactis* NCDO763. Cette construction permet suite à une double recombinaison d'enlever 161 paires de bases du gène *pepF* qui incluent le site actif de PepF.

7.1.2 Construction d'une souche surproductrice de PepF et des mutants complémentés

La souche surproductrice de PepF a été obtenu en insérant le gène *pepF* avec son propre promoteur dans le plasmide pILN13 (Renault *et al.*, 1996). Ce plasmide est un plasmide à haut nombre de copies qui permet, suite à une digestion par l'enzyme de restriction KpnI, un switch à faible nombre de copies. La même construction a donc pu être utilisée pour complémenter les mutants *pepF*.

7.1.3 Construction du mutant *yvjB (eep)* et sa complémentation

Pour la délétion du gène *eep* entier, sans affecter les gènes en amont et en aval, la construction suivante (*Figure 7-1***) a été réalisé dans le plasmide pGHOST9 (voir**

Tableau 7-3) pour la séquence des amorces):

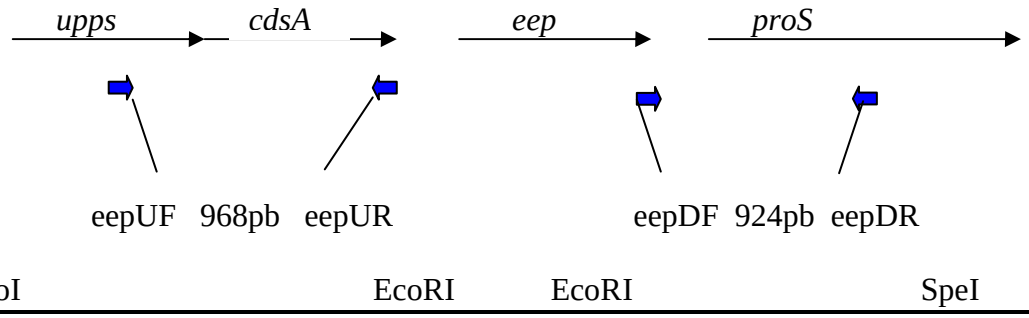


Figure 7-1 : Schéma de la construction du mutant eep.

Pour la complémentation du mutant *eep*, le plasmide pILN13 a été utilisé dans sa version faible nombre de copies. Le mutant *eep* a été complémenté par l'opéron *uppS-cdsA-eep* et son propre promoteur.

7.1.4 Construction du mutant *lspA*

Pour la délétion du gène *lspA* entier, sans affecter les gènes en amont et en aval, la construction suivante (Figure 7-2) a été réalisé dans le plasmide pGHOST9 (voir

Tableau 7-3 pour la séquence des amorces):

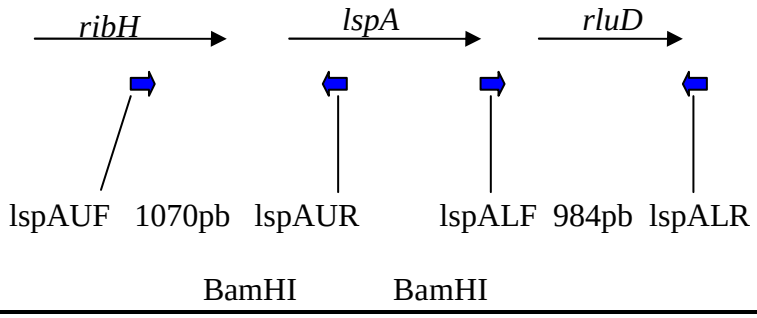


Figure 7-2 : Schéma de la construction du mutant lspA

7.1.5 Construction d'un gène *optA* contenant le peptide signal de *yvdF*

Le but de cette construction a été de créer un fragment du gène *optA* contenant le peptide signal de *yvdF* pour étudier le comportement de *optA* dans un mutant *eep.* Le fragment du peptide signal a été inséré par PCR selon la technique de (Ho *et al.*, 1989). Cette technique permet de réaliser une PCR sur deux produits de PCR si une extrémité des produits est complémentaire et s'aligne à la même température que les amorces utilisés pour la PCR (***Figure 7-1***).

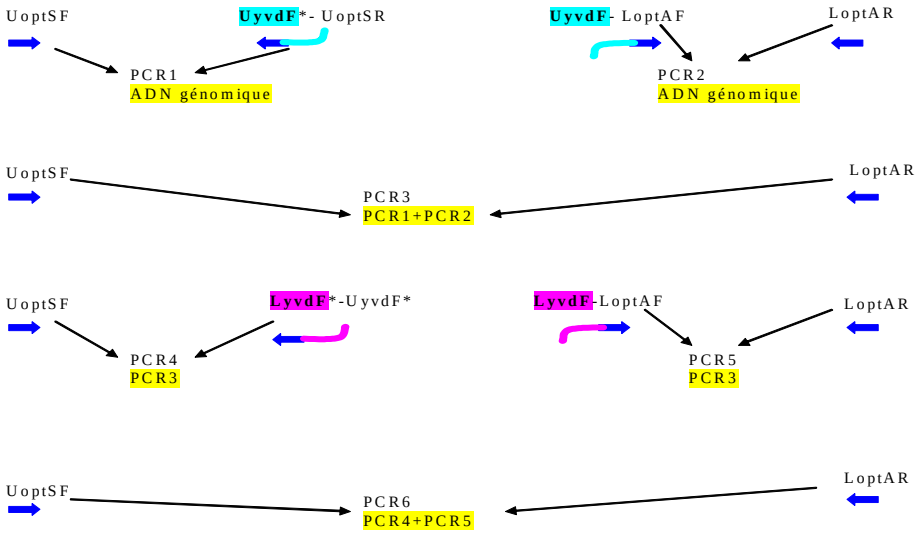


Figure 7-3 : *Schéma des PCR successives réalisées pour insérer le peptide signal de yvdF.*

Dans un premier temps la partie UyvdF a été insérée. Une fois ce fragment obtenu, la partie LyvdF a été insérée par une deuxième étape de PCR (**Figure 7-3** et **Figure 7-4**).

UoptSR UyvdF LyvdF LoptAF

M K T W K K V T L G T V A L G S A A L L A A C

GAG GGA AAT TAC ATG AAG ACT TGG AAA AAA GTT ACC TTA GGT ACA GGT GCA CTT GGT TCG GCA GCT TTG CTT GCA GCA TGC GGC

ATG AAA AAA ATA AAA TTG TTG ACT CTT GCA ACA GTT GGA ATT GCC TCA ATT GGC TTA CTT GCT GCT TGC

M K K I K L L T L A T V G I A S I G L L A A C

Figure 7-4 : *Séquence du peptide signal de optA (en rouge) et du peptide signal de yvdF qui illustre le choix des amorces.*

7.2 Protocoles de biologie moléculaire

Les techniques de biologie moléculaire ont été utilisées selon les protocoles de (Sambrook & Russel, 2001).

7.2.1 Préparation d'ADN plasmidique

Pour la préparation d'ADN plasmidique, le kit commercial Qiagen a été utilisé. Pour l'extraction à partir de cultures de *L. lactis*, du lysozyme (10 mg/ml) est additionné à la solution de resuspension et les cellules sont incubées 30 min à 37°C avant l'étape de lyse. Cette étape de lyse n'est pas réalisée pour l'extraction d'ADN plasmidique à partir de *E. coli.*

7.2.2 Electrophorèse d'ADN

Les électrophorèses ont été réalisées sur des gels d'agarose (SeaKem LE) 0,7% pour vérification et agarose (SeaKem GTG) 0,7% pour purification. Le gel a été préparé dans du tampon du migration TAE 0,5%. 5 µL d'échantillon sont mélangés avec 5µL de bleu de charge l'ADN migre à 100V pendant 30 minutes. Pour une purification d'ADN, la totalité de l'échantillon migre à 50V pendant 4 heures; pour un transfert d'ADN sur membrane 50µL d'échantillon migrent à 15V pendant 13 heures.

7.2.3 Purification d'un fragment d'ADN à partir d'un gel

Après séparation sur gel d'agarose, les fragments d'ADN ont été purifiés à l'aide d'un kit commercial Qiagen qui utilise le principe d'adsorption spécifique de l'ADN sur une membrane de silice après dissolution de l'agarose à 50°C, les fragments ont été purifiés avant les ligations.

7.2.4 Préparation de cellules électrocompétentes de *Lactococcus lactis*

La technique est basée sur une fragilisation de la paroi de *L. lactis* par la glycine ou la thréonine. Les deux acides aminés ont été testés et les meilleurs taux de compétence (10^6) ont été obtenus avec la thréonine pour *L. lactis* IL1403 et NZ9000. 400 mL de M17-sacharose 0.5M et thréonine 2% ont été ensemencés au 1/100 à partir d'une pré-culture de la souche en M17-glucose. A une DO_{600nm} de 0,6 la culture a été arrêtée et les cellules ont été lavées trois fois dans un tampon saccharose-0,5M-Glycérol-10%. Elles ont été finalement reprises dans le même tampon, congelées rapidement dans l'azote liquide et conservées à -80°C avant la transformation.

7.2.5 Préparation de cellules thermocompétentes de *E. coli*

100 mL de LB + Mg avec ou sans antibiotique selon la souche ont été ensemencés. La culture, arrêtée à une DO_{600nm} de 0,6 a été incubée dans du tampon d'acétate de potassium 30mM final, KCl 100mM final, $CaCl_2$ 10mM final, glycérol 15 %, $MnCl_2$ 50mM final (ajusté à pH 5,8 avec l'acide acétique) pendant 30 minutes. La culture a été lavée et incubée dans du tampon contenant du MOPS 10mM final, KCl 10mM, CaCl2 75mM et glycérol 15 % final (ajusté à pH 6,8) pendant 15 minutes. Les cellules sont congelées dans l'azote liquiede avant la conservation à -80°C.

7.2.6 Transformation par électroporation ou choc thermique

Avant la transformation, les cellules compétentes ont été incubées avec l'ADN plasmidique dans la glace pendant 30 minutes. Pour le lactocoque en présence de 100 à 200 µg d'ADN plasmidique les bactéries électrocompétentes sont soumises à un choc électrique de (200Ω , 25µF et 2.5KV) qui rend transitoirement perméable la membrane plasmique, et qui permet la transformation de ces cellules par un plasmide. Pour *E. coli,* les bactéries sont soumise pendant 90 secondes à un choc thermique à 42°C. Après transformation, les bactéries sont incubées en milieu riche non sélectif pour permettre l'expression du gène de sélection puis étalées sur milieu sélectif ne permettant la croissance que des bactéries possédant le plasmide transféré (la sélection est une résistance à un

antibiotique).

7.2.7 Préparation d'ADN génomique *L. lactis*

Nous avons utilisé le protocole de (Hoffman & Winston, 1987). Les cellules sont lysées dans du tampon de lyse (2% TritonX100 ; 1% SDS ; 100mMNaCl, 10mM Tris/HCl pH8, 1mM EDTA pH8) par cassage mécanique aux billes de verre. Les débris cellulaires et les protéines sont séparés de l'ADN par une extraction au phénol. L'ADN total est précipité à l'éthanol et récupéré avec une pipette Pasteur coudée.
Selon l'utilisation prévue de l'ADN, un traitement RNAse à une concentration de 100µg/mL à été fait.

7.2.8 Ligation

Les ligations ont été effectuées par l'ADN-ligase du bactériophage T4 utilisant le kit Fast-link ligation (Epicentre) avec une incubation de la réaction de ligation d'une heure, puis inactivation à 70°C durant 15 minutes. Les ligations ont été utilisées le jour même pour transformer des bactéries compétentes. Les ligations intermédiaires des fragments PCR ont été réalisées dans le vecteur pGEMT (Promega). La ligation dans ce vecteur utilise les TA débordants produits lors de la PCR.

7.2.9 Digestion

Les enzymes de restriction et tampons utilisés proviennent de New England Biolabs. Les digestions ont été réalisées dans un volume final de 50 µL avec 2 µL d'enzyme. Dans le cas de deux réactions de restriction successives, la deuxième a été réalisé après vérification de la première et purification dans le cas où la deuxième enzyme n'utilise pas le même tampon que la première.

7.2.10 Southern blot

Après digestion enzymatique, les fragments d'ADN sont séparés sur gel d'agarose à 0,7% par l'application d'un voltage de 20V pendant 14 heures. Le gel est placé sous agitation pendant 15 minutes dans une solution de HCl 0,12 N pour dépurination, rincé deux fois à l'eau, puis l'ADN subit une dénaturation dans la solution I (NaOH 0,5 M; NaCl 5 M) pendant 40 minutes. Après deux rinçages à l'eau, l'ADN est neutralisé par 2 bains de 20 minutes dans la solution II (Tris 0,5 M; NaCl 3 M; pH 7,2). Le transfert de l'ADN sur une membrane de nitrocellulose est réalisé de manière active sous vide utilisant une VacuGene Pomp (Amersham Biosciences). Après transfert, l'ADN est fixé de façon covalente au support par traitement aux UV (Stratalinker). Le kit ECL (Direct nucleic acid labeling and detection system) Amersham Biosciences a été utilisé pour la hybridation et la révélation. La membrane est placée à 42°C, pendant 1 à 4 heures, sous agitation, dans 50 ml d'une solution de pré-hybridation ECL. Pour l'hybridation, la sonde marquée avec le réactif de marquage ECL est dénaturée 5 minutes à 100°C puis directement rajoutée dans le milieu de pré-hybridation. L'ensemble est placé sous agitation, à 42°C, pendant une nuit. La membrane est lavée avec du tampon urée (urée 360 g/L, SDS 0,4%, SSC 0,5%) deux fois et une fois avec du

tampon SSC 2X durant 20 minutes. La membrane est exposée face à un film autoradiographique. Le temps d'exposition est d'environ 5 minutes et peut varier selon l'intensité du signal.

7.2.11 Quantification de l'ADN et des plasmides sur gel d'agarose ou par spectrométrie

L'ADN est quantifié après une dilution au 1/200ème dans l'eau et la DO est mesurée à 260 nm, 280 nm et 310 nm. La concentration en ADN est calculée sachant qu'une DO_{260nm}=1 correspond à une concentration d'ADN double brin de 50µg/mL. La pureté des ADNs est vérifié sachant que l'ADN pur se caractérise par : $1.8<DO_{260}/DO_{280}<2$ et DO_{310} faible (de l'ordre de 0.001). L'ADN plasmidique linéarisé par une enzyme de restriction appropriée ou des fragments linéaires provenant des PCR ont été quantifiés approximativement sur gel d'agarose par comparaison à un marqueur composé de fragments d'ADN de concentration connue (Smart Ladder, Eurogentec).

7.2.12 Réaction en chaîne de la polymérase (PCR)

L'ADN polymérase MP (Qbiogene) ou le TripleMaster PCR system (Eppendorf) ont été utilisées pour les PCR dans les conditions suivantes : 100 ng d'ADN, 100 pmol de chaque oligonucléotide, 200 nM de désoxyribonucléotides (dATP, dCTP, dGTP et dTTP), tampon du fournisseur et 1 unité d'ADN polymérase. Le programme comporte une dénaturation de 5 minutes à 95°C. L'amplification se compose de 30 à 35 cycles de: i) dénaturation de l'ADN, 30 secondes à 95°C; ii) alignement des amorces, 30 secondes à la température d'alignement (50°C à 62°C); iii) élongation à 72°C à partir des amorces pendant 1 à 2 minutes (1 minute par kb). Les réactions ont été faites avec un thermo cycleur Mastercycler (Eppendorf).

7.2.13 Séquençage d'ADN

Les constructions ont été vérifiées par séquençage utilisant le séquenceur « Applied Biosystems 310 automated DNA sequencer » avec le ABI PRISM Dye Terminator Cycle Sequencing Kit (Perkin Elmer) v.3.1.

7.2.14 Utilisation du vecteur pGhost9 pour la construction de mutants

Des cellules électrocompétentes de *L. lactis* sont transformées avec 10 µg de vecteur et incubées à 28°C dans du M17-saccharose-glucose sans antibiotique pendant 3-5 heures, étalées sur boîtes contenant du milieu M17-glu-Ery puis incubées à 28°C pendant 36 heures. Les transformants sont testés par PCR. Un clone correct est repiqué pour réaliser l'intégration du pGhost9. La culture est ensemencée au à 1/100ème et incubée pendant 2h30 à 28°C. La culture est placée 2h30 à 37°C. L'intégration est testée par PCR : deux bandes correspondantes à la taille du gène délété et non délété doivent être présentes. L'excision du plasmide se fait par une culture de 24 heures à 28°C sans antibiotique. Souvent plusieurs repiquages sont nécessaires pour exciser le plasmide. Après cette culture, les bactéries sont étalés sur boîtes et les clones Ery-sensibles (ayant perdu le plasmide) sont testés par PCR. Un mutant correct, c'est à dire, présentant par PCR un fragment correspondant

au gène délété est séquencé puis conservé.

7.3 Protocoles de microbiologie

7.3.1 Milieux et culture des bactéries

Les souches de lactocoque ont été cultivées dans du milieu M17 (Difco)-glucose 0,5% à 30°C sans agitation. Les souches de *E. coli* ont été cultivées sur milieu LB (Difco) sous agitation. Ces milieux sont utilisés sous forme liquide ou avec de l'agar pour cultiver les souches sur boîtes de Pétri. Pour la sélection des clones résistants, l'antibiotique est additionné au milieu à des concentrations entre 5 à 100µg/mL selon la souche: érythromycine (5 $\mu g\ ml^{-1}$ pour *L. lactis*, 150 $\mu g\ ml^{-1}$ pour *E. coli*), chloramphenicol (5 $\mu g\ ml^{-1}$ pour *L. lactis*; 20 $\mu g\ ml^{-1}$ pour *E. coli*), ampicilline (100 $\mu g\ ml^{-1}$ pour *E. coli*), kanamycine (20 $\mu g\ ml^{-1}$ pour *E. coli*). Pour le test de résistance à la pénicilline, cet antibiotique a été rajouté sur des disques placés sur la boîte ne contenant pas d'antibiotique après l'étalement des cellules. Les disques ont été irradiés pendant 10 minutes aux UV pour les stériliser.

Le milieu chimiquement défini (MCD) contient tous les acides aminés, le MCD minimum contient uniquement les 9 acides aminés essentiels pour *L. lactis* IL1403 (Cocaign-Bousquet *et al.*, 1995).

Pour les expériences de transport des peptides, un des acides aminées essentiels à été rajouter au MCD sous forme de peptide de telle manière à ce que la quantité de cet acide aminé soit la même que pour le MCD complet. Le peptide utilisé a été la Leu-Enkephalin (Bachem).

Pour révéler l'activité des autolysines, des micrococques (*Micrococcus luteus*) ATCC 4698 Sigma autoclavés ont été rajoutés dans le milieu M17-glu-agar à 0,2% (poids/volume) final.

7.3.2 Microscopie

Les observations au microscope électronique et la détection de PepF dans la cellule ont été faites à la plateforme de microscopie à l'INRA de Jouy en Josas (MIMA2). Pour la localisation de PepF dans la cellule, des anticorps polyclonaux ont été fabriqués par PARIS (Production d'Anticorps, Réactifs Immunologiques & Services, Compiegne, France). Une fois les anticorps testés par Western Blot, les cellules ont été fixées dans du paraformaldéhyde 4%, dehydratées avec de l'éthanol 30% (à 4°C) pendant une heure, 50%, 70%, (à -20°C), 90% and 100% (à – 35°C) et lavées pendant 3 heures à 35°C dans du Lowicryl K4M (Delta microscopies – Labège – France)/éthanol 100% (1v/2v, 1v/1v and 2v/1v, respectivement). Ensuite, deux lavages de Lowicryl K4M (16h et 2h) ont été réalisés. La polymérisation a été faite à 320 nm pendant 48 h à 35°C, et la température augmentée pendant 3 jours jusqu'à 20°C avant de suivre la procédure Leica AFS procédure (Leica – Microsystems Rueil-Malmaison – France). Des fines coupes de 90 nm ont été montées sur des grilles de nickel. Après la réaction d'inclusion dans du tampon PBS – 1% BSA – 0.1% eau froide et gélatine de peau de poisson, les coupes ont été incubées pendant 2h à température ambiante avec les anticorps PARIS contre PepF, dilution 1:100. Les coupes ont été lavées avec le même tampon et incubées pendant 30 minutes avec la protéine A (1:20) couplée à des particules d'or de 10 nm (Aurion – BioValley – France).

Pour l'étude de la division cellulaire, des cultures de *L. lactis* NZ9000-pSEC1 sauvage et du mutant

PepF induits et non induits à la nisine ont été fixés chimiquement dans une solution de 2% glutaraldéhyde et 0,1M de cacodylate de sodium, incluses dans la résine Epon et coupées à température ambiante.

7.4 Protocoles de biochimie et protéomique

7.4.1 Gel SDS-page

Des gels sodium dodécyl sulfate-polyacrylamide (4-12%) NuPage (Invitrogen) ont été utilisés pour l'électrophorèse en une dimension. Après migration des protéines, le gel est placé dans une solution de coloration (Simply Bleu safe stain, Invitrogen) pendant 1 h sous agitation après 1 lavage de 5 minutes à l'eau. Il est ensuite décoloré par plusieurs lavages à l'eau avant d'être scanné.

7.4.2 Extraction des protéines cytoplasmiques, d'enveloppe et sécrétées

Une fois la culture arrêtée par centrifugation, les cellules sont lavées avec du tampon phosphate 0,2M et cassées dans du tampon de cassage contenant un inhibiteur de protéase au disrupteur de cellules (Cell D, Basic Z 0.75KW) à 2400 bar. Les débris cellulaires sont séparés par centrifugation à 5000 rpm. La fraction cellulaire est séparée de la fraction d'enveloppe par ultracentrifugation à 50000 rpm, le culot est repris dans le tampon de cassage. Les protéines sécrétées sont concentrées 60 fois par précipitation avec l'acide trichloroacétique (TCA) 20% final et deux lavages à l'acétone pour éliminer complètement l'acide.

7.4.3 Dosage des protéines Bradford

La quantité de protéines est mesurée à 595 nm, longueur d'onde à laquelle le complexe formé entre protéines et bleu de Coomassie présente l'absorbance maximale. La quantité de protéines est déterminé par comparaison à une courbe étalon (solutions de BSA de concentration connue) réalisée en même temps que les échantillons sont mesurés.

7.4.4 Chromatographie

Les muropeptides obtenus lors de l'extraction du peptidoglycane ont été analysés par HPLC utilisant une colonne C18 Hypersyl PEP100. Ils ont été identifiés par comparaison à un standard pour lequel les peptides avaient été identifiés par spectrométrie de masse. Une fois la quantité relative de chaque pic déterminé (utilisant les surfaces des pics), les quantités de dimères (d), trimères (t) et tétramères (te) de muropeptides ont été utilisées pour calculer l'indice de réticulation (crosslinking index) de la paroi (Glauner, 1988) :

$$Crosslinkage = \frac{\frac{1}{2\sum d} + \frac{2}{3\sum t} + \frac{3}{4\sum te}}{\sum muropeptides}$$

Les acides tels que pyruvate, acétate, lactate et formate ont été quantifiés par HPLC sur

une colonne Aminex-HPX-87H (BioRad) avec une élution isocratique de H_2SO_4 5mM à un débit de 0,35 ml/minute à 35°C. Les protéines avaient été précipitées précédemment avec du H_2SO_4 2%. Les quantités d'acides présents ont été calculées par comparaison à une courbe de calibration standard de chaque acide analysé.

7.4.5 Extraction de peptidoglycane

Les peptidoglycanes de *L. lactis* IL1403 et son mutant *pepF* ont été préparés à partir d'une culture à DO_{600nm} =0,3 selon le protocole de (Atrih *et al.*, 1999). Les cellules sont portées à ébullition dans du SDS 4% durant 30 minutes. La paroi est lavée six fois avec de l'eau distillée. Pour éliminer les protéines, la paroi a été traitée avec la pronase (200 µg ml^{-1}) durant 16 h à 37 °C, puis, avec la trypsine (200 µg ml^{-1}) pendant 16 h à 37 °C. Ensuite, la paroi est traitée avec l'acide fluorohydrique pour éliminer les acides teichoiques. Après la digestion avec la muramidase, les muropeptides obténus sont réduits avec du borate avant l'analyse par HPLC.

7.4.6 Induction à la nisine

Une solution de nisine à 0,02mg/mL est préparée le jour de son utilisation et rajoutée aux cultures en phase exponentielle pour obtenir des concentrations finales de 5 à 10 ng/mL. Les cultures sont incubées de nouveau pendant 3 heures avant d'être analysées.

7.4.7 Extraction de peptides à partir des gels SDS-PAGE

Après la découpe des bandes d'intérêt à partir d'un gel SDS-PAGE, les morceaux de gel ont été lavés avec 100 µL de solution de lavage (50% bicarbonate d'ammonium 50mM et 50% d'acétonitrile). Une fois cette solution retirée et les morceaux de gel séchés à 37°C, une solution contenant la trypsine (40µL d'acide acétique 50mM + 20µg de trypsine PROMEGA) est rajoutée pour digérer les protéines durant une nuit à 37°C. Dans le cas où les peptides de masse comprise entre 0 et 6 kDa ont été analysés, cette étape de digestion n'a pas été fait.

Une fois la digestion finie et le surnageant contenant le lysat de la digestion retiré (ou les morceaux de gels séchés, dans le cas de peptides entre 0 et 6 kDa), 3 lavages successives des morceaux des gels sont faits avec une solution de bicarbonate d'ammonium 50mM pour extraire les peptides. Les surnageants sont poolés et séchés au speed-vac. Le culot, contenant les peptides concentrés est repris dans 20µL d'un tampon pour analyse au LC-MS-MS (acide trifluoroacétique 0,08%, acétonitrile 2%).

7.4.8 Analyse de peptides de taille inférieur à 3kDa

Les peptides ont été obtenus par ultrafiltration des extraits cytoplasmiques utilisant des filtres Microcon (Millipore) de <3 kDa. Ensuite, cette fraction a été séparée par HPLC et la fraction éluée entre 0 et 35% acétonitrile a été collectée pour enrichir l'échantillons en peptides hydrophobes

pour les cas où un peptide hydrophobe comme les morceaux de peptides signaux étaient recherchés (mutant PepF). Dans le cas où des peptides divers provenant des lipoprotéines étaient recherchés (mutant YvjB), une fraction plus hydrophile, éluée à >60% d'acétonitrile a été collectée. Ces échantillons ont été séchés, repris dans du tampon acide trifluoroacétique 0,08%, acétonitrile 2% et analysés par LC-MS-MS.

7.4.9 Production et purification de PepF-histidine-tagged produite chez *E. coli* BL21 (DE3) Gold

Une culture de *E. coli* BL21 (DE3) Gold contenant le plasmide permettant d'induire la production hétérologue de PepF-histidine-tagged sous le promoteur T7 a été réalisé dans du milieu LB-Km. La production a été induite à l'IPTG à une concentration finale de 1 mM au moment où la culture se trouvait à une DO_{650nm} de 0,5. Les bactéries ont été cultivées à 37 °C avant l'induction à l'IPTG et à 30 °C sous agitation de 200 rpm durant la production de PepF-histidine-tagged (4 h) pour éviter la formation de corps d'inclusion. Les cellules ont été récupérées par centrifugation et cassées au disrupteur de cellules (Cell D, Basic Z 0.75KW) à 1600 bars. La protéine PepF-histidine-tagged a été purifiée en utilisant des colonnes Ni^{2+}-*nitrilotriacetic acid* (Ni-NTA) (Qiagen).

7.4.10 Analyse au LC-MS-MS

Les peptides ont été analysés par un gradient de 2-80% d'acétonitrile dans l'eau contenant 0,1% d'acide formique pendant 50 minutes. Un débit de 300 nL/min a été utilisé pour l'élution des peptides de la colonne C-18 PepMap100 reversed-phase nanocolumn (75 μm ID x 15cm, 3 μm, 100 Å) (LC Packings, Amsterdam, Netherlands). L'ionisation et la fragmentation des peptides ont été réalisées dans un spectromètre de masse de type trappe d'ions. Les spectres ont été acquis entre 200 et 2000 m/z.

7.4.11 Dosage de l'activité de PepF par flourescence

L'activité de PepF a été mesurée sur un substrat fluorescent quenché : Mc-Pro-Leu-Gly-Pro-Lys-(DNP)OH. La fluorescence est émise quand le peptide est hydrolysé (Tisljar *et al.*, 1990). Cette activité a été mesurée pendant 100 secondes en utilisant un spectrofluoromètre (Kontron instruments, Uvikon 931).

7.4.12 Western Blot

Le transfert de protéines d'un gel SDS-PAGE sur une membrane PVDF a été fait dans un système BIORAD à 100 V pendant 2 heures. La membrane a été colorée avec du rouge Ponceau pour vérifier le transfert. Ensuite la membrane a été saturée avec du tampon PLT (PBS 10mM, Tween 20 0,1%, BSA 1%) pendant 2 heures avant de l'incubation avec le premier l'anticorps. Après une nuit d'incubation à 4°C et 3 lavages au PLT, le deuxième anticorps a été incubé pendant 1 heure. Les dilutions des anticorps ont varié entre 1/1000 et 1/3000. Ces concentrations ont été déterminées

empiriquement. La révélation a été faite par le kit ECL (Plus Western Blotting Detection System) Amersham Biosciences.

7.5 Logiciels et obtention des données

7.5.1 Analyse de séquences d'ADN et protéiques

Les séquences d'ADN et protéiques nécessaires pour le design d'amorces et la comparaison des séquences ont été obtenues en interrogant les bases des données KEGG et NCBI. L'alignement des séquences a été réalisé avec Clustal W 1.83 (Thompson *et al.*, 1994). Le design et l'analyse des amorces ont été effectués avec le logiciel Oligo Explorer 1.1.0. La prédiction de peptides signal a été faite avec SignalP (Bendtsen *et al.*, 2004). L'alignement multiple pour la recherche de motifs dans la séquence a été réalisé avec AlignACE version 3.0 (Roth *et al.*, 1998). Clone Manager © a été utilisé pour planifier la construction des plasmides. La traduction des séquences d'ADN, la prédiction de masse des séquences protéiques et le calcul des indices GRAVY ont été faits avec les outils Expasy (http://www.expasy.org/tools/).

7.5.2 Analyses des résultats MSMS

Les spectres MS/MS ont été analysés avec Bioworks 3.2 et les bases de *L. lactis* MG1363 ou IL1403 ont été interrogées en utilisant SEQUEST.

7.5.3 Mise en forme des données

Nous avons utilisé cinq types de données : un graphe des voies métaboliques (1), des profils phylogénétiques (2), des données sur la distance sur le chromosome (3), des données transcriptomiques (4) pour le KCCA. Les données issues de gels 2D (5) ont pu être incluses grâce à une stratégie de complétion de noyaux que nous avons développée (article I). Pour les noms des fichiers correspondant à chaque type de données voir le Tableau 7-4**.** Ces données sont fournies dans les données supplémentaires.

1) Le graphe des voies métaboliques

Pour obtenir un graphe des toutes les voies métaboliques, les fichiers xml qui décrivent chaque voie métabolique de *L. lactis* ont été téléchargés utilisant la page ftp.genome.ad.jp correspondant à la version KGML_v0.6. 87 fichiers ont été récupérés. Ensuite les relations deux à deux entre gènes ont été extraites de ces fichiers avec le script list2adj écrit en langage ruby et fournie par JP Vert. A partir de ce fichier, une matrice laplacienne est créée. Il s'agit d'une matrice contenant des « 1 » si un lien existe entre deux protéines et des « 0 » s'il n'y a pas de lien. Sur la diagonale se trouve la valeur négative du degré de chaque protéine. Dans la première colonne les identifiants des protéines sont indiqués utilisant l'identifiant de la base KEGG (L166370 par exemple).

2) Les profils phylogénétiques

Pour construire les profils phylogénétiques, il est nécessaire de récupérer les séquences de tous les gènes du lactocoque et de réaliser des BLAST contre d'autres bactéries. La base de données utilisée est celle des séquences protéiques codées par chaque gène. Dans cette base de données les noms de protéines sont donnés par NP_ suivi d'un numéro caractéristique. C'est cet identifiant qui a été choisi pour les protéines. Le logiciel public ARCT (de Magalhães *et al.*, 2005) a été utilisé pour cette tâche. Il est programmé en langage Perl et s'appuie sur des libraries BioPerl pour l'interrogation de la base NCBI et pour réaliser les BLAST automatiques. Des BLAST ont été faits contre 276 bactéries entièrement séquencées avec comme seuil de détection une E-value de à 10^{-5}. La sortie est un tableau de 2180 lignes (protéines) et 278 colonnes. Dans la première colonne se trouve le nom des gènes (NP_267883 par exemple), dans la deuxième colonne le nombre d'a.a. de chaque protéine; les 276 colonnes restantes correspondent aux 278 bactéries.

3) La distance sur le chromosome

Les positions des gènes ont été récupérées dans http://genome.jouy.inra.fr/cgi-bin/micado/index.cgi. La distance entre gènes a été calculée entre la fin d'un gène et le début du suivant directement dans dans le programme « kernel » une fois l'information sur les positions chargée.

4) Les données transcriptomiques

Les données transcriptomiques ont été récupérées à trois endroits différents : i) La base des données transcriptomiques http://genome.jouy.inra.fr/efp/base/www provenant du projet européen *Express Fingerprint*, ii) d'E. Guedon et iii) Sur le GEO du NCBI http://www.ncbi.nlm.nih.gov/geo/. Les différents fichiers ont été fusionnés en un seul fichier utilisant les gènes présents sur tous les fichiers et les noms des gènes (*pepF*) ont été changé pour obtenir l'identifiant NCBI (NP_267883). Le résultat de cette opération est un fichier avec les noms de gènes dans la première colonne et les données correspondant à 115 expériences (avec deux valeurs, rouge et vert, par gène et par expérience), soit 230 colonnes de données. Le log ratio normalisé entre les deux conditions a été calculé pour obtenir une seule colonne par expérience : $\log_2(vert/rouge) - \log_2(\sum vert / \sum rouge)$. Pour les expériences comportant des répétitions par gènes, la moyenne a été prise.

5) Les données des gels 2D

Ils proviennent de 13 expériences différentes (24 gels au total). Elles ont été toutes réalisées par C. Gitton (Unité BioBac). Pour les cas des répétitions la moyenne du volume détecté pour chaque protéine a été utilisé. Il s'agit des données incomplètes parce que uniquement les protéines solubles et avec un point isoélectrique compris entre pH 4 et 7 sont détectées. Une autre difficulté s'est présenté pour ces données. Une partie des expériences avait été faites sur la souche *L. lactis* NCDO763. Un test statistique de *maximum descrepancy* (Borgwardt *et al.*, 2006) a été réalisé pour confirmer que les données provenant des deux souches pouvaient être utilisées ensemble. Les données des gels 2D sont des volumes des spots qui représentent chaque protéine. Cette quantité est normalisée par le volume total. Ce volume normalisé V a été encore transformé selon les recommandations de Chich et collaborateurs : $tV = \sqrt[3]{V}$ (Chich *et al.*, 2007).

6) Le fichier de correspondances

Il a été nécessaire de créer un fichier de correspondances utilisant l'information des bases des données comme http://www.infobiogen.fr/ (actuellement fermé) et http://genome.jouy.inra.fr/cgi-

bin/micado/index.cgi, puis complété manuellement. Ce fichier contient les correspondances entre les numéros de locus (annotation originale du génome), les numéros d'accession SWISSPROT, les numéros NCBI de la base des protéines (NP_267883), les identifiants KEGG (L166370), ainsi que les noms de gènes (*pepF*).

Une fois les fichiers construits, les noms des gènes ont été harmonisés à l'aide du code lit_fic.m écrit sous matlab.

Ceci fait, les gènes communs à tous les types de données ont été recherchés pour constituer le lot d'apprentissage (programme : donnees.m). A l'aide du même programme 333 gènes ont été identifiés dans tous les sources de données. Ensuite, les gènes communs aux sources des données 2-5 ont été recherchés pour constituer l'ensemble des gènes candidats soit 1756 gènes. Les différentes étapes sont schématisées sur la Figure 7-5 et les jeux de données résultants sont listés dans le **Tableau 7-4**.

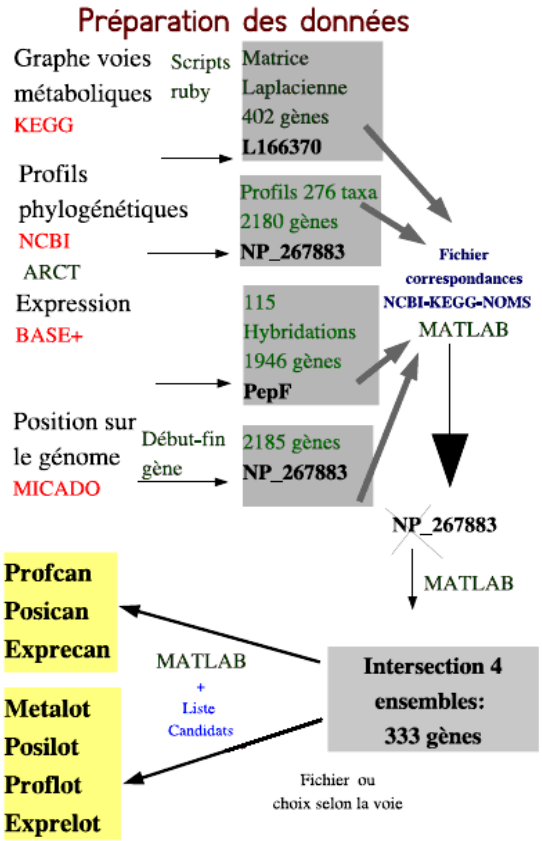


Figure 7-5 : Etapes de récupération et mise en forme des données.

Tableau 7-4 : *Données*

Type de données	Provenance	caractéristiques	Fichier
Réseaux des voies métaboliques	KEGG : KGML_v0.6	333×333	metatout.txt
Profils phylogénétiques	BLAST avec ARCT 0.9 contre 276 bactéries avec un seuil de E-value=10^{-5}	333×276 1756×276	proftout.txt profcan.txt
Distance entre gènes en p.b.	Calculée à partir des positions des gènes	333×2 (début et fin) 1756×2	positout.txt posican.txt
Données transcriptomiques	http://genome.jouy.inra.fr/efp/base/www Gene Expression Omnibus (GEO), NCBI (Guedon *et al.*, 2001)	333×115 (115 hybridisations) 1756×115	expretout.txt exprecan.txt
Gels 2D	24 gels (Gitton *et al.*, 2005) (Guillot *et al.*, 2003)	216×13	2D.txt
Noms de gènes	Fichier des correspondances	2284×5	ORFs.xls

7.5.4 Programmes réalisés

Plusieurs codes ont été écris sous Matlab v. 7.2.0.232 (The Mathworks ©) pour réaliser les prédictions des liens par KCCA, ces codes utilisent 4 sources ou 5 sources de données pour réaliser les prédictions, permettent de faire la détermination des paramètres par validation croisée, la mise en forme des noms de gènes et l'analyse de sensibilité. Les noms de fichiers correspondants aux codes sont listés dans le **Tableau 7-5**.

Le **Tableau 7-5** contient la liste des codes développés sous Matlab v. 7.2.0.232 (The Mathworks ©) ainsi que d'autres codes utilisés pour la construction et l'analyse des données.

Tableau 7-5 : *Codes utilisés*

Code	Référence	Langage	Utilisation
kernel	ce travail	matlab	KCCA avec 4 sources des données
kernelprot	ce travail	matlab	KCCA avec 5 sources des données
crossv	ce travail	matlab	validation croisée pour déterminer les paramètres
donnes	ce travail	matlab	construction des jeux de données
lit fic	ce travail	matlab	harmonisation des noms de gènes
anal_sens	ce travail	matlab	analyse de sensibilité du KCCA
ARCT	(de Magalhães *et al.*, 2005)	perl	BLAST automatiques
list2adj	JP Vert	ruby	retrouve les relations 2 à 2 du graphe KEGG
make net	JP Vert	ruby	Construction d'un graphe des toutes les voies métaboliques de KEGG pour *L. lactis*

7.6 Analyses statistiques

Dans ce travail nous avons utilisé une méthode d'inférence supervisée reposant sur l'analyse de corrélation canonique. Nous avons étendue cette méthode à des données incomplètes. Nous avons aussi proposé une approche originale d'identification de données essentielles pour répondre à l'inférence des rôles de protéines cibles. Cette méthode se base sur les plans d'expériences fractionnaires. La manière dont les méthodes ont été adaptées et appliquées dans le cadre de ce travail est détaillée dans les articles. Uniquement des détails concernant l'exploration de modèles pour la recherche de sous-ensembles de facteurs qui conduisent aux mêmes prédictions que la référence, réalisée dans le cadre de l'analyse de sensibilité sera détaillée ici.

7.6.1 Enumération de modèles parcimonieux en facteurs

Une fois les simulations réalisées et un modèle linéaire ajusté, nous nous retrouvons avec un certain nombre de facteurs classés par ordre d'importance. Ces facteurs représentent des facteurs principaux $E_{\mathrm{Pr}} = \{x_i, i \in I\}$, c'est à dire une seule variable (par exemple une expérience transcriptomique) ou des interactions d'ordre deux entre variables $E_{Int} = \{(x_u, x_v), (u,v) \in U)\}$. Soit $P_{u,v} = \{(x,y) \sim (x_u, x_v)\}$ l'ensemble des alias de l'interaction (x_u, x_v). Du fait que les effets d'ordre deux sont confondus, nous nous retrouvons en fait avec un grand nombre de facteurs possiblement impliqués dans l'effet donnée par le modèle. Ils est nécessaire d'énumérer les modèles possibles tenant compte de tous les facteurs impliqués et donc de la structure d'alias de facteurs d'ordre deux. On s'intéresse alors aux modèles comportant les facteurs principaux E_{Pr} et un représentant de chaque paquet d'alias $P_{u,v}(u,v) \in U$. Soit *M* l'ensemble de tous les modèles correspondant à E_{Pr} et $E_{Int.}$ Soit *F(m)* le nombre de facteurs différents intervenants dans le modèle $m \in M$. On a $|E_{\mathrm{Pr}}| \le F(m) \le |E_{\mathrm{Pr}}| + 2 \times |E_{Int}|$. On dira qu'un modèle *m* est parcimonieux lorsque $F(m) = \min_{m \in M} F(m)$. Il n'est pas possible de calculer toutes les possibilités. Nous avons proposé l'algorithme suivant pour rechercher un modèle parcimonieux. Il s'agit d'un algorithme sous-optimal:

On affecte une étiquette à chaque facteur de E_{Pr}, de -1 à $|E_{\mathrm{Pr}}|$. On affecte aussi les étiquettes 1 à $|E_{Int}|$ à chaque paquets d'interactions. On construit alors une matrice *S* de taille $T \times T$, où *T* est le nombre de facteurs apparaissant dans $E_{\mathrm{Pr}} \cup_{(u,v) \in U} P_{u,v}$:

Un exemple de matrice *S* est montré en Figure 7-6 et un résumé de l'algorithme est présenté plus bas.

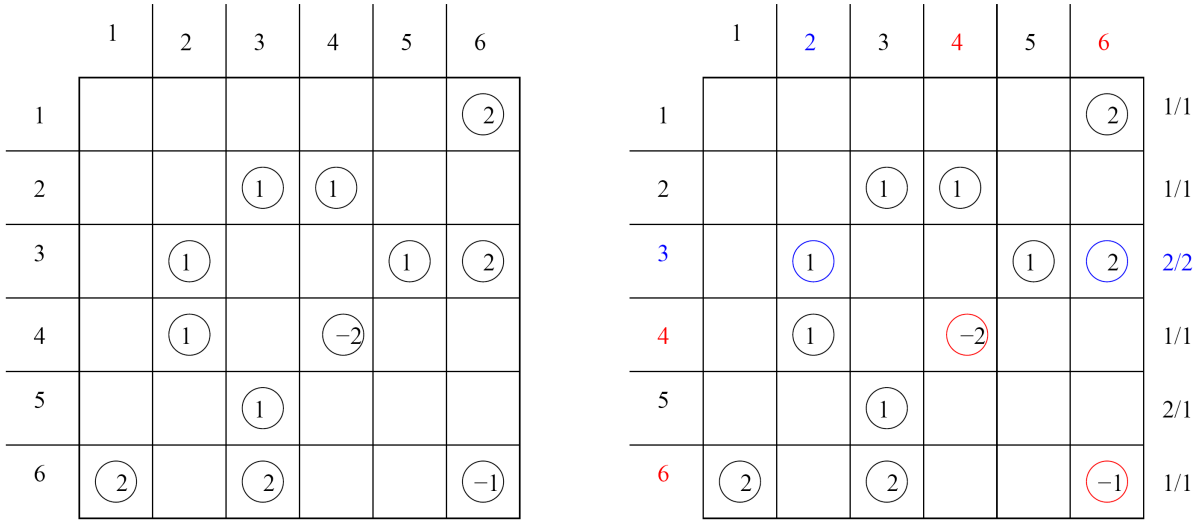


Figure 7-6 : Construction de la matrice *S* et sélection des facteurs. Les facteurs en rouge sont d'abord sélectionnés puis ceux en bleu, car ils correspondent au coût optimal : Le coût optimal est 2/2 dans cette exemple.

Algorithme

1. sélectionner tous les facteurs correspondant aux effets principaux
2. regarder s'il y a des interactions n'impliquant que les facteurs principaux
3. faire le point des paquets non encore couverts
4. S'il y des paquets non couverts,
   (a) calculer pour chaque ligne de la matrice le nombre maximum ***nbfact*** d'interactions différentes qu'on peut couvrir en considérant $x_{i(l)}$ et d'autres facteurs, calculer aussi le nombre *nbinteract* minimum de facteurs supplémentaires nécessaires pour cela.
   (b) selectionner la ligne de rapport $\frac{nbfact}{nbinter}$ minimal. S'il y a des lignes pour lesquelles ce rapport est identique, on choisit celle pour laquelle *nbinter* est le plus grand.
   (c) Ajouter le facteur $x_{i(l)}$ et les facteurs correspondant aux etiquettes couvertes et non déjà inclus dans la liste. Il peut y avoir plusieurs possibilités équivalentes en terme d'interactions couvertes et de nombre de facteurs. Dans ce cas, soit on construit plusieurs listes en parallèle, soit on choisit d'une manière ou d'une autre
   (d) modifier la matrice S : enlever les etiquettes couvertes.
   (e) Retourner en 3.
5. Terminé : on a une liste de facteurs

8 Discussion

Le travail réalisé au cours de cette thèse est un travail multidisciplinaire qui a apporté des résultats et des réflexions sur, d'une part, l'inférence du rôle des protéines par des méthodes à noyaux et d'autre part, l'implication d'enzymes protéolytiques dans l'export des protéines des lactocoques. Notre approche est une des rares à combiner une approche statistique et des tests expérimentaux sur des protéines dont les rôles ne sont pas connus. Les autres études rapportées valident généralement les modèles par des jeux de données déjà connus.
Nous allons discuter chacun des aspects abordés : les difficultés et précautions à prendre lors de la construction et l'obtention des données, l'analyse de corrélation bâtie sur des noyaux (KCCA) utilisés pour prédire des relations entre protéines et sa position parmi d'autres méthodes d'inférence, la qualité des prédictions et le type de liens obtenus, les hypothèse biologiques issues des liens prédits et la manière de les poser ainsi que les résultats des validations expérimentales à la lumière des connaissances sur les protéines cibles chez le lactocoque et d'autres bactéries.

Notre méthode de prédiction

Si nous reprenons les étapes de cette étude, nous retrouvons à l'origine les données qui ont été utilisées pour l'inférence statistique. Les données concernant la distance entre gènes sur le chromosome et les profils phylogénétiques ont été obtenues en interrogeant des bases de données avec des outils automatiques disponibles développés sous différents langages de programmation. Les données transcriptomiques proviennent également de différentes sources. Ce fait n'a pas facilité la mise en forme des données, puisque une harmonisation des noms des gènes a été nécessaire. Il a été également nécessaire d'éliminer une partie des gènes de l'ensemble de sources de données quand ils étaient absents d'une des expériences transcriptomiques. Il s'agit pour la plupart de petits gènes ou de gènes de phages. En conséquence, uniquement 1756 gènes (sur 2321 catalogués dans le NCBI) ont été gardés pour l'analyse. Ce travail est souvent négligé ou simplement pas mentionné dans la plupart des études mais il nécessite un temps non négligeable. Cette étape du travail aurait était plus aisée si les noms des gènes et des formats standardisés avaient été utilisés dans les différentes bases de données. Très liée à la récupération de données est la transformation de ces données, qui, en général, cherche à ramener à la même échelle des données provenant de différentes répétitions, conditions et même laboratoires. Nous avons utilisé des transformations couramment rapportées comme le logarithme pour les données transcriptomiques ou la troisième racine pour les données des gels 2D, mais plusieurs autres possibilités existent.

Il est également très important de se poser la question de la pertinence et de la qualité des données utilisées, compte tenu du fait qu'il s'agit de données obtenues à haut-débit ou de manière automatique. Il est très difficile de répondre à cette question, puisque il est rarement possible de réaliser une analyse qui permette d'évaluer toutes les sources des données et variables utilisées. Pour cette raison, il est nécessaire de prendre des précautions dans le but de trier d'une certaine manière les données de bonne et de mauvaise qualité. Nous avons surmonté cette difficulté en effectuant une analyse statistique supervisée. Cette analyse permet de réaliser un apprentissage par rapport à une source de données dans laquelle les relations entre protéines sont connues.

Le fait de connaître les relations entre certaines protéines grâce au réseau métabolique KEGG nous a permis également de valider la qualité des prédictions sur la base de ce réseau. Nous savions que

nos prédictions étaient de bonne qualité (environ 17% d'erreur) pour la reconstruction du graphe métabolique connu.

Notre but était de prédire des liens de tous types et non pas uniquement des liens métaboliques (c'est à dire les liens qui lient deux protéines quand elles réalisent deux réactions successives dans une voie métabolique) mais aussi des liens représentant la participation à une même fonction cellulaire, une même cascade de signalement, une régulation commune, etc. La question qui se posait alors, était de savoir si les liens inférés étaient applicables aussi aux relations de tous types. Nous avons abordé cette question en essayant de reconstruire des liens connus entre protéines qui ne sont pas liées métaboliquement comme nous l'avons défini ci-dessus et n'avaient pas été utilisés dans l'apprentissage. Nous avons comparé les liens inférés avec des interactions physiques entre protéines déterminées par la technique du double-hybride, ainsi qu'avec des régulons connus chez le lactocoque. Même si les liens directs entre deux protéines interagissant physiquement n'ont pas été retrouvés, un réseau très voisin du réseau d'interactions physiques a été retrouvé. Cependant, les relations de protéines faisant partie du même régulon (CodY et CcpA) ne sont pas retrouvées dans les relations prédites. Trois hypothèses peuvent être tirées de ces résultats : (i) les protéines participant à une même fonction ne sont pas forcement régulées de la même manière ; (ii) la régulation commune d'un groupe de gènes est observable dans des conditions très précises et très difficiles à retrouver lors de l'intégration de plusieurs conditions transcriptomiques au même temps (cette hypothèse semble pertinente au vu des profils d'expressions des gènes du régulon CodY dans les différentes expériences de transcriptomique utilisées (voir figure); (iii) une protéine peut être régulée de plusieurs manières ou faire partie de plusieurs régulons.

Parmi les méthodes de prédiction du rôle des protéines disponibles au moment où nous avons réalisé les prédictions pour PepF et YvjB, nous avons choisi l'analyse de corrélation canonique bâtie sur des noyaux (KCCA) proposée par Yamanishi et collaborateurs (2003). C'est une méthode supervisée qui présente l'avantage de permettre l'intégration de différentes sources de données qui sont toutes représentées sous forme de noyaux. Ceci facilite l'utilisation de sources de données hétérogènes grâce à la flexibilité de représentation par des noyaux.

La méthode KCCA présente l'inconvénient de ne pas accepter des sources contenant des données manquantes. Néanmoins, des sources de données très informatives existent et celles-là ne peuvent pas être obtenues pour tous les gènes ou toutes les protéines d'un organisme. Tel est le cas des gels de protéines à deux dimensions. Même si l'abondance relative de protéines dans une condition donnée n'est pas disponible pour toutes les protéines, il s'agit d'une source d'information complémentaire aux données transcriptomiques. Nous avons proposé une stratégie pour compléter les données provenant des gels 2D pour pouvoir les inclure dans nos sources de données. Pour inclure ces données il faut surmonter la difficulté des données manquantes. Deux possibilités existent : i) compléter les données ou ii) compléter le noyau. Nous avons décidé de construire un noyau avec les données existantes et de compléter le noyau lui même avec les connaissances disponibles sur les paires des protéines absentes. Ce prend compte la différence entre l'absence de protéines pour causes liées à la technique et les protéines vraiment absentes. Nous avons constaté que l'information supplémentaire apportée par cette source de données est très faible, ce qui est dû sûrement à la quantité de protéines absentes (environ 85%). Ce qui est intéressant, c'est que nous avons montré que même les sources de données incomplètes peuvent être utilisées dans la KCCA.

Les résultats des prédictions de la KCCA se sont avérés très utiles pour émettre des hypothèses biologiques sur le rôle des protéines. Tout d'abord, le fait d'avoir des distances ordonnées permet de se concentrer en priorité sur les protéines prédites comme étant les plus proches. La protéine la plus proche nous a guidé vers le rôle principal de PepF et de YvjB. Les autres prédictions étaient très variées pour PepF, comportant des protéines appartenant à plusieurs fonctions différentes. Au contraire, pour YvjB une grande majorité de protéines participait dans la sécrétion de protéines. Cela indique que la KCCA est une analyse qui permet de repérer les fonctions principales tout en tenant compte des fonctions secondaires ou pléïotropes.

PepF

Les prédictions obtenues nous ont permis de déceler la participation de PepF et YvjB à différentes fonctions cellulaires et nous avons réussi, dans la plupart des prédictions testées, à confirmer leur implication dans ces fonctions ce qui valide notre approche prédictive. Néanmoins, les prédictions sur l'implication d'une protéine donnée dans une fonction ne sont pas suffisantes pour déterminer les mécanismes précis mis en jeu au niveau moléculaire, ni pour proposer un modèle pour intégrer les différentes fonctions cellulaires affectées.

Pour PepF nous avons déduit des résultats expérimentaux qu'elle participe dans la sécrétion de protéines plus vraisemblablement comme signal peptide peptidase (SPPase), fonction qui correspond bien à son rang d'action sur les oligopeptides (entre 7 et 17 acides aminés, (Nardi *et al.*, 1997)). En son absence et en conditions de surproduction de protéines exportées, la translocation de protéines est affectée. Les résultats obtenus nous font penser que le peptide signal n'est pas efficacement dégradé. Il encombre possiblement la machinerie de sécrétion, phénomène qui avait déjà été observé chez *E. coli* (Chen *et al.*, 1987; Wickner *et al.*, 1987) *in vitro*. De la même manière que chez *B. subtilis* (Bolhius *et al.*, 1999) et chez *E. coli* (Wickner *et al.*, 1987) il est nécessaire de surproduire des protéines sécrétées pour observer un blocage de la sécrétion dans les mutants de signal peptide peptidases SppA et OpdA respectivement. Cela est dû principalement au fait que d'autres peptidases peuvent prendre le relais pour hydrolyser des peptides signaux libérés par les SPases. OpdA chez *E. coli*, une protéine similaire à PepF a aussi été décrite comme participant à la sécrétion, puisque son absence affecte la sécrétion de plusieurs protéines (Conlin *et al.*, 1992; Emr, 1982).

Néanmoins, nous ne savons pas comment se fait l'encombrement de la machinerie de sécrétion. Une étude réalisé par Ahn et collaborateurs a montré que l'interaction du peptide signal des protéines exportées avec SecA provoque une ouverture de sa structure (Ahn & Kim, 1996). Il est alors possible d'imaginer qu'en cas de surproduction de protéines sécrétées, SecA est amené à interagir plus que dans des conditions normales avec les protéines possédant un peptide signal. Dans ces conditions, les morceaux non dégradés du peptide signal (en absence de PepF) pourraient encombrer SecA et empêcher le transfert de protéines dans le canal de translocation SecYEG.

D'autres explications alternatives existent pour expliquer les observations faites. Il est possible, qu'en absence de PepF les deux types de protéines exportées (la protéine sécrétée NucB et la liporptoéine OptA) soient dégradées, possiblement par la dérégulation d'une protéase causée par l'absence d'un peptide produit par PepF. Comme ce peptide ne serait pas produit en absence de PepF, une cascade de régulation pourrait provoquer cette dérégulation et la dégradation des protéines sécrétées. Une dérégulation de la sporulation, engendrée par l'absence de PepF avait été

observée chez *B. subtilis*. Néanmoins, cette explication est peu plausible puisque une protéase capable de dégrader deux protéines si différentes pourrait théoriquement dégrader également d'autres protéines, ce qui provoquerait le dysfonctionnement de plusieurs mécanismes dans la bactérie et le mutant *pepF* ne présente aucune diminution de la viabilité ni de la croissance.

Nous avons également observé une implication de PepF dans le métabolisme du pyruvate, l'observation principale étant une diminution de la quantité de lactate. La participation de PepF dans ce métabolisme reste inexpliqué, mais il est possible que la modification du cycle du pyruvate, notamment, la diminution en production de lactate, provoque indirectement un dysfonctionnement de la machinerie de sécrétion par un possible manque d'ATP. Cependant, le lactocoque est capable de réaliser une fermentation mixte. Il est capable de se procurer de l'ATP par d'autres voies, notamment en produisant de l'acétate, qui est produit en plus grande quantité dans la souche mutant *pepF* NZ9000 déjà en conditions normales (sans surproduction d'une protéine sécrétée). De plus, la croissance n'est pas affectée, comme cela a déjà été mentionné. Pour ces raisons nous favorisons comme rôle principal de PepF son implication dans la sécrétion via la dégradation de peptides signaux avec un effet pléïotrope sur le métabolisme du pyruvate.

Nous postulons que la réticulation plus forte de la paroi observée chez le mutant *pepF*, est le résultat d'une réponse au stress provoqué par l'absence de PepF. Le renforcement de la paroi cellulaire en réponse au stress osmotique et nutritionnel a déjà été observé chez *L. casei* (Piuri *et al.*, 2005) et *E. coli* (Gyaneshwar *et al.*, 2005) respectivement.

Nous n'avons pas observé de différences morphologiques entre le mutant *pepF* et la souche sauvage lors de la division cellulaire comme cela aurait pu être attendu par des liens prédits avec des protéines de la division cellulaire. Il est possible que les différences ne soient pas visibles dans nos conditions d'expérience ou que simplement le lien prédit avec la division soit incorrect. Néanmoins, une implication de PepF dans la sporulation chez *B. subtilis* a été reportée par Kanamaru et collaborateurs (Kanamaru *et al.*, 2002), puisque sa surproduction inhibe la sporulation. Compte tenu que plusieurs protéines de la sporulation participent aussi à la division cellulaire, un lien avec cette fonction semble correct, mais n'a pas pu être prouvé par nos validations.

YvjB

Nous avons observé qu'YvjB est nécessaire à la localisation correcte et la maturation d'au moins une lipoprotéine (OptA). Selon nos observations YvjB est nécessaire pour que les lipoprotéines sur lesquelles elle agit, atteignent l'enveloppe cellulaire et ensuite, pour qu'elles soient maturées par la SPase II. Elle ne participe pas uniquement au clivage mais aussi à la reconnaissance et à la mise en place de la lipoprotéine, puisqu'en son absence elle n'est même pas présente dans l'enveloppe. Cette reconnaissance dépend de la nature du peptide signal de la lipoprotéine. Seules les lipoprotéines reconnues par YvjB sont affectées en son absence (dans le mutant). De plus, YvjB est nécessaire pour le clivage lui même de ces lipoprotéines. Nous n'avons pas pu mettre en évidence précisément son action sur la lipoprotéine. Le fait que YvjB ne prenne pas en charge toutes les lipoprotéines, nous amène à nous interroger sur l'existence d'une autre enzyme protéolytique ayant le même rôle que YvjB qui prendrait en charge les autres lipoprotéines. Cette protéine pourrait être YueF qui possède un domaine protéolytique et un segment transmembranaire prédit. Néanmoins, nous n'avons pas trouvé de caractéristiques communes aux lipoprotéines, qui ne sont pas reconnues par YvjB. Par contre, nous avons trouvé des caractéristiques communes dans la séquence des

peptides signaux des lipoprotéines prises en charge par YvjB, notamment une lysine et une leucine en positions 6 et 9 du peptide signal, ainsi qu'un acide aminé hydrophobe en postion -2 de la lipobox, ce qui favoriserait le fait que les autres lipoprotéines ne soient pas prises en charge par aucune protéine. Les protéines appartenant à la famille S2P, les homologues de YvjB, participent à la protéolyse intermembranaire et pour une d'entre elles, Eep de *Enterococcus faecalis*, son action sur le peptide signal d'une lipoprotéine a été montrée. Cependant, la conversion du peptide signal est réalisée après la libération par la signal peptidase II.

Egalement, comme pour le cas de PepF, il est possible d'imaginer que la disparition d'OptA de l'enveloppe est due à une dégradation et que la présence de YvjB rend plus stable OptA face aux possibles protéases. Néanmoins, deux arguments vont à l'encontre de cette hypothèse. D'abord, le fait qu'un dérivé de la même protéine, se différenciant uniquement au niveau du peptide signal de YvdF, soit plus stable. Ensuite le fait qu'aucune trace de dégradation de OptA n'ai été trouvé ni par Western blot, ni par LC-MSMS. Il est aussi possible qu'OptA soit libérée dans le milieu extérieur à cause d'un défaut de maturation qui empêcherait son attachement à la membrane. Néanmoins, nous n'avons pas détecté la protéine dans le surnageant de culture lors de la phase exponantielle de croissance.

Sécrétion et maturation des protéines chez Lactococcus lactis

Les observations que nous avons faites sur PepF et YvjB complètent les connaissances sur l'export de protéines chez *L. lactis*. Le plus probable est que PepF hydrolyse la partie hydrophile (N-terminale) des peptides signaux d'au moins deux types de protéines exportés, les protéines sécrétées possédant un peptide signal comme Usp45 et les lipoprotéines possédant un peptide signal comme OptA. Il s'agit de la partie qui, de part sa nature hydrophile, regagne le cytoplasme (Tjalsma *et al.*, 2000) et serait donc accessible à PepF (localisée dans le cytoplasme). Compte tenu du fait que tous les peptides signaux ont des caractéristiques chimiques similaires (von Heijne, 1989) et que la spécificité de clivage de PepF est très large (Monnet *et al.*, 1994) il est tout à fait possible que cette enzyme soit chargée d'hydrolyser tous les peptides signaux des protéines exportées.

Concernant YvjB nous avons montré que sa présence est nécessaire pour la localisation et la maturation correcte de certaines lipoprotéines et que la reconnaissance de ces lipoprotéines dépend de la nature du peptide signal. Compte tenu des résultats des travaux réalisés sur Eep (l'homologue de YvjB) chez *Enterococcus faecalis* (An & Clewell, 2002; An *et al.*, 1999; Antiporta & Dunny, 2002; Buttaro *et al.*, 2000) il est très probable que YvjB agisse sur la partie hydrophobe (C-terminale) du peptide signal. Malgré un lien apparent entre la participation de PepF et YvjB sur les peptides signaux de protéines exportées, nous n'avons pas prédit de liens entre ces deux protéines ni entre PepF et la signal peptidase II ou inversement entre YvjB et la signal peptidase I. Néanmoins, nous avons montré que l'absence de PepF affecte aussi l'export de lipoprotéines, puisque dans un mutant PepF, la quantité de OptA diminue, lors de la surproduction d'une protéine sécrétée.

Ce que nous pouvons conclure est que l'action des deux protéines est indépendante et que le point de rencontre est la voie de translocation. D'ailleurs, les prédictions avaient mis en évidence des liens de PepF et YvjB avec la voie de translocation Sec (liens prédits entre PepF et SecA, et YvjB et SecY). Les connaissances sur les diverses voies de translocation chez le lactocoque sont insuffisantes pour savoir si les différents types de protéines exportées que nous avons analysées passent par les mêmes voies d'exportation. Néanmoins, à la lumière des connaissances actuelles

nous pouvons imaginer que les deux types de protéines passent par la voie de translocation Sec, la voie encombrée par l'absence de PepF. Nos résultats montrent que dans des conditions normales, en absence de PepF, les lipoprotéines prises en charge par YvjB ne sont pas affectées. Cependant, lors de la surproduction de protéines sécrétées la quantité de lipoprotéines prises en charge par YvjB diminue dans l'enveloppe, montrant que les deux types de protéines passent par la même voie de translocation. Les résultats des nos observations sont résumés en Figure 8-1.

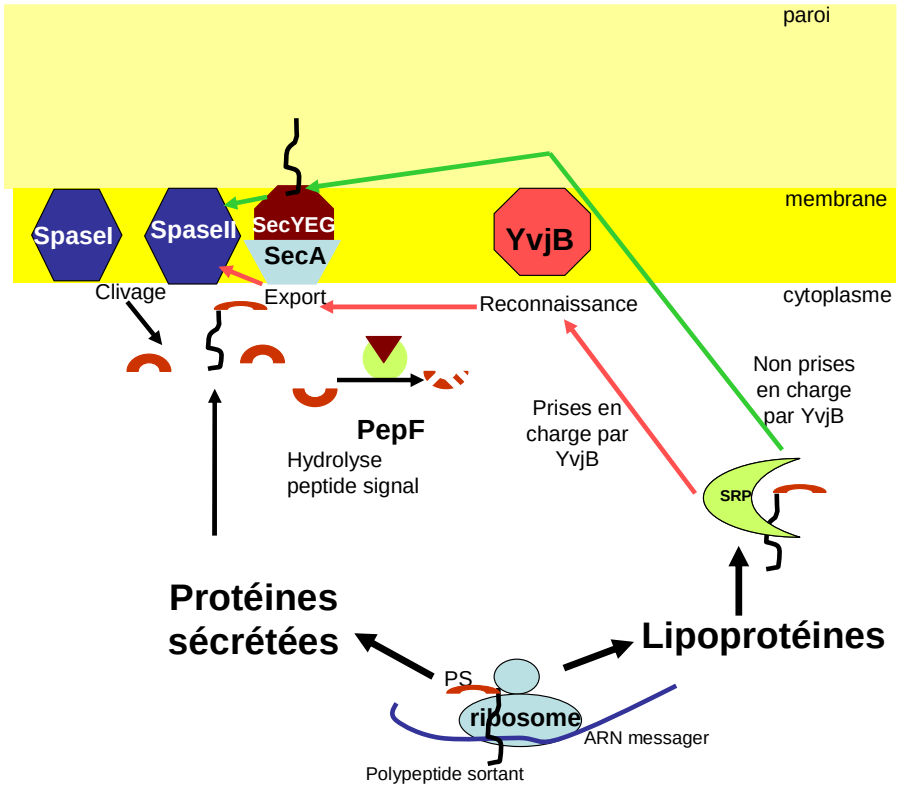


Figure 8-1 : Modèle global de l'export de protéines qui résume le rôle de PepF et YvjB. PepF participe dans l'export de protéines comme signal peptide peptidase. YvjB participe à la reconnaissance et maturation de lipoprotéines de peptides signaux hydrophobes, de grande taille qui contiennent le motif KXXL en position 6 et 9 du peptide signal. Les étapes sont 1) La reconnaissance ; cette étape existe uniquement pour les protéines prise en charge par YvjB, 2) L'export, 3) Le clivage du peptide signal, 4) L'hydrolyse du peptide signal.

L'approche que nous avons utilisée a été très efficace pour la détermination des liens entre PepF et YvjB et d'autres protéines, parce que les prédictions nous ont permis d'établir des hypothèses biologiques qui nous ont bien guidés dans nos tests expérimentaux. Les connaissances chez d'autres organismes sur les homologues de ces protéines, ainsi que des expériences qui n'auraient pas été guidées par les prédictions auraient peut être abouti aux mêmes résultats. Cependant, la découverte des rôles pléïotropes aurait été très difficile et le temps et le coût investi auraient été plus élevés.

CAAX

Nous avons concentré nos efforts sur la validation expérimentale des liens prédits pour PepF et YvjB, néanmoins, des liens ont été prédits également pour la prényl-peptidase CAAX. Cette protéine présente plusieurs particularités. Tout d'abord nous avons constaté la présence de plusieurs copies plus ou moins similaires dans le génome de *L. lactis* IL1403 et un l'indice CAI pour le gène diffère de la moyenne chez cette bactérie (résultats préliminaires présentés dans la section de revue bibliographique). Quant aux liens prédits nous constatons des liens avec un grand nombre protéines de prophages. Ensuite, le voisinage des gènes codant pour les différentes copies de la prényl-

peptidase CAAX sont très variables et souvent entourés de transposons chez le lactocoque, mais aussi chez d'autres organismes. Ces résultats indiquent que CAAX est probablement acquise par transferts horizontaux. Si nous avions une hypothèse à tester ce serait la participation de CAAX à l'acquisition d'ADN étranger issu d'un transfert horizontal.

Analyse de sensibilité

Le fait que l'analyse statistique ait été réalisée avec une grande quantité de données post-génomiques nous amène à nous poser la question de l'importance de chacune de ces données. L'analyse de sensibilité réalisée à la fin de ce travail nous a permis de retrouver un sous-ensemble de données important pour la prédiction des liens d'une protéine cible donnée. Le fait de connaître l'apport de chaque variable est utile pour l'interprétation biologique des prédictions. L'analyse de sensibilité nous semble une démarche intéressante et originale pour déterminer l'importance de données d'entrée. Nous l'avons réalisée sur une source de données, conditionnellement à l'autre source de données. Néanmoins, il aurait été possible d'une part de perturber plus d'une source de données en même temps. Cela aurait été possible en modifiant le plan d'expériences. De la même manière nous aurions pu étudier l'influence des autres sources de données, par exemple du graphe des voies métaboliques, en enlevant certaines protéines de l'ensemble d'apprentissage.

9 Conclusions

- L'inférence statistique de liens entre protéines par analyse de corrélation canonique bâtie sur des noyaux (KCCA) est utile pour émettre des hypothèses biologiques sur le rôle de protéines et guider les validations expérimentales de rôles prédits.
- Le rôle principal plus probable de l'endopeptidase PepF chez *L. lactis* est la dégradation des peptides signaux des protéines sécrétées.
- PepF possède des rôles pléïotropes ; elle est impliquée dans la synthèse du peptidoglycane et le métabolisme du pyruvate.
- YvjB participe dans la mise en place de certaines lipoprotéines chez *L. lactis* et sa présence est nécessaire pour la maturation de ces lipoprotéines.
- L'analyse de sensibilité permet de définir l'importance des données utilisées pour l'inférence statistique et d'établir des sous-ensembles de données qui prédisent les mêmes liens entre protéines. Cela peut aider à la formulation d'hypothèses biologiques d'une parte et inviter à la vérification de la qualité des données, d'autre part.
- D'une manière générale cette étude confirme l'intérêt d'un travail commun et complémentaire entre biologistes et mathématiciens.

10 Perspectives

Ce travail ouvre des perspectives en relation avec l'inférence statistique du rôle des protéines et avec le rôle des protéines en soi.

Les résultats de la méthode KCCA ont permis l'obtention de distances entre toutes les protéines de *L. lactis* IL1403. Des prédictions de liens sont alors disponibles pour toutes ces protéines. La

première perspective de ce travail est alors d'explorer ces données existantes pour proposer des hypothèses biologiques et les valider expérimentalement.

Compte tenu du fait que les prédictions obtenues sont satisfaisantes et que la flexibilité de la KCCA permet l'utilisation de plusieurs types de données, elle devrait être étendue chez le même organisme modèle pour d'autres types de données comme les connaissances sur les régulons, les motifs conservés de protéines, données d'interaction physique entre protéines, etc., mais surtout à d'autres organismes. Pour cela il faudrait automatiser la récupération, construire et mettre en forme des données et développer de nouveaux noyaux si de nouvelles sources de données étaient utilisées. Actuellement un CDD a été recruté à MIA pour automatiser l'outil et étendre son utilisation à *Bacillus subtilis.*

Maintenant que d'autres méthodes d'inférence de rôle de protéines, comparables à la KCCA, existent (Aerts *et al.*, 2006; Bleakley *et al.*, 2007; De Bie *et al.*, 2007; Werhli & Husmeier, 2007), il serait intéressant de comparer les prédictions de ces méthodes concernant nos deux protéines cibles PepF et YvjB.

Ce travail conduit à des avancées dans les connaissances sur PepF et YvjB chez le lactocoque. Les mécanismes précis restent à déterminer. Pour le cas de PepF, son mode de participation à la synthèse du peptidoglycane et au métabolisme du pyruvate reste à établir. Des analyses plus précises sur la composition de la paroi et les métabolites produits à chaque étape du métabolisme du pyruvate pourraient nous éclairer sur ces mécanismes.

De la même manière nous ne connaissons pas le mécanisme exact par lequel YvjB agit sur les lipoprotéines qu'elle pend en charge. Il paraît difficile de reproduire la réaction entre YvjB et les lipoprotéines *in vitro,* parce que YvjB est une protéine membranaire interagissant fortement avec la signal peptidase II. Un autre moyen de comprendre le mécanisme d'action d'YvjB serai de caractériser les fragments libérés ce qui n'est pas simple non plus du fait de leur petite taille et leur nature hydrophobe.

11 Références

Aerts S., Lambrechts D., Maity S., Van Loo P., Coessens B., De Smet F., Tranchevent L. C., De Moor B., Marynen P., Hassan B., Carmeliet P., et Moreau Y. (2006). Gene priotitization throgh genomic data fusion. Nature Biotechnology 24: 537-544.

Ahn T., and Kim H. (1996). Differential effect of precursor ribose binding protein of Escherichia coli and its signal peptide on the SecA penetration of lipid bilayer. The Journal of Biological Chemistry 271: 12372-12379.

Akiyama Y., Kanehara K., et Ito K. (2004). RseP (YaeL), an Escherichia coli RIP protease, cleaves transmembrane sequences. The EMBO Journal 23: 4434-4442.

An F. Y., et Clewell D. B. (2002). Identification of the cAD1 sex pheromone precursor in Enterococcus faecalis. Journal of Bacteriology 184: 1880-1887.

An F. Y., Sulavik M. C., et Clewell D. B. (1999). Identification and characterization of a determinant (eep) on the Enterococcus faecalis chromosome that is involved in production of the peptide sex pheromone cAD1. J. Bacteriol. 181: 5915-5921.

Antiporta M. H., et Dunny G. M. (2002). ccfA, the genetic determinant for the cCF10 peptide pheromone in Enterococcus faecalis OG1RF. Journal of Bacteriology 184: 1155-1162.

Aronszajn N. (1950). Theory of kernels. Transactions of the American Mathematical Society 68: 337-404.

Atrih A., Bacher G., Allmaier G., Williamson M. P., et Foster S. (1999). Analysis of peptidoglycan structure from vegetative cells of Bacillus subtilis 168 and role of PBP 5 in peptidoglycan maturation. Journal of Bacteriology 181: 3956-3966.

Bach F. R., et Jordan M. I. (2002). Kernel independent component analysis. Machine Learning Research 3: 1-48.

Barrett A. J., et Rawlings N. D. (2001). Evolutionary lines of cystein peptidases. Biol Chem 382: 727-733.

Ben-Bassat A., Bauer K., Chang S. Y., Myambo K., Boosman A., et Chang S. (1987). Processing of the initiation methionine from proteins: properties of the Escherichia coli methionine aminopeptidase and its gene structure. Journal of Bacteriology 169: 751-757.

Bendtsen J. D., Kiemer L., Fausboll A., et Brunak S. (2005). Non-classical protein secretion in bacteria. BMC Microbiol 5, doi:10.1186/1471-2180-5-58: 58.

Bendtsen J. D., Nielsen H., von Heijne G., et Brunak S. (2004). Improved prediction of signal peptides: SignalP. Journal of Molecular Biology 340: 783-795.

Bhasvar A. P., et Brown E. D. (2006). Cell wall assembly in Bacillus subtilis: how spirals and spaces challenge paradigms. Molecular Biology 60: 1077-1090.

Bleakley K., Biau G., et Vert J. P. (2007). Supervised reconstruction of biological networks with local models. Bioinformatics 23: i57-i65.

Bohn C., Collier J., et Bouloc P. (2004). Dispensable PDZ domain of Escherichia coli YaeL essential protease. Molecular Microbiology 52: 427-435.

Bolhius A., Matzen A., Hyyryläinen H. L., Kontinen V. P., Meima R., Chapuis J., Venema G., Bron S., Freudl R., et Van Dijl J. M. (1999). Signal Peptide Peptidase- and ClpP-like proteins of Bacillus subtilis required for efficient translocation and processing of secretory proteins. The Journal of Biological Chemistry 274: 24585-24592.

Bolotin A., Wincker P., Mauger S., Jaillon O., Malarme K., Weissenbach J., Ehrlich S. D., et Sorokin A. (2001). The complete genome sequence of the lactic acid bacterium Lactococcus lactis ssp. lactis IL1403. Genome Research 11: 731-753.

Borgwardt K. M., Gretton A., Rasch M. J., Kreigel H. P., Schölkopf B., et Smola A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics 22: 49-57.

Box G. E. P., Hunter W. G., et Hunter J. S. (1978). "Statistics for experimenters, an intriduction to design, data analysis and model building."

Bramkamp M., Weston L., Richard D., et Errington J. (2006). Regulated intramembrane proteolysis of FtsL protein and the control of cell division in Bacillus subtilis. Molecular Microbiology 62: 580-591.

Braz V. S., Lang E. A. S., et Marques M. V. (2002). Cloning and characetrization of the gene encoding the PepF endopeptidase from the aquatic bacterium Caulobacter crescentus. Brazilian Journal of Microbiology 33: 84-91.

Brown M. S., Jin Y., Rawson R. B., et Goldstein J. L. (2000). Regulated Intramembrane Proteolysis: A Control Mechanism Conserved from Bacteria to Humans. Cell 100: 391-398.

Bühlmann P. (2006). Boosting for high-dimensional linear models. Ann. Statist. 34: 559-583.

Bühlmann P. (2007). "Variable selection."

Buist G., Ridder A. N. J. A., Kok J., et Kuipers O. P. (2006). Different subcellular locations of secretome components of Gram-positive bacteria. Microbiology 152: 2867-2874.

Bunai K., Yamada K., Hayashi K., Nakamura K., et Yamane A. (1999). Enhancing effect of Bacillus subtilis Ffh, a homologue of the SRP54 subunit of the mammalian signal recognition particle, on the binding of SecA to precursors of secretory proteins in vitro. Journal of Biochemistry (Tokio) 125: 151-159.

Buttaro B. A., Antiporta M. H., et Dunny G. M. (2000). Cell-associated pheromone peptide (cCF10) production and pheromone inhibition in Enterococcus faecalis. Journal of Bacteriology 182: 4926-4933.

Cadiñanos J., Schmidt W. K., Fueyo A., Varela I., Lopez-Otin C., et Freije J. M. P. (2003a). Identification, functional expression and enzymic analysis of two distinct CaaX proteases from Caenorhabditis elegans. Biochem. J. 370: 1047-1054.

Cadiñanos J., Varela I., Mandel D. A., Schmidt W. K., Diaz-Perales A., Lopez-Otin C., et Freije J. M. P. (2003b). AtFACE-2, a functional prenylated protein protease from Arabidopsis thaliana related to mammalian Ras-converting enzymes. J. Biol. Chem. 278: 42091-42097.

Carlsson F., Stahlhammar-Carlemalm M., Flärdh K., Sandin C., Carlemalm E., et Lindahl G. (2007). Signal sequence directs localized secretion of bacterial surface proteins. Nature 442: 943-946.

Chandler J. R., et Dunny G.M. (2008). Characerization of the sequence specificity determinants required for processing and control of sex pheromone by the intramembrane protease Eep and the plasmid-encoded protein PrgY. Journal of Bacteriology 190: 1172-1183.

Chao S. H., Cheng T. H., Shaw C. Y., Lee M. H., Hsu Y. H., et Tsai Y. C. (2006). Characterization of a Novel PepF-Like Oligopeptidase Secreted by Bacillus amyloliquefaciens 23-7A. Applied and Environmental Microbiology 72: 968-971.

Chapot-Chartier M. P., Nardi M., Chopin M. C., Chopin A., et Gripon J. C. (1993). Cloning and sequencing of pepC, a cystein aminopeptidase from Lactococcus lactis subsp cremoris AM2. Appl Environ Microbiol 59: 330-333.

Chatel J. M., Langella P., Adel-Patient K., Commissaire J., Wal J. M., et Corthier G. (2001). Induction of mucosal immune response after intranasal or oral inoculation of mice with Lactococcus lactis producing bovine beta-lactoglobulin. Clin. Diagn. Lab. Immunol. 8: 545-551.

Chen J. C., P.H. V., et Shapiro L. (2005). A membrane metalloprotease participates in the sequential degradation of a Caulobacter polarity determinant. Molecular Biology 55: 1085-1103.

Chen L., Tai P. C., Briggs M. S., et Gierasch L. M. (1987). Protein Translocation into Escherichia coli Membrane Vesicles is Inhibited by Functional Synthetic Signal Peptides. Journal of Biological Chemistry 262: 1427-1429.

Chich J. F., David O., Villers F., Schaeffer B., Lutomski D., et Huet S. (2007). Statistics for proteomics: Experimental design and 2-DE differential analysis. Journal of Chromatography 849: 261-272.

Chopin A., Chopin M.-C., Moillo-Batt A., and Langella P. (1984). Two plasmid-determined restriction and modification systems in Streptococcus lactis. Plasmid 11: 260-263.

Chung Y. S., et Dubnau D. (1995). ComC is required for the processing and translocation of comGC, a pilin-like competence protein of Bacillus subtilis. Mol. Microbiol. 15: 543-551.

Cladman W. M., Watt M. A., Dini J. P., et Mellors A. (1996). The Pasteurella haemolytica O-sialoglycoprotein endopeptidase is inhibited by zinc ions and does not cleave fetuin. Biochem Biophys Res Commun 7: 141-146.

Clarke S. (1992). Protein isoprenylation and methylation at carboxy-terminal cystein residues. Annu Rev Biochem 61: 355-386.

Cocaign-Bousquet M., Garrigues C., Loubière P., et Lindley N. D. (1996). Physiology of pyruvate metabolism in Lactococcus lactis. Antonie Van Leeuwenhoek 70: 253-267.

Cocaign-Bousquet M., Garrigues C., Novak L., Lindley N. D., et . P. L. (1995). Rational development of a simple synthetic medium for the sustained growth of Lactococcus lactis. Journal of Applied Bacteriology 79: 108-116.

Conlin C. A., et Miller C. G. (2000). opdA, a Salmonella enterica Serovar Typhimurium Gene Encoding a Protease, Is Part of an Operon Regulated by Heat Shock. Journal of Bacteriology 182: 518-521.

Conlin C. A., Trun N., Silhavy T. J., et Miller C. G. (1992a). Escherichia coli prlC Encodes an Endopeptidase and Is Homologous to the Salmonella typhimurium opdA Gene.

Conlin C. A., Vimr E. R., et Miller C. G. (1992b). Oligopeptidase A is required for normal phage P22 development. Journal of Bacteriology 174: 5869-5880.

Courtin P., Miranda G., Guillot A., Wessner F., et Mézange C. (2006). Peptidoglycan Structure Analysis of Lactococcus lactis Reveals the presence of an L,D-Carboxypeptidase Involved in Peptidoglycan Maturation. Journal of Bacteriology 188: 5293-5298.

Covert M. W., Knight E. M., Reed J. L., Herrgard M. J., et Palsson B. O. (2004). Integrating high-throughput et computational data elucidates bacterial networks. Nature 429: 92-96.

Dalbey R., et Wickner S. (1992). Leader Peptidase catalyses the release of exported proteins from the outer surface of the Escherichia coli plasma membrane. J. Biol. Chem. 260: 15925-15931.

De Bie T., Tranchevent L. C., van Oeffelen L. M. M., et Moreau Y. (2007). Kernel-based data fusion for gene prioritization. Bioinformatics 23: i125-i132.

de Gier J. W., and Luirink J. (2001). Biogenesis of inner membrane proteins in Escherichia coli. Molecular Microbiology 40: 314-322.

de Jong H. (2002). Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. Journal of Computational Biology 9: 67-103.

de Magalhães J. P., Costa J., et Toussaint O. (2005). "HAGR: the Human Aging genomic Resources". Nucleic Acids Res 33: D537-D543.

De Ruyter P. G. G. A., Kuipers O. P., Beerthuyzen M. M., van Alen-Boerrigter I., et Vos W. M. (1996). Functional Analysis of Promoters in the Nisin Gene Cluster of Lactococcus lactis. Journal of Bacteriology 178: 3434-3439.

den Hengst C. D., Curley P., Larsen R., Buist G., Nauta A., van Sinderen D., Kuipers O. P., et Kok J. (2005a). Probing Direct Interactions between CodY and the oppD Promoter of Lactococcus lactis. Journal of Bacteriology 187: 512-521.

den Hengst C. D., van Hijum S., Geurts J. M. W., Nauta A., Kok J., et Kuipers O. P. (2005b). The Lactococcus lactis CodY Regulon. The Journal of Biological Chemistry 280: 34332-34342.

Denham E. L., Ward P. N., et Leigh J. A. (2008). Lipoprotein Signal Peptides Are Processed by Lsp and Eep of Strepotococcus uberis. Journal of Bacteriology 190: 4641-4647.

Dev I. K., et Ray P. H. (1990). Signal Peptidases and Signal Peptide Hydrolases. Journal of Bioenergetics and Biomembranes 22: 271-290.

Dridi N. (2008). Prédiction du rôle de Protéine cible à l'aide de méthodes d'apprentissage local, Unité MIA, INRA, Jouy en Josas (France), Tunis (Tunisie).

Dudley E. G., Husgen A. C., He W., et Steele J. L. (1996). Sequencing, distribution, and inactivation of the dipeptidase A gene (pepDA) from Lactobacillus helveticus. Journal of Bacteriology 178: 701-704.

Eisner G., Moser M., Schäfer U., Beck K., et Müller M. (2006). Alternate Recruitment of Signal Recognition Particle and Trigger Factor to the Signal Sequence of a Growing Nascent Polypeptide. The Journal of Biological Chemistry 281: 7172-7179.

Emr S. D. (1982). Localization and Processing of Outer Membrane and Periplasmic Proteins in Escherichia coli Strains Harbouring Export-specific Suppressor Mutations. The Journal of Biological Chemistry 257: 5852-5860.

Emr S. D., et Silhavy T. J. (1980). Mutations affecting localization of Escherichia coli outer membrane protein, the becteriophage lambda receptor. Journal of Molecular Biology 141: 63-90.

Esnouf R. M. (1999). Further additions to MolScript version 1.4, including reading and contouring of electron-density maps. Acta Crystallogr D Biol Crystallogr 55: 938-940.

Frees D., Varmanen P., et Ingmer H. (2001). Inactivation of a gene that is highly conserved in Gram-positive bacteria stimulates degradation of non-native proteins and concomitantly increases stress tolerance in Lactococcus lactis. Mol Microbiol 93-103.

Freund Y., et Schapire R. E. (1999). A short introduction to boosting. Journal of Japanese Society of Artificial Intelligence 14: 771-780.

Friedman N., Linial M., Nachman I., et Pe'er D. (2000). Using Bayesian networks to analyze expression data. Journal of Computational Biology 7: 601-620.

Fuh G., Pisabarro M. T., Li Y., Quan C., Lasky L. A., et Sidhu S. S. (2000). Analysis of PDZ domain-ligand interactions using carboxy-terminal phage display. J. Biol. Chem. 275: 21486-21491.

Fukuchi S., et Nishikawa K. (2004). Estimation of the number of orphan genes in bacterial genomes. DNA Research 11: 219-231.

Gillespie J. R., Yokoyama K., Lu K., Eastman R. T., Bollinger J. G., Van Voorhis W. C., Gelb M. H., et Buckner F. S. (2007). C-terminal proteolysis of prenylated proteins in trypanosomatids and RNA interference of enzymes required for the post-translational processing pathway of farnesylated proteins. Mol. Biochem. Parasitol. 153: 115-124.

Gitton C., Meyrand M., Wang J., Caron C., Trubuil A., Guillot A., et Mistou M.-Y. (2005).

Proteomic Signature of Lactococcus lactis NCDO763 Cultivated in Milk. Appl Environ Microbiol 71: 7152-7163.

Glauner B. (1988). Separation and quantification of muropeptides with high-performance liquidi chromatography. Analytical Biochemistry 172: 451-464.

Gonzales T., et Robert-Baudony J. (1996). Bacterial aminopeptidases: Properties and functions. FEMS Micorbiology Reviews 18: 319-344.

Gottesman S. (1996). Proteases and their targets in Escherichia coli. Annu Rev Genet 30: 465-506.

Guedon E., Renault P., Ehrlich D., et Delorme C. (2001a). Transcriptional Pattern od genes coding for the proteolytic system of Lactococcus lactis and evidence for coordinated regulation of key enzymes by peptide supply. Journal of Bacteriology 183: 3614-3622.

Guedon E., Serror P., Ehrlich D., Renault P., et Delorme C. (2001b). Pleiotropic transcriptional repressor CodY senses the intracellular pool of branched-chain amino acids in Lactococcus lactis. Mol. Microbiol. 40: 1227-1239.

Guillot A., Gitton C., Anglade P., et Mistou M. Y. (2003). Proteomic analysis of Lactococcus lactis, a lactic acid bacterium. Proteomics 3: 337-354.

Gutierrez J. A., Crowley P. J., Cvitkovitch D. G., Brady L. J., Hamilton I. R., Hillman J. D., et Bleiweis A. S. (1999). Streptococcus mutans ffh, a gene encoding a homoligue of the 54 kDa subunit of the signal recognition particle, is involved in resistance to acid stress. Microbiology 145: 357-366.

Gyaneshwar P., Paliy O., McAuliffe J., Popham D. L., Jordan M. I., et Kustu S. (2005). Sulfur and Nitrogen Limitation in Escherichia coli K-12: Specific Homeostatic Responses. Journal of Bacteriology 187: 1074-1090.

Hiron A., Borezée-Durant E., Piard J. C., et Juillard V. (2007). Only one of four oligopeptide transport systems mediates nitrogen nutrition in Staphylococcus aureus. Journal of Bacteriology 189: 5119-5129.

Hlavacek O., et Vachoa L. (2002). ATP-dependent proteinases in bacteria. Folia Microbiologica 47: 203-212.

Ho S. N., Hunt H. D., Horton R. M., Pullen J. K., et Pease L. R. (1989). Site-directed mutagenesis by overlap extention using the polymerase chain reaction. Gene 77: 51-59.

Hoffman C. S., et Winston F. (1987). A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of Escherichia coli. Gene

57: 267-272.

Kanamaru K., Stephenson S., and Perego M. (2002). Overexpression of the PepF oligopeptidase inhibits sporulation initiation in Bacillus subtilis. J. Bacteriol. 184: 43-50.

Kato T., Tsuda K., et Asai K. (2005). Selective integration of multiple biological data for supervised network inference. Bioinformatics 21: 2488-2495.

Kauffman S. A. (1993). "The origins of order: self-organization and selection in evolution," Oxford University Press, New York.

Khandavilli S., Homer K. A., Yuste J., Basavanna S., Mitchell T., et Brown J. S. (2008). Maturation of Streptococcus pneumoniae lipoproteins by a type II signal peptidase is required for ABC transporter function and full virulence. Molecular Microbiology 67: 541-557.

Kim S., Yoon J., et Yang J. (2008). Kernel approaches for genic interaction extraction. Bioinformatics 24: 118-126.

Kinch L. N., Ginalski K., et Grishin N. V. (2006). Site-2 protease regulated intramembrane proteolysis: Sequence homologs suggest an ancient signaling cascade. Protein Science 15: 84-93.

Kondor R. I., et Lafferty J. (2002). Diffusion kernels on graphs and other discrete structures. In "Machine learning: Proceedings of the 19th International conference" (C. Sammut, and A. G. Hoffmann, Eds.), pp. 315-322, Morgan Kaufmann, San Francisco.

Kraut J. (1977). Serine Proteases: Structure and mechanism of catalysis. Ann. Rev. Biochem. 46: 331-358.

Kunji E. R., Fang G., Margot P., Bruins A. P., Poolman B., et Konings W. N. (1998). Reconstruction of the proteolytic pathway for use of beta-casein by Lactococcus lactis. Molecular Microbiology 27: 1107-1118.

Lamarque M., Charbonnel P., Aubel D., Piard J. C., Atlan D., et Juillard V. (2004). A Multifunction ABC Transporter (Opt) Contributes to Diversity of Peptide Uptake Spacificity within the genus Lactococcus. Journal of Bacteriology 186: 6492-6500.

l'Anson K. J., Movahedi S., Griffin H. G., Gasson M. J., et Mulholland F. (1995). A non-essential glutamyl aminopeptidase is required for optimal growth of Lactococcus lactis MG1363 in milk. Microbiology 141: 2873-2881.

Lauritzen S. (1996). Graphical models. In "Oxford Statistical Science Series, 17", Oxford.

Lee W., McDonough M. A., Kotra L. P., Li Z. H., Silvaggi N. R., Takeda Y., Kelly J. A., et Mobashery S. (2001). A snapshot of the final step of bacterial cell wall synthesis. Proc Natl Acad Sci 98: 1427-1431.

Lee H. C., and Bernstein H. D. (2001). The targeting pathway of Escherichia coli presecretory and integral membrane proteins is specified by the hydrophobicity of the targeting signal. Proc Natl Acad Sci 98: 3471-3476.

Loessner M. J. (2006). Bacteriophage endolysins-current state of research and applications. Curr. Opin. Microbiol. 8: 480-487.

Mars I., et Monnet V. (1995). An aminopeptidase P from Lactococcus lactis with original specificity. Biochim Biophys Acta 1243: 209-215.

Mazmanian S. K., Ton-That H., et Schneewind O. (2001). Sortase-catalysed anchoring of surface proteins to the cell wall of Staphylococcus aureus. Mol Microbiol 40: 1049-1057.

Meinshausen N., et Bühlmann P. (2006). High-Dimensional Graphs and Variable Selection with the Lasso. The Annals of Statistics 34: 1436-1462.

Metrione R. M., Neves A. G., et Fruton J. S. (1966). Purification and properties of dipeptidyl transferase (cathepsin C). Biochemistry 5: 1597-1604.

Mierau I., Haandrickman A. J., Velterop O., Tan P. S. T., Leenhoust K., Konings W. N., et Venema G. (1994). Tripeptidase gene (pepT) from Lactococcus lactis: molecular cloning and nucleotide sequencing of pepT and construction of a chromosomal deletion mutant. Journal of Bacteriology 176: 2854-2861.

Mierau I., Kunji E. R. S., Leenhouts K. J., Hellendorn M. A., Haandrikman A. J., Poolman B., Konings W. N., Venema G., and Kok J. (1996). Multiple-peptidase mutants of Lactococcus lactis are severely impaired in their ability to grow in milk. Journal of Bacteriology 178: 2794-2803.

Mierau I., Tan P. S. T., Haandrickmaan A. J., Kok J., Leenhouts K. J., Konings W. N., et Venema G. (1993). Cloning and sequencing of the gene for a lactococcal endopeptidase, an enzyme with sequence similarity to mammalian enkephalinase. Journal of Bacteriology 175: 2087-2096.

Miyake R., Shigheri Y., Tatsu Y., Yumoto N., Umekawa M., Tsujimoto Y., Matsui H., et Watanabe K. (2005). Two Thimet Oligopeptidase-Like Pz Peptidases Produced by a Collagen-Degrading Thermophile, Geobacillus collagenovorans MO-1. Journal of Bacteriology 187: 4140-4148.

Miyoshi S., et Shinoda S. (2000). Microbial metalloproteases and pathogenesis. Microbes Infect 2: 91-98.

Mogk A., Schmidt R., et Bukau B. (2007). The N-end rule pathway for regulated proteolysis: prokaryotic and eukaryotic strategies. Trends Cell Biol 17: 165-172.

Monnet V., Nardi M., Chopin A., Chopin M. C., et Gripon J. C. (1994). Biochemical and genetic characterization of PepF, an oligopeptidase from Lactococcus lactis. Journal of Biological Chemistry 268: 32070-32076.

Nardi M., Chopin M. C., Chopin A., Cals M. M., et Gripon J. C. (1991). Cloning and DNA sequence analysis of an X-prolyl dipetidyl aminopeptidase gene from Lactococcus lactis subsp. lactis NCDO763. Appl Environ Microbiol 57: 45-50.

Nardi M., Renault P., et Monnet V. (1997). Duplication of the pepF gene and shuffling of DNA fragments on the lactose plasmid of Lactococcus lactis. J. Bacteriol. 179: 4164-4171.

Narita S. I., Matsuyama S. I., et Tokuda H. (2004). Lipoprotein trafficking in Escherichia coli. Arch Microbiol 182: 1-6.

Nelson D., et Cox M. (2000). "Lehninger Principles of Biochemistry."

Neviani E., Boquien C. Y., Monnet V., Phan Than L., et Gripon J. C. (1989). Purification and characterization of an aminopeptidase from Lactococcus lactis subsp. cremoris AM2. Appl Environ Microbiol 55: 2308-2314.

Niven G. W., Holder S. A., et Stroman P. (1995). A study of the substrate specificity of aminopeptidases N from Lactococcus lactis subsp. cremoris Wg2. Appl Microbiol Biotechnol 44: 100-105.

Novak P., Ray P. H., et Dev I. K. (1986). Localization et Purification of Two Enzymes from Escherichia coli Capable of Hydrolyzing a Signal Peptide. The Journal of Biological Chemistry 261: 420-427.

Paetzel M., Karla A., Strynadka N., et Dalbey R. (2002). Signal peptidases. Chem. Rev. 102: 4549-4580.

Papanikou E., Karamanou S., et Economou A. (2007). Bacterial protein secretion through the translocase nanomachine. Nature Reviews, Microbiology 5: 839-851.

Pei J., et Grishin N. V. (2001). Type II CAAX prenyl endopeptidases belong to a novel superfamily of putative membrane-bound metalloproteases. Trends in Biochemical Sciences 26: 275-277.

Pestova E. V., Havarstein L. S., et Morrison D. A. (1996). regulation of competence for genetic transformation in Streptococcus pneumoniae by an auto-induced peptide pheromone and a two-component regulatory system. Mol.

Microbiol. 21: 853-862.

Piard J. C., Hautefort I., Fischetti V. A., Ehrlich D., Fons M., et Gruss A. (1997). Cell Wall Anchoring of the Streptococcus pyogenes M6 Protein in Varoious Lactic Acid Bacteria. Journal of Bacteriology 179: 3068-3072.

Piuri M., Sanchez-Rivas C., et Ruzal S. M. (2005). Cell wall modifications during osmotic stress in Lactococcus caseii. J. Appl. Microbiol. 98: 84-95.

Poolman B., Kunji E. R., Hagting A., Juillard V., et Konings W. N. (1995). The proteolytic pathway of Lactococcus lactis. j Appl Bacteriol Sym Suppl 79: 65S-75S.

Poquet I., Saint V., Seznec E., Simoes N., Bolotin A., et Gruss A. (2000). HtrA is the unique surface housekeeping protease in Lactococcus lactis and is required for natural protein processing. Microbiology 35: 1042-1051.

Price C. E., and Driessen A. J. M. (2008). YidC is involved in the biogenesis of anaerobic respiratory complexes in the inner membrane of Escherichia coli. J Biol Chem in press.

Qi F., Klein-Seetharaman J., et Bar-Joseph Z. (2005). Random forest similarity for protein-protein interaction prediction from multiple sources. In "Biocomputing, Proceeding of the Pacific Symposium", Hawaii, USA.

Rasmussen M., et Bjorck L. (2002). Proteolysis and its regulation at the surface of Streptococcus pyogenes.

Rawlings N. D., et Barrett A. J. (1995). Evolutionary families of metallopeptidases. Methods in Enzymology 248: 183-228.

Redko Y., Courtin P., Mézange C., Huard C., et Chapot-Chartier M. P. (2007). Lactococcus lactis Gene yjgB Encodes a gamma-D-Glutaminyl-L-Lysyl-Endopeptidase Which Hydrolyzes Peptidoglycan. Appl Environ Microbiol 73: 5825-5831.

Renault P., Corthier G., Goupil N., Delorme C., et Ehrlich S. D. (1996). Plasmid vectors for Gram-positive bacteria switching from high to low copy number. Gene 183: 175-182.

Rhazi N., Charlier P., Dehareng D., Engher D., Vermeire M., Frère J. M., Nguyen-Distèche M., et Fonzé E. (2003). Catalytic mechanism of the Streptomyces K15 DD-transpeptidase/penicillin-binding protein probed by site-directed mutagenesis and structural analysis. Biochemistry 42: 2895-2906.

Roth F. P., Hughes J. D., Estep P. W., et Church G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol 16: 939-945.

Rudner D. Z., Fawcette P., et Losick R. (1999). A family of membrane-embedded metalloproteases involved in regulated proteolysis of membrane-associated transcription factors. Proc Natl Acad Sci 96: 1007-1018.

Saltelli A., Chan K., et Scott E. M. (2000). "Sensitivity Analysis," Wiley.

Sambrook J., et Russel D. W. (2001). "Molecular cloning: a laboratory manual," Cold Spring Harbor Laboratory Press.

Samuelson J. C., Chen M., Jiang F., Moller I., Wiedmann M., Kuhn A., Phillips G. J., and Dalbey R. E. (2000). YidC mediates membrane protein insertion in bacteria. nature 406: 637-641.

Savijoki K., Ingmer H., et Varmanen P. (2006). Proteolytic systems of lactic acid bacteria. Appl Microbiol Biotechnol 71: 394-406.

Schäfer J., et Strimmer K. (2005). An empirical Bayes approach to inferring larg-scale gene association networks. Bioinformatics 21: 754-764.

Schleifer K. H., et Kandler O. (1972). Peptidoglycan types of bacterial cell walls and their taxonomic implications. Bacteriol Rev 36: 407-477.

Schöbel S., Zellmeier S., Schumann W., et Wiegert T. (2004). The Bacillus subtilis sigmaW anti-sigma factor RsiW is degraded by intramembrane proteolysis through YluC. Molecular Microbiology 52: 1091-1105.

Scott J. R., et Barnett T. C. (2006). Surface Proteins of Gram-Positive Bacteria and How They Get There. Annu. Rev. Microbiol. 60: 397-423.

Seydel A., Gounon P., and Pugsley A. (1999). Testing the `.2 rule' for lipoprotein sorting in the Escherichia coli cell envelope with a new genetic selection. Molecular Microbiology 34: 810-821.

Siezen R. J. (1999). Multi-domain, cell-envelope proteinases of lactic acid bacteria. Antonie Van Leeuwenhoek 76: 139–155.

Siezen R. J., Renckens B., van Swam I., Peters S., van Kranenburg R., Kleerebezem M., et de Vos W. M. (2005). Complete Sequences of Four Plasmids of Lactococcus lactis subsp. cremoris SK11 Reveal Extensive Adaptation to the Dairy Environment. Appl Environ Microbiol 71: 8371-8382.

Stoll H., Dengjel J., Nerz C., et Götz F. (2005). Staphylococcus aureus Deeficient in Lipidation of Prelipoproteins Is Attenuated in Growth and Immune Activation. Infection and immunity 73: 2411-2423.

Storer C., et Ménard R. (1994). Catalytic Mechanism in Papain Family of Cystein

Peptidases. Methods in Enzymology 224: 486-500.

Suguna K., Padlan E. A., Smith C. W., Carlson W. D., et Davies D. R. (1987). Binding of a reduced peptide inhibitor to the aspartic proteinase from Rhizopus chinensis: implications for a mechanism of action. Proc Natl Acad Sci 84: 7009-7013.

Sutcliffe I. C., et Russel R. R. B. (1995). Lipoproteins of Gram-Positive Bacteria. Journal of Bacteriology 177: 1123-1128.

Tan P. S. T., Chapot-Chartier M. P., Pos K. M., Rousseau M., Boquien C. Y., Gripon J. C., and Konings W. N. (1992). Localization of peptidases in lactococci. Appl Environ Microbiol 58: 285-290.

Tan P. S. T., Poolman B., et Konings W. N. (1993). Proteolytic enzymes of Lactococcus lactis. J. Dairy Res. 60: 269-286.

Tan P. S. T., Pos K. M., et Konings W. N. (1991). Purification and characterization of an endopeptidase from Lactococcus lactis subsp. cremoris Wg2. Applied and Environmental Microbiology 57: 3593-3599.

Thomas T. D., Turner K. W., et Crow V. L. (1980). Galactose fermentation in Streptococcus lactis and Streptococcus cremoris: pathways, products and regulation. Journal of Bacteriology 144: 672-682.

Thompson J. D., Higgins D. G., et Gibbson T. J. (1994). Improving the sensitivity of progressive multiple sequence alignement through sequence, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.

Tisljar U., Knight C. G., et Barrett A. J. (1990). An Alternative Quenched Flourescence Substrate for Pz-peptidase. Anal. Biochem. 64: 112-115.

Tjalsma H., Bolhuis A., Jongbloed J. D., Bron S., et van Dijl J. M. (2000). Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome. Microbiol. Mol. Biol. Rev. 64: 515-547.

Tjalsma H., Kontinen V. P., Pragai Z., Wu H., Meima R., Venema G., Bron S., Sarvas M., et van Dijl J. M. (1999a). The role of lipoprotein processing by signal peptidase II in the Gram-positive Eubacterium Bacillus subtilis. The Journal of Biological Chemistry 274: 1689-1707.

Tjalsma H., Zanen G., Venema G., Bron S., et van Dijl J. M. (1999b). The Potential Active Site of the Lipoprotein-specific (Type II) Signal Peptidase of Bacillus subtilis. J Biol Chem 274: 28191-28197.

Tynkkynen S., Buist G., Kunji E. R., Kok J., Poolman B., Venema G., et Haandrickman A. J. (1993). Genetic and biochemical characterization of the oligopeptide transport system of Lactococcus lactis. J. Bacteriol. 175: 7523-7532.

Varmanen P., Ingmer H., et Vogensen F. K. (2000). CtsR of Lactococcus lactis encodes a negative regulator of clp gene expression. Microbiology 146: 1447-1455.

Venema R., Tjalsma H., van Dijl J. M., de Jong A., Leenhoust K., Buist G., et Venema G. (2003). Active Lipoprotein Precursors in the Gram-positive. The Journal of Biological Chemistry 278: 14739-14746.

Vert J. P., Tsuda K., et Schölkopf B. (2004). A Primer on Kernel Methods. In "Kernel Methods in Computational Biology" (B. Schölkopf, K. Tsuda, and J. P. Vert, Eds.), pp. 35-70, The MIT Press, Campbridge.

Verzelen N., et Villers F. (2007). Tests for gaussien graphical models. Computational Statistics ans Data Analysis Submitted.

Vimr E. R., Green L., et Miller C. G. (1983). Oligopeptidase-Deficient Mutants of Salmonella typhymurium. Journal of Bacteriology 153: 1259-1265.

Vogt V. M. (1970). Purification and properties of an aminopeptidase from Escherichia coli. Journal of Biological Chemistry 245: 4760-4769.

von Heijne G. (1989). The structure of signal peptides from bacterial lipoproteins. Protein Eng 2: 531-4.

von Heijne G., et Abrahamsen L. (1989). Species-specific variation in signal peptide design; implications for protein secretion in foreign hosts. FEBS Letters 244: 439-446.

Von Mering C., Jensen L. J., Kuhn M., Chaffron S., Doerks T., Krüger B., Snel B., et Bork P. (2007). STRING 7-recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 35: D358-D362.

Vongerichten K. F., J.R. K., Matern H., et Plapp R. (1994). Cloning and DNA sequence analysis of pepV, a carnosinase gene from Lactobacillus delbruecki subsp. bulgaricus. Int Dairy J 2: 345-361.

Walsh N. P., Alba B. M., Bose B., Gross C. A., et Sauer R. T. (2003). OMP peptide signals initiate the envelope-stress response by activating DegS protease via relief of inhibition mediated by its PDZ domain. Cell 113: 61-71.

Werhli A. V., et Husmeier D. (2007). Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge. Statistical Applications in Genetics and Molecular

Biology 6.

Westra R. L., Hollanders G., Bex G. J., Gyssens K., et Tuyls K. (2007). The Identification of Dynamic Gene-Protein Networks. In "Knowledge Discovery and Emergent Complexity in Bioinformatics", pp. 157-170, Springer, Berlin / Heidelberg.

Wickner W., Moore K., Dibb N., Geissert D., et Rice M. (1987). Inhibition of purified Escherichia coli leader peptidase by the leader (signal) peptide of bacteriophage M13 procoat. Journal of Bacteriology 169: 3821-3822.

Wydau S., Dervyn E., Anba J., Ehrlich D., et Maguin E. (2006). Conservation of key elements of natural competence in Lactococcus lactis ssp. FEMS Mcrobiology Letters 263: 223-228.

Yamaguchi K., Yu F., and Inouye M. (1988). A single amino acid determinant of the membrane localization of lipoproteins in E. coli. Cell 53: 423-432.

Yamanishi Y., Vert J. P., et Kanehisa M. (2004a). Heterogeneous Data Comparison and Gene Selection with Kernel Canonical Correlation Analysis. In "Kernel Methods in Computational Biology" (B. Schölkopf, K. Tsuda, et J. P. Vert, Eds.), pp. 209-229, The MIT Press, Cambridge.

Yamanishi Y., Vert J. P., et Kanehisa M. (2004b). Protein network inference from multiple genomic data: a supervised approach. Bioinformatics 20: i363-i370.

Yamanishi Y., Vert J. P., Nakaya A., et Kanehisa M. (2003). Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. Bioinformatics 19: i323-i330.

Zanen G., Houben E. N. G., Meima R., Tjalsma H., Jongbloed J. D., Westers H., Oudega B., Luirink J., van Dijl J. M., et Quax W. J. (2005). Signal peptide hydrophobicity is critical for early stages in protein export by Bacillus subtilis. FEBS Journal 272: 4617-4630.

Zwickl P., Seemüller E., Kapelari B., et Baumeister W. (2001). The proteasome: a supramolecular assembly designed for controlled proteolysis,. Adv. Protein Chem. 59: 187-222.

12 Présentations du travail

12.1 Présentations orales

1) Journée du groupe SSB (Statistics for Biosystems Biology), Juin 2008, Jouy en Josas.
Détermination du rôle de certaines peptidases bactériennes par inférence à partir de données hétérogènes et incomplètes. **Liliana Lopez, Véronique Monnet, Alain Trubuil**

2) Séminaire du LBBE (Laboratoire de Biométrie et Biologie Evolutive), Mai 2008, Lyon.
Détermination du rôle de certaines peptidases bactériennes par inférence à partir de données hétérogènes et incomplètes. **Liliana Lopez, Véronique Monnet, Alain Trubuil**

3) 15ème colloque des Bactéries Lactiques, Rennes, Novembre 2007.
Détermination du rôle de la peptidase bactérienne PepF. **Liliana Lopez, Alain Trubuil, Véronique Monnet.** Actes.

4) Journée MICALIS, Cellules en tant que système, Octobre 2007, Jouy en Josas.
Détermination du rôle de la peptidase bactérienne PepF. **Liliana Lopez, Alain Trubuil, Véronique Monnet.**

5) Tri-Partite Meeting, MIA(INRA)-BIOSS-Biometris, Juin 2007, Wageningen.
Determination of the role of a bacterial protein using a kernel method and further experimental validation. **Liliana Lopez, Alain Trubuil, Véronique Monnet.**

6) Journées de l'école doctorale ABIES, Mars 2007, Paris.
Détermination du rôle de la peptidase bactérienne PepF par inférence statistique. **Liliana Lopez, Alain Trubuil, Véronique Monnet.**

12.2 Posters

1) LAB9, Septembre 2008, Egmond aan Zee.
Role of bacterial peptidases inferred by statistical analysis and further experimental validation
Lopez L., Trubuil A., Guillot A., Pechoux C., Monnet V.

2) JOBIM (Journées Ouvertes de Biologie, Informatique et Mathématique), Juin 2008, Lille.
Finding important data subsets for kernel-based bacterial. **Liliana Lopez, Alain Trubuil, Véronique Monnet.** Actes.

1) JOBIM (Journées Ouvertes de Biologie, Informatique et Mathématique), Juillet 2007, Marseille.
Determination of the role of certain bacterial peptidases by inference from heterogeneous and incomplete data. **Liliana Lopez, Alain Trubuil, Véronique Monnet.** Actes.

2) IPG (Integrative Post-Genomics): International Multidisciplinary Meeting on Post-Genomics, Novembre 2006, Lyon.
Determination of the role of certain bacterial peptidases by inference from heterogeneous and incomplete data. **Liliana Lopez, Alain Trubuil, Véronique Monnet.**

3) 4èmes Rencontres des Microbiologistes INRA. Dourdan, Juin 2006.
Détermination du rôle de certaines peptidases bactériennes par inférence à partir de données hétérogènes et incomplètes. **Liliana Lopez, Véronique Monnet, Alain Trubuil**

4) 14ème colloque des Bactéries Lactiques, Paris, Mai 2006.
Détermination du rôle régulateur de certaines peptidases bactériennes par inférence à partir de données hétérogènes et incomplètes. **Liliana Lopez, Véronique Monnet, Alain Trubuil.** Actes.

13 Annexes

13 Annexes

13.1 Liens prédits par KCCA

13.1.1 PepF, données supplémentaires article I

Supplementary data

gene	function	K2K3K4	K2K3K4K5	distances	shared relationsh. with PepF on the network*
sipL	signal peptidase I	1	1	1.77**	rpiA, pyrC,glyA,yjhF
yibC	hypothetical protein (sulfatase domain)	1	1	1.78**	glyA, yjhF,pyrC, glK, rpiA
yufA	hypothetical protein	1	1	2.08**	rpiA
ycgG	oxidoreductase	1	1	2.10**	glyA, pyrC, glK
ysgC	hypothetical protein (polysaccharide synthetase domain)	1	1	2.11**	glyA, pyrC, glmU, glK
yfjE	flavodoxin	1	1	2.11**	pyrC,glmU
dltD	D-alanine transfer protein DltD	1	1	2.13**	glyA, pyrC, glmU, glK
ftsA	cell division protein FtsA	1	1	2.18**	glyA, pyrC, glmU
glK	glucose kinase	1	1	2.18**	
bmpA	basic membrane protein A	1	1	2.20**	glyA, yjhF, pyrC, glmU, glK
pknB	serine/threonine protein kinase	1	1	2.22**	pyrC,glmU
ylcC	similar to sortase	1	1	2.24**	glyA,pyrC
ydbF	putative transcription regulator	1	1	2.24**	glyA, pyrC, glmU, glK
yiaC	hypothetical protein (membrane domain)	1	1	2.24**	glyA, yjhF, pyrC, glK

exoA	exodeoxyribonuclease A	1	1	2.24**	glyA, yjhF, rpiA
yudL	hypothetical protein (ABC transp. permease do-main)	1	1	2.25**	pyrC,glmU
yiiB	hypothetical protein (methyltransferase domain)	1	1	2.25**	pyrC,glmU
recN	DNA repair protein RecN	1	1	2.26**	glyA, yjhF, glK, rpiA
ftsW	cell division protein FtsW	1	1	2.28**	glyA, pyrC, glmU
pfs	5'methylthioadenosine/S-adenosylhomocyseine nucl.	1	1	2.29**	glyA, yjhF, glK, rpiA
infB	translation initiation factor IF-2	1	1	2.30**	glyA
yeeG	transcription regulator	1	1	2.31**	
ylfD	putative autolytic lysozyme	1	1	2.33**	yjhF,glK
secA	preprotein translocase SecA subunit	1	1	2.34**	
pepM	methionine aminopeptidase	1	1	2.36**	glyA, pyrC, glmU, glK
nagB	glucosamine 6 P isomerase	1	1	2.37**	glyA,pyrC
ylfI	unknown protein	1	1	2.40**	murE,rpiA
uvrB	excinuclease ABC subunit B	1	1	2.41**	pyrC,glmU
mtsA	manganese ABC transporter substrate binding protein	1	1	2.42**	
plpD	outer membrane lipoprotein precursor	1	1	2.42**	glyA, murE, rpiA
nifZ	pyridoxal-phosphate dependent aminotransfe-rase	1	1	2.44**	pyrC,glmU
snf	SWI/SNF family helicase	1	1	2.45**	glyA, yjhF, glK, rpiA
ykiC	hypothetical protein (methyltransferase domain)	1	1	2.47**	
yeeF	hypothetical protein (acetyltransferase domain)	1	1	2.49**	pyrC, aldB, glmU
ywdF	PDZ domain-containing protein	1	1	2.51**	pyrC,glmU
ybiE	oxidoreductase	1	1	2.53**	
yieH	hypothetical protein	1	1	2.54**	
gyrB	DNA gyrase subunit B	1	1	2.54**	pyrC, glmU, glK
pyrP	uracil permease	1		2.57**	
ygjB	hypothetical protein (stress response domain)	1	1	2.58**	
ftsK	cell division protein FtsK	1	1	2.59**	pyrC,glmU
kinC	sensor protein kinase	1	1	2.60**	pyrC,glmU
rgrB	GntR family transcriptional regulator	1	1	2.62**	glyA, pyrC, glmU

yjbC	hypothetical protein (hydrolase HAD domain)	1	1	2.62**	
yngG	sugar ABC transporter permease protein	1	1	2.63**	
sunL	rRNA methylase	1	1	2.64**	
htrA	exported serine protease	1	1	2.64**	
yudJ	hypothetical protein (ABC transporter permease domain)	1	1	2.64**	
comFA	competence protein ComFA	1	1	2.64**	pyrC,glmU
glmU	glucosamine-1-phosphate N-acetyltransferase	1	1	2.03**	
glyA	serine, glycine hydroxymethyltransferase	1	1	2.13**	
yjhF	Phosphoglycerate mutase	1	1	2.26**	
pyrC	dihydroorotase	1	1	2.37**	aldB, pyk
argS	Argynil tRNA synthetase	1		2.48**	
hasC	UTP-glucose-1-Phosphate uridylyltransferase	1	1	2.50**	
atpD	ATP synthetase	1	1	2.54**	
murE	UDP-MurNAc-tripeptide synthetase	1	1	2.57**	murB, rpiA, deoB
murB	UDP-N-Acetylenolpyruvoylglucosamine reductase	1	1	2.57**	murE, rpiA, deoB
rpiA	Ribose-5-Phosphate isomerase	1	1	2.58**	murB, murE, glyA, pyrC
holB	DNA polymerase	1	1	2.61**	
aldB	alpha-acetolactate decarboxylase	1	1	2.62**	pyrC, yjhF, pyk, rpiA
pyk	pyruvate kinase	1	1	2.63**	pyrC, yjhF,
deoB	phosphopentomutase	1	1	2.65**	murB, murE
malE	Maltose ABC transporter		1	2.28 +	
dfpB	Hypotetical protein		1	2.34 +	
arsC	Arsenate reductase		1	2.62 +	
ftsQ	Cell division protein		1	2.62 +	

**distances given for the results with K2K3K4.
+ distances obtained with K2K3K4K5

*Shared predicted relationships on the known protein metabolic network (data used for training) strengthen the relationships found between two genes belonging to the candidate data set. Taking this observation into account can help to avoid making mistakes in the case where small distances calculated between genes could have been obtained simply due to the two genes having small similarities to all genes rather than

reflecting any real small distance between them.

Cross-validation

combination	features	dRBFK3	dRBFK4	dRBF2K5	delta	diffusion constant
K2	50	/	/	/	0.001	0.1
K3	50	0.0005	/	/	0.001	0.1
K4	50	/	0.011	/	0.001	0.1
K2K3	50	0.0005	/	/	0.001	0.06
K3K4	50	0.0005	0.001	/	0.001	0.05
K2K4	50	/	0.001	/	0.001	0.01
K2K3K4	50	0.0005	0.001	/	0.001	0.01
K5	50	/	/	0.55	0.001	0.01
K4K5	50	/	0.001	0.55	0.001	0.01
K2K3K4K5	50	0.0005	0.001	0.55	0.001	0.01
tested values	10:100 by 10	0.0005:0.005 by 0.0505	0.001:0.1 by 0.0101	0.05:0.55 by 0.25	0.001:0.01 by 0.001	0.01:0.1 by 0.01

13.1.2 YvjB ou Eep

position	distance	gène	fonction de la protéine
1	1.1438	ffh	signal recognition particle protein Ffh
2	1.1984	yhhG	hypothetical protein
3	1.2474	yuaA	putative CMP-binding factor
4	1.2947	ylxQ	hypothetical protein
5	1.3231	prfA	peptide chain release factor RF-1
6	1.3584	yccI	hypothetical protein
7	1.3663	gidA	glucose inhibited division protein GidA
8	1.3778	nusG	transcription antitermination protein
9	1.4027	yhhC	hypothetical protein
10	1.5610	rplE	50S ribosomal protein L5
11	1.6201	glmS	glucosamine--fructose-6-phosphate aminotransferase
12	1.6762	fusA	elongation factor EF-G
13	1.6968	rheB	ATP-dependent RNA helicase
14	1.6970	ftsY	cell division protein FtsY
15	1.7141	prmA	ribosomal protein L11 methyltransferase
16	1.7232	rnc	ribonuclease III
17	1.7236	rbfA	ribosome-binding factor A
18	1.7553	rheA	ATP-dependent RNA helicase
19	1.7758	rplI	50S ribosomal protein L9
20	1.7865	yueF	putative protease
21	1.8190	rplL	50S ribosomal protein L7/L12
22	1.8576	ftsH	cell division protein FtsH
23	1.8801	gidB	glucose-inhibited division protein GidB
24	1.8889	sunL	rRNA methylase; K03500 Sun protein
25	1.9179	cdsa	phosphatidate cytidylyltransferase; Glycerophospholipid metabolism
26	1.9419	ecsA	ABC transporter ATP binding protein
27	1.9446	rnhB	ribonuclease HII
28	1.9488	yedA	hypothetical protein
29	1.9608	lspA	lipoprotein signal peptidase
30	1.9687	yljI	permease
31	1.9718	yfjD	tRNA/rRNA methyltransferase
32	1.9956	rplJ	50S ribosomal protein L10
33	1.9960	uppS	undecaprenyl pyrophosphate synthetase
34	2.0012	rimM	16S rRNA processing protein
35	2.0062	yjaI	unknown protein
36	2.0211	fhuD	ferrichrome ABC transporter substrate binding protein
37	2.0320	ftsQ	cell division protein FtsQ
38	2.0541	secY	preprotein translocase SecY subunit
39	2.0691	ysfB	ABC transporter ATP-binding protein
40	2.0711	rplM	50S ribosomal protein L13
41	2.0983	rgrB	GntR family transcriptional regulator
42	2.1076	recJ	single-stranded DNA specific exonuclease
43	2.1347	ecsB	ABC transporter permease protein
44	2.1382	yhbE	unknown protein
45	2.1521	obgL	GTP-binding protein Obg
46	2.1554	infC	translation initiation factor IF-3
47	2.1581	clpC	ATP-dependent protease ATP-binding subunit
48	2.1645	priA	primosomal protein N'(replication factor Y) (superfamily II helicase)

49	2.1684	prfB	peptide chain release factor RF-2
50	2.1739	yjaD	transcription regulator
51	2.1808	acpS	holo-[acyl-carrier protein] synthase
52	2.1942	yccE	unknown protein
53	2.1982	rpsJ	30S ribosomal protein S10
54	2.2000	yfdB	hypothetical protein
55	2.2071	yacB	hypothetical protein
56	2.2327	lgt	prolipoprotein diacylglycerol transferase
57	2.2361	pheS	phenylalanyl-tRNA synthetase alpha chain
58	2.2916	tig	trigger factor
59	2.3031	apt	adenine phosphoribosyltransferase; Purine metabolism
60	2.3040	rplB	50S ribosomal protein L2;
61	2.3052	yccL	hypothetical protein
62	2.3097	gatB	Glu-tRNA amidotransferase subunit B
63	2.3139	dnaA	replication initiation protein DnaA
64	2.3337	clpX	ATP dependent Clp protease
65	2.3410	xpt	xanthine phosphoribosyltransferase
66	2.3439	ywdG	hypothetical protein
67	2.3593	yebE	hypothetical protein
68	2.3626	ydgI	hypothetical protein
69	2.3830	ptpL	protein-tyrosine phosphatase
70	2.4024	yhfD	hypothetical protein
71	2.4035	yqdA	hypothetical protein
72	2.4313	ygcC	oxidoreductase
73	2.4409	yofM	hypothetical protein
74	2.4453	ksgA	dimethyladenosine transferase
75	2.4502	ybaF	hypothetical protein
76	2.4577	lysS	lysyl-tRNA synthetase
77	2.4626	ezrA	cell division regulator
78	2.5570	smpB	tmRNA-binding protein SmpB

13.1.3 Prénilpeptidase CAAX

position	distance	gène	fonction de la protéine
1	1.0449	yoic	unknown, pas de domaines connus
2	1.148	pi345	prophage
3	1.282	ps301	prophage
4	1.3589	yhge	unknown
5	1.3728	ps351	prophage
6	1.4172	cbr	carbonyl reductase
7	1.4624	ykbc	hypotetical
8	1.4991	yveb	unknown
9	1.5044	pi108	prophage
10	1.5097	ysca	unknown
11	1.5427	yheg	hypotetical
12	1.5498	ywdd	unknown
13	1.5539	yphk	unknown protein
14	1.5709	yoia	hypotetical
15	1.5785	ymdc	kanamycin kinase
16	1.603	pi251	prophage

17	1.612	yfic	unknown
18	1.6125	ycdf	transcription regulator
19	1.62	yree	unknown
20	1.629	ytjF	unknown
21	1.6297	yqbk	unknown
22	1.6329	yhgd	unknown
23	1.6333	yted	transmembrane efflux protein
24	1.6395	yxde	oxidoreductase
25	1.64	yvaD	unknown
26	1.6435	yldB	hypotetical
27	1.6509	rmaH	transcription regulator
28	1.652	yphc	oxidoreductase
29	1.6645	pi144	prophage
30	1.6757	rmeC	transcription regulator
31	1.6761	yihD	unknown
32	1.6813	yqfe	hypotetical
33	1.6895	ycgc	hypotetical
34	1.6931	ycfH	hypotetical
35	1.6985	yogI	hypotetical
36	1.7018	ytgD	unknown
37	1.7023	ywfh	unknown
38	1.715	yqia	multidrug transporter
39	1.7241	pi357	prophage
40	1.7305	pi349	prophage
41	1.7365	yudF	unknown
42	1.7439	yxba	hypotetical
43	1.744	yfbi	unknown
44	1.7486	ypac	hypotetical
45	1.7553	yrbb	hypotetical
46	1.7581	yhcb	unknown
47	1.7644	yfcc	hypotetical
48	1.7685	yxdd	transcription regulator
49	1.7783	yfjG	transcription regulator
50	1.7797	ps215	prophage
51	1.7811	ymgc	hypotetical
52	1.7813	yccb	unknown
53	1.7942	ypgb	oxidoreductase
54	1.7991	pi202	prophage
55	1.8147	yjjd	ABC transporter permease protein;
56	1.8197	yejD	unknown
57	1.8224	yndd	unknown
58	1.8226	yndc	unknown
59	1.8229	tra905	transposase of IS905
60	1.8257	yijh	unknown
61	1.8295	yfgh	hypotetical
62	1.8326	ygig	hypotetical
63	1.8361	yoaH	hypothetical protein
64	1.8364	yhcg	hypotetical
65	1.8398	yoib	unknown
66	1.8401	ypcc	unknown
67	1.8434	pi322	prophage

68	1.8469	ylcf	unknown
69	1.8535	yhbf	hypotetical
70	1.8592	pi327	prophage
71	1.8616	pi201	prophage
72	1.862	ybdk	hypotetical
73	1.8632	ycgI	hypotetical
74	1.8679	yseh	unknown
75	1.8685	rara	transcription regulator
76	1.8707	ynii	hypotetical
77	1.8707	yqeh	unknown
78	1.8736	ykhk	unknown
79	1.8749	ybdJ	hypotetical
80	1.878	yica	unknown
81	1.8789	yneg	unknown
82	1.8825	yjgb	hypotetical
83	1.8836	yliB	hypotetical
84	1.8884	tra981	transposase of IS981
85	1.8885	ynef	unknown
86	1.8891	pi229	prophage
87	1.89	pi347	prophage
88	1.8913	ygih	hypotetical
89	1.8967	ymjE	glycosyl transferase
90	1.8991	ybeH	hypotetical

13.1.4 CodY

position	distances	gène	fontion de la protéine
1	1.21	guaB	inositol-5-monophosphate dehydrogenase
2	1.28	ctsR	CtsR
3	1.56	yugD	protease
4	1.76	rplC	50S ribosomal protein L3
5	1.80	hprT	hypoxanthine-guanine phosphoribosyltransferase
6	1.80	ybiD	hypothetical protein L184159
7	1.85	yjjA	hypothetical protein L190464
8	1.89	yriA	hypothetical protein L165684
9	1.92	ykjA	hypothetical protein L90693
10	1.94	rplF	50S ribosomal protein L6
11	2.06	rplB	50S ribosomal protein L2
12	2.08	yheA	hypothetical protein L140288
13	2.12	greA	transcription elongation factor
14	2.19	fabZ1	hydroxymyristoyl-acyl carrier protein dehydratase
15	2.20	ytgF	hypothetical protein L150593
16	2.27	ahrC	transcription regulator
17	2.28	yccF	hypothetical protein L26054
18	2.35	rplK	50S ribosomal protein L11
19	2.41	yeiF	hypothetical protein L86338
20	2.43	ybjK	hypothetical protein L195257
21	2.48	nifU	NifU

22	2.52	pyrH	UMP-kinase
23	2.53	dnaA	chromosomal replication initiation protein
24	2.54	yahG	ABC transporter ATP binding protein
25	2.55	rplN	50S ribosomal protein L14
26	2.56	yhhG	hypothetical protein L175136
27	2.60	clpX	ATP-dependent protease ATP-binding subunit
28	2.62	hpt	hypoxantine-guanine phosphorybosyltransferase
29	2.63	accC	acetyl-CoA carboxylase biotin carboxylase subunit
30	2.66	rplE	50S ribosomal protein L5
31	2.70	ygdA	hypothetical protein L32389
32	2.71	dfpA	phosphopantothenoylcysteine decarboxylase
33	2.73	ybeB	hypothetical protein L141748
34	2.76	upp	uracil phosphoribosyltransferase
35	2.77	tuf	elongation factor Tu
36	2.78	tig	trigger factor
37	2.79	tgt	queuine tRNA-ribosyltransferase
38	2.79	ytfB	hypothetical protein L133761
39	2.81	yacB	hypothetical protein L1001
40	2.82	lplL	lipoate-protein ligase
41	2.82	zitP	zinc ABC transporter permease protein
42	2.88	llrD	two-component system regulator
43	2.91	yxfC	hypothetical protein L142355
44	2.94	hflX	HflX
45	2.95	rplA	50S ribosomal protein L1
46	2.95	yuhB	protease
47	2.96	yofM	hypothetical protein L57401
48	2.98	fusA	elongation factor G
49	2.98	yueE	putative protease
50	3.02	zitQ	zinc ABC transporter ATP binding protein
51	3.03	dnaK	molecular chaperone DnaK
52	3.06	frr	ribosome recycling factor
53	3.07	yjjG	hypothetical protein L196206
54	3.10	nusG	transcription antitermination protein NusG
55	3.13	uvrC	excinuclease ABC subunit C
56	3.13	glnR	glutamine synthetase repressor
57	3.15	yhfD	hypothetical protein L155396
58	3.15	infC	translation initiation factor IF-3
59	3.15	ysxL	GTPase EngB
60	3.17	rbgA	ribosomal biogenesis GTPase
61	3.19	llrC	two-component system regulator
62	3.19	fur	ferric uptake regulator
63	3.20	cysK	cysteine synthase
64	3.20	ynaE	hypothetical protein L106755
65	3.23	ygjB	hypothetical protein L92295

13.1.5 PepN

position	distances	gène	fonction de la protéine
1	1.05	yeeG	transcription regulator
2	1.52	htrA	exported serine protease

3	1.69	ygbD	hypothetical protein L15964
4	1.82	gyrA	DNA gyrase subunit A
5	1.96	yshB	hypothetical protein L61866
6	2.03	rmlB	dTDP-glucose 4,6-dehydratase
7	2.05	pmpA	maturation protein
8	2.18	pepDB	dipeptidase
9	2.21	fhuR	fhu operon transcription regulator
10	2.22	trpS	tryptophanyl-tRNA synthetase II
11	2.23	yieH	hypothetical protein L48341
12	2.26	yacG	hypothetical protein L22496
13	2.26	yhbF	hypothetical protein L115789
14	2.27	pydA	dihydroorotate dehydrogenase 1A
15	2.32	pepC	aminopeptidase C
16	2.53	yedE	hypothetical protein L37338
17	2.55	parC	DNA topoisomerase IV subunit A
18	2.62	dfrA	dihydropholate reductase
19	2.63	yacC	hypothetical protein L21634
20	2.64	ytcE	hypothetical protein L112263
21	2.69	pbp2A	penicillin-binding protein 2a
22	2.72	rgpA	rhamnosyltransferase
23	2.74	yrjG	hypothetical protein L179243
24	2.77	ylfF	hypothetical protein L155662
25	2.85	ybgB	hypothetical protein L160937
26	2.86	ybaA	hypothetical protein L101209
27	2.88	noxE	NADH oxidase
28	2.88	yxdB	hypothetical protein L117205
29	2.89	pbp1B	penicillin-binding protein 1B
30	2.92	xylH	4-oxalocrotonate tautomerase
31	2.93	atpH	ATP synthase delta subunit
32	2.96	groEL	chaperonin GroEL
33	2.96	ps123	integrase
34	2.99	ygbE	hypothetical protein L16806
35	3.00	recA	recombinase A
36	3.01	ymgI	hypothetical protein L66407
37	3.04	pbpX	penicillin-binding protein
38	3.06	ybjB	hypothetical protein L196216
39	3.07	ysiA	transport protein
40	3.07	rmlC	dTDP-L-rhamnose synthase
41	3.10	yliE	hypothetical protein L184675
42	3.12	yeiE	hypothetical protein L85854
43	3.14	ygaI	hypothetical protein L6768
44	3.14	xerS	site-specific tyrosine recombinase XerS
45	3.19	comGD	ComGD
46	3.19	pepDA	dipeptidase
47	3.19	yviI	hypothetical protein L169106
48	3.21	ytgA	hypothetical protein L143459
49	3.24	yliF	hypothetical protein L184880

13.1.6 AtpA

position	distance	gène	fonction de la protéine
1	1.37	dnaE.	DNA polymerase III DnaE
2	1.40	secA.	preprotein translocase subunit SecA
3	1.41	atpD	ATP synthase alpha subunit
4	1.58	pheT.	phenylalanyl-tRNA synthetase subunit beta
5	1.63	glmS	glucosamine--fructose-6-phosphate aminotransferase
6	1.69	nifS.	pyridoxal-phosphate dependent aminotransferase
7	1.76	ycjB.	hypothetical protein L91807
8	1.77	dnaN.	DNA polymerase III subunit beta
9	1.78	priA.	primosome assembly protein PriA
10	1.79	dnaD.	DnaD
11	1.81	glmU	glucosamine-1-phosphate N-acetyltransferase
12	1.83	atpE	ATP synthase epsilon subunit
13	1.87	yccG.	hypothetical protein L26400
14	1.97	yheB.	hypothetical protein L141547
15	2.00	dnaJ.	DnaJ
16	2.05	nifZ.	pyridoxal-phosphate dependent aminotransferase
17	2.07	atpF	ATP synthase subunit b
18	2.08	recD.	exodeoxyribonuclease V alpha chain
19	2.09	prfB.	peptide chain release factor 2
20	2.13	mutS or hexA.	MutS
21	2.17	yrgF.	hypothetical protein L148945
22	2.26	ykiC.	hypothetical protein L84257
23	2.27	rnpA.	ribonuclease P
24	2.28	rnz or ygcA.	ribonuclease Z
25	2.33	trmE or thdF.	tRNA modification GTPase TrmE
26	2.34	ftsK.	FtsK
27	2.35	clpC.	ATP-dependent protease ATP-binding subunit
28	2.36	ruvA.	Holliday junction DNA helicase motor protein
29	2.40	ruvB.	Holliday junction DNA helicase B
30	2.43	yqeL.	GTP-binding protein
31	2.43	ygiK.	hypothetical protein L87336
32	2.45	yujA.	hypothetical protein L74738
33	2.46	trxB1.	thioredoxin reductase
34	2.47	yjbE.	general stress protein GSP13
35	2.51	yudL.	hypothetical protein L22691
36	2.51	yfjE.	flavodoxin
37	2.51	ykaC.	hypothetical protein L5517
38	2.52	ysiG.	hypothetical protein L1889726
39	2.52	lacR.	lactose transport regulator
40	2.53	yuhI.	hypothetical protein L60959
41	2.54	ftsY.	FtsY
42	2.55	uvrB.	excinuclease ABC subunit B
43	2.56	zitQ.	zinc ABC transporter ATP binding protein
44	2.57	rheB.	ATP-dependent RNA helicase
45	2.58	yrjA.	hypothetical protein L173313
46	2.58	yqjB.	hypothetical protein L93855
47	2.59	radA.	DNA repair protein RadA
48	2.60	dacB.	D-alanyl-D-alanine carboxypeptidase
49	2.61	xpt.	exodeoxyribonuclease VII large subunit
50	2.64	ylfH.	N-acetylglucosamine catabolic protein

51	2.64	ftsQ.	FtsQ
52	2.64	rimM.	16S rRNA-processing protein
53	2.66	gcp.	O-sialoglycoprotein endopeptidase
54	2.66	def.	peptide deformylase
55	2.68	ywgA or recX.	recombination regulator RecX
56	2.69	yciD.	hypothetical protein L86677
57	2.71	yseF.	hypothetical protein L29491
58	2.71	ylxQ.	hypothetical protein L175450
59	2.72	polC.	DNA polymerase III PolC
60	2.76	yqdA.	hypothetical protein L39365
61	2.78	aroC.	chorismate synthase
62	2.79	cysM.	cysteine synthase
63	2.79	ylgC.	hypothetical protein L161988
64	2.81	yraE.	hypothetical protein L106425
65	2.81	ftsZ.	cell division protein FtsZ
66	2.82	engC or yuaD.	ribosome-associated GTPase
67	2.82	ptpL.	tyrosine phosphatase
68	2.82	lspA.	lipoprotein signal peptidase
69	2.83	yjaD.	transcription regulator
70	2.83	yqaC.	hypothetical protein L4747
71	2.84	hslO or yudG.	Hsp33-like chaperonin
72	2.85	smc.	chromosome segregation SMC protein
73	2.86	yraC.	DNA polymerase III subunit delta
74	2.87	rgrB.	GntR family transcription regulator
75	2.87	yogG or cfa.	hypothetical protein L65498
76	2.88	ykiG.	hypothetical protein L87113
77	2.88	prfA.	peptide chain release factor 1
78	2.88	yudK.	hypothetical protein L21717
79	2.89	purM.	phosphoribosylaminoimidazole synthetase
80	2.89	yfdE.	hypothetical protein L133367
81	2.89	ybbE.	hypothetical protein L114325
82	2.91	comGA.	ComGA
83	2.91	yqjE.	hypothetical protein L100263
84	2.91	mutY.	A/G-specific adenine glycosylase
85	2.92	sunL.	rRNA methylase
86	2.94	yofM.	hypothetical protein L57401
87	2.95	ywdF.	hypothetical protein L20937
88	2.96	yuaA.	hypothetical protein L184033
89	2.97	trxH.	thioredoxin H-type
90	2.99	pcrA.	ATP-dependent helicase PcrA
91	2.99	ybiE.	oxidoreductase
92	2.99	kinC.	sensor protein kinase
93	3.01	rplI.	50S ribosomal protein L9
94	3.01	ecsB.	ABC transporter permease protein
95	3.01	ysjC.	hypothetical protein L76216
96	3.02	ygjD.	4-alpha-glucanotransferase
97	3.05	yfcI.	hypothetical protein L128550
98	3.06	smpB.	SsrA-binding protein
99	3.06	gltX.	glutamyl-tRNA synthetase
100	3.07	ffh.	signal recognition particle protein
101	3.08	yeaA.	hypothetical protein L582
102	3.09	pknB.	serine/threonine protein kinase

103	3.10	ylcC.	hypothetical protein L125196
104	3.11	kinD.	sensor protein kinase
105	3.13	cshA.	recombination factor protein RarA
106	3.13	pyrP.	pyrimidine regulatory protein PyrR
107	3.14	yhhE.	hypothetical protein L173151
108	3.14	ywbA.	hypothetical protein L193121
109	3.14	ydjD.	hypothetical protein L195751
110	3.15	dltD.	D-alanine transfer protein
111	3.16	dltB.	peptidoglycan biosynthesis protein
112	3.17	obgL.	GTPase ObgE
113	3.17	comFA.	ComFA
114	3.19	yqgA.	hypothetical protein L61727
115	3.19	nifU.	NifU
116	3.20	infB.	translation initiation factor IF-2
117	3.20	ytbE.	hypothetical protein L101219
118	3.20	yecE.	putative lipid kinase
119	3.21	yciH.	hypothetical protein L89418
120	3.21	yejH.	hypothetical protein L98583
121	3.22	nagB.	glucosamine-6-P isomerase
122	3.22	yxfB.	hypothetical protein L141634
123	3.23	trxB2.	thioredoxin reductase
124	3.23	yigC.	hypothetical protein L862989
125	3.23	uvrC.	excinuclease ABC subunit C
126	3.24	ytfB.	hypothetical protein L133761
127	3.25	yfdB.	hypothetical protein L131937

13.1.7 AtpD

position	distance	gène	fonction de la protéine
1	0.95	ftsA.	FtsA
2	1.02	ftsW.	FtsW1
3	1.14	sigX	RNA polymerase ECF sigma factor
4	1.25	pknB.	preprotein translocase subunit SecA
5	1.32	secA.	hypothetical protein L26400
6	1.41	atpA	ATPase
7	1.60	atpE	ATP synthase epsilon subunit
8	1.64	nifS.	pyridoxal-phosphate dependent aminotransferase
9	1.74	gyrB.	DNA gyrase subunit B
10	1.75	aptG	ATP synthase gamma subunit
11	1.96	nifZ.	pyridoxal-phosphate dependent aminotransferase
12	1.99	gcp.	O-sialoglycoprotein endopeptidase
13	2.00	yjiF.	hypothetical protein L189428
14	2.05	glmU.	glucosamine-1-phosphate N-acetyltransferase
15	2.19	ybdD.	hypothetical protein L132712
16	2.20	ccpA.	catabolite control protein A
17	2.21	mutY.	A/G-specific adenine glycosylase
18	2.22	ptsI.	phosphoenolpyruvate-protein phosphotransferase
19	2.24	pyrC.	dihydroorotase
20	2.24	mfd.	transcription-repair coupling factor
21	2.24	ytdC.	hypothetical protein L118668

22	2.25	aspC.	aspartate aminotransferase
23	2.26	leuS.	leucyl-tRNA synthetase
24	2.26	uvrB.	excinuclease ABC subunit B
25	2.26	pyrP.	pyrimidine regulatory protein PyrR
26	2.33	comFA.	ComFA
27	2.34	pepM.	methionine aminopeptidase
28	2.44	yheB.	hypothetical protein L141547
29	2.44	purM.	phosphoribosylaminoimidazole synthetase
30	2.45	yjbC.	hypothetical protein L112952
31	2.46	yudJ.	hypothetical protein L20683
32	2.47	mutS or hexA.	MutS
33	2.48	kinC.	sensor protein kinase
34	2.54	pepF	oligoendopeptidase F
35	2.58	yeeE.	hypothetical protein L44542
36	2.60	priA.	primosome assembly protein PriA
37	2.60	pheT.	phenylalanyl-tRNA synthetase subunit beta
38	2.60	murI.	glutamate racemase
39	2.63	yhfB.	hypothetical protein L151062
40	2.69	yudL.	hypothetical protein L22691
41	2.71	aldB.	alanyl-tRNA synthetase
42	2.77	atpB.	F0F1 ATP synthase subunit alpha
43	2.80	rgrB.	GntR family transcription regulator
44	2.87	proS.	prolyl-tRNA synthetase
45	2.87	pstC.	phosphate ABC transporter permease protein
46	2.88	mtsB.	manganese ABC transporter ATP binding protein
47	2.89	clpE.	ATP-dependent protease ATP-binding subunit
48	2.93	miaA.	tRNA delta(2)-isopentenylpyrophosphate transferase
49	2.97	ftsK.	FtsK
50	2.98	ftsQ.	FtsQ
51	2.99	yvaB.	hypothetical protein L85091
52	3.01	ywdF.	hypothetical protein L20937
53	3.01	ygiK.	hypothetical protein L87336
54	3.03	yqgE.	transporter
55	3.07	ftsE.	cell-division ATP-binding protein
56	3.07	pg.	glucose-6-phosphate isomerase
57	3.11	dnaJ.	DnaJ
58	3.11	yfjE.	flavodoxin
59	3.15	serS.	seryl-tRNA synthetase
60	3.16	holB.	DNA polymerase III subunit delta'
61	3.16	ygcC.	oxidoreductase
62	3.21	efp.	elongation factor P
63	3.23	aroA.	3-phosphoshikimate 1-carboxyvinyltransferase
64	3.24	yudK.	hypothetical protein L21717
65	3.24	ywdG.	hypothetical protein L22498

13.1.8 AtpG

position	distance	gène	fonction de la protéine
1	1.60	sigX	RNA polymerase ECF sigma factor
2	1.75	atpD	ATP synthase alpha subunit

3	1.87	ponA.	penicillin-binding protein 1A
4	1.92	pbpX.	penicillin-binding protein
5	2.04	rgpA.	rhamnosyltransferase
6	2.07	atpH	ATP synthase delta subunit
7	2.14	yshB.	hypothetical protein L61866
8	2.19	pepDB.	dipeptidase
9	2.30	glmU	glucosamine-1-phosphate N-acetyltransferase
10	2.34	ygbD.	hypothetical protein L15964
11	2.41	glmS	glucosamine--fructose-6-phosphate aminotransferase
12	2.41	pyk.	pyruvate kinase
13	2.44	yacG.	hypothetical protein L22496
14	2.45	ytgA.	hypothetical protein L143459
15	2.52	ykiH.	hypothetical protein L88637
16	2.54	busAB.	betaine ABC transporter permease and substrate binding protein
17	2.60	ymhA.	hypothetical protein L73572
18	2.65	rmlC.	dTDP-L-rhamnose synthase
19	2.67	ytcB.	hypothetical protein L104285
20	2.72	parC.	DNA topoisomerase IV subunit A
21	2.75	rmlA.	glucose-1-phosphate thymidylyltransferase
22	2.76	yhcA.	ABC transporter ATP-binding and permease protein
23	2.77	rmlB.	dTDP-glucose 4,6-dehydratase
24	2.79	noxD.	NADH oxidase
25	2.82	ynaB.	transcription regulator
26	3.02	yxdB.	hypothetical protein L117205
27	3.11	pbp1B.	penicillin-binding protein 1B
28	3.14	ygbE.	hypothetical protein L16806
29	3.17	ynbA.	
30	3.21	gyrA.	DNA gyrase subunit A
31	3.21	yqjD.	hypothetical protein L98109
32	3.21	yliF.	hypothetical protein L184880
33	3.24	yqaD.	hypothetical protein L5610
34	3.25	yqfG.	hypothetical protein L58460

13.1.9 AtpH

position	distance	gène	fonction de la protéine
1	1.67	yieH.	exported serine protease
2	1.73	pepC.	hypothetical protein L115789
3	1.91	ywaG.	glutamine ABC transporter permease and substrate binding protein
4	2.06	yufA.	branched-chain amino acid aminotransferase
5	2.07	atpG	ATP synthase gamma subunit
6	2.10	hasC.	outer membrane lipoprotein precursor
7	2.11	htrA.	tryptophanyl-tRNA synthetase II
8	2.13	yhcA.	ABC transporter ATP-binding and permease protein
9	2.15	glnP.	hypothetical protein L148513
10	2.24	bcaT.	transport protein
11	2.24	pbp1B.	hypothetical protein L183216
12	2.39	plpC.	hypothetical protein L110441
13	2.40	trpS.	hypothetical protein L6768
14	2.43	ywaF.	ATP-dependent protease ATP-binding subunit

15	2.44	yrgE.	fructose-bisphosphate aldolase
16	2.45	ysiA.	hypothetical protein L196216
17	2.48	ybiC.	exonuclease VII small subunit
18	2.54	ynbB.	transcription regulator
19	2.56	ygaI.	glyceraldehyde 3-phosphate dehydrogenase
20	2.57	clpE.	multidrug resistance efflux pump
21	2.58	fbaA.	dTDP-L-rhamnose synthase
22	2.59	ybjB.	hypothetical protein L110588
23	2.60	xseA.	hypothetical protein L179243
24	2.61	yeeG.	hypothetical protein L112263
25	2.73	gap.	hypothetical protein L96658
26	2.78	pmrA.	oxidoreductase
27	2.82	rmlC.	NADH oxidase
28	2.85	yjbB.	amino acid ABC transporter permease protein
29	2.85	yrjG.	aminopeptidase N
30	2.86	ytcE.	maturation protein
31	2.87	usp45.	glycosil transferase
32	2.88	ybiE.	ABC transporter permease protein
33	2.90	noxD.	hypothetical protein L102093
34	2.91	yjgD.	xanthine permease
35	2.93	pnpA.	dipeptidase
36	2.96	ybaI.	fhu operon transcription regulator
37	2.98	ylbB.	outer membrane lipoprotein precursor
38	3.00	yjaB.	hypothetical protein L5610
39	3.00	pbuX.	zinc ABC transporter substrate binding protein
40	3.03	nagA.	penicillin-binding protein 2a
41	3.03	pepDB.	hypothetical protein L131806
42	3.04	fhuR.	regulatory protein
43	3.05	plpB.	hypothetical protein L160937
44	3.09	yqaD.	hypothetical protein L56431
45	3.09	zitS.	heat-inducible transcription repressor
46	3.09	pbp2A.	hypothetical protein L22496
47	3.13	yveI.	hypothetical protein L149781
48	3.15	mecA or ysfD.	amino acid ABC transporter substrate binding protein
49	3.16	ybgB.	hypothetical protein L155662
50	3.16	yuhE.	hypothetical protein L114054
51	3.19	hrcA.	hrcA
52	3.19	yacG.	yacG
53	3.20	yrgG.	yrgG
54	3.22	yvdF.	yvdF
55	3.23	ylfF.	ylfF
56	3.23	yvdA.	yvdA

13.1.10 DnaA

position	distance	gène	fonction de la protéine
1	0.54	hpt.	hypoxantine-guanine phosphorybosyltransferase
2	1.00	yacB.	hypothetical protein L1001
3	1.14	ybjK.	hypothetical protein L195257
4	1.25	ybcG.	hypothetical protein L122849

5	1.43	rplC.	50S ribosomal protein L3
6	1.43	sdaA.	alpha-subuni L-serine dehydratase
7	1.47	rplE.	50S ribosomal protein L5
8	1.47	yccF.	hypothetical protein L26054
9	1.47	yjjA.	hypothetical protein L190464
10	1.49	yxaC.	hypothetical protein L86471
11	1.52	yyaL.	translation-associated GTPase
12	1.56	ybeB.	hypothetical protein L141748
13	1.57	ygdA.	hypothetical protein L32389
14	1.58	fusA.	elongation factor G
15	1.59	dnaC.	replicative DNA helicase
16	1.60	purA.	adenylosuccinate synthetase
17	1.65	ykhF.	ABC transporter ATP binding protein
18	1.69	ylxQ.	hypothetical protein L175450
19	1.73	ytgF.	hypothetical protein L150593
20	1.73	yhhG.	hypothetical protein L175136
21	1.74	radA.	DNA repair protein RadA
22	1.75	rplB.	50S ribosomal protein L2
23	1.76	femD.	phosphoglucosamine mutase
24	1.78	ysxL or engB.	GTPase EngB
25	1.78	nusG.	transcription antitermination protein NusG
26	1.82	hprK or ptsK or hprT.	hypoxanthine-guanine phosphoribosyltransferase
27	1.82	pth.	peptidyl-tRNA hydrolase
28	1.84	nifU.	NifU
29	1.84	nusA.	transcription elongation factor NusA
30	1.87	pyrH.	UMP-kinase
31	1.87	yjaF.	hypothetical protein L106356
32	1.94	mraY.	phospho-N-acetylmuramoyl-pentapeptide-transferase
33	1.94	ykjI.	hypothetical protein L98095
34	1.96	tig.	trigger factor
35	1.99	rplK.	50S ribosomal protein L11
36	2.02	folC.	folylpolyglutamate synthase
37	2.04	yofM.	hypothetical protein L57401
38	2.05	comFC.	ComFC
39	2.07	pg.	glucose-6-phosphate isomerase
40	2.08	efp.	elongation factor P
41	2.11	rplN.	50S ribosomal protein L14
42	2.14	yhbE.	hypothetical protein L92295
43	2.17	greA.	transcription elongation factor
44	2.18	yhfF.	hypothetical protein L157023
45	2.20	pyrG.	CTP synthetase
46	2.20	udk.	uridine kinase
47	2.22	rplJ.	50S ribosomal protein L10
48	2.22	yfdB.	hypothetical protein L131937
49	2.23	ybeC.	hypothetical protein L142332
50	2.23	frr.	ribosome recycling factor
51	2.24	ylgC.	hypothetical protein L161988
52	2.25	yeaA.	hypothetical protein L582
53	2.25	lepA.	GTP-binding protein LepA
54	2.26	ylqL.	GTP-binding protein
55	2.27	ctsR.	CtsR
56	2.28	yvaB.	hypothetical protein L85091

57	2.29	yueE.	putative protease
58	2.31	gidB.	glucose-inhibited division protein B
59	2.31	alaS.	transcription regulator
60	2.33	rplA.	50S ribosomal protein L1
61	2.34	ftsH or tma.	FtsH
62	2.34	smpB.	SsrA-binding protein
63	2.35	rplM.	50S ribosomal protein L13
64	2.36	rplF.	50S ribosomal protein L6
65	2.36	lysS.	lysyl-tRNA synthetase
66	2.36	rluB.	pseudouridine synthase
67	2.36	tuf.	elongation factor Tu
68	2.36	yjjG.	hypothetical protein L196206
69	2.38	hslO or yudG.	Hsp33-like chaperonin
70	2.38	yccH.	hypothetical protein L26998
71	2.40	dfpA.	phosphopantothenoylcysteine decarboxylase
72	2.42	ydgF.	hypothetical protein L164461
73	2.44	yseI.	hypothetical protein L32195
74	2.45	yuaA.	hypothetical protein L184033
75	2.46	yjiE.	hypothetical protein L188550
76	2.46	ykiG.	hypothetical protein L87113
77	2.48	yheA.	hypothetical protein L140288
78	2.49	purB.	adenylosuccinate lyase
79	2.50	ysiG.	hypothetical protein L1889726
80	2.50	rluD.	pseudouridine synthase
81	2.52	dnaN.	DNA polymerase III subunit beta
82	2.53	yccK.	hypothetical protein L28696
83	2.53	codY.	transcriptional repressor CodY
84	2.54	trmH.	tRNA-guanosine methyltransferase
85	2.54	llrD.	two-component system regulator
86	2.55	yqaC.	hypothetical protein L4747
87	2.58	gmk.	guanylate kinase
88	2.58	prfA.	peptide chain release factor 1
89	2.58	clpX.	ATP-dependent protease ATP-binding subunit
90	2.58	fabZ1	hydroxymyristoyl-acyl carrier protein dehydratase
91	2.58	ytbE.	hypothetical protein L101219
92	2.60	ksgA.	dimethyladenosine transferase
93	2.61	yedA.	hypothetical protein L33187
94	2.61	yseF.	hypothetical protein L29491
95	2.62	purN.	phosphoribosylglycinamide formyltransferase
96	2.62	ykhE.	hypothetical protein L72684
97	2.63	rnhB.	ribonuclease HII
98	2.64	infC.	translation initiation factor IF-3
99	2.65	tgt.	queuine tRNA-ribosyltransferase
100	2.66	rnpA.	ribonuclease P
101	2.66	yciH.	hypothetical protein L89418
102	2.67	ykjA.	hypothetical protein L90693
103	2.67	purF.	amidophosphoribosyltransferase
104	2.68	comGA.	ComGA
105	2.70	guaB.	inositol-5-monophosphate dehydrogenase
106	2.70	apt.	adenine phosphoribosyltransferase
107	2.70	glmS	glucosamine--fructose-6-phosphate aminotransferase
108	2.71	ysfB.	ABC transporter ATP-binding protein

109	2.72	ytfB.	hypothetical protein L133761
110	2.75	secY.	preprotein translocase subunit SecY
111	2.75	guaA	bifunctional GMP synthase/glutamine amidotransferase protein
112	2.78	yheB.	hypothetical protein L141547
113	2.78	ylaF.	nicotinate phosphoribosyltransferase
114	2.81	trxH.	thioredoxin H-type
115	2.83	comEA.	ComEA
116	2.84	typA.	TypA
117	2.85	hflX.	HflX
118	2.88	cmk.	cytidine monophosphate kinase
119	2.89	yccJ.	hypothetical protein L28204
120	2.90	rimM.	16S rRNA-processing protein
121	2.91	yccG.	hypothetical protein L26400
122	2.91	ftsQ.	FtsQ
123	2.91	uvrC.	excinuclease ABC subunit C
124	2.92	rplI.	50S ribosomal protein L9
125	2.93	yfdE.	hypothetical protein L133367
126	2.93	fur.	ferric uptake regulator
127	2.95	llrC.	two-component system regulator
128	2.96	ezrA.	septation ring formation regulator EzrA
129	2.97	yncB.	hypothetical protein L122632
130	2.97	cysM.	cysteine synthase
131	2.98	yhhE.	hypothetical protein L173151
132	2.98	ydgI.	putative GTP pyrophosphokinase
133	2.98	ynbC.	hypothetical protein L110933
134	2.99	ykaC.	hypothetical protein L5517
135	3.03	ywbA.	hypothetical protein L193121
136	3.04	yhfD.	hypothetical protein L155396
137	3.05	yqgA.	hypothetical protein L61727
138	3.05	mutS or hexA.	MutS
139	3.06	yahG.	ABC transporter ATP binding protein
140	3.12	rnc.	ribonuclease III
141	3.13	glnR.	glutamine synthetase repressor
142	3.14	ffh.	signal recognition particle protein
143	3.14	yhjF.	hypothetical protein L196017
144	3.14	yjbE.	general stress protein GSP13
145	3.15	optD.	oligopeptide ABC trasporter ATP binding protein
146	3.15	gatC.	aspartyl/glutamyl-tRNA amidotransferase subunit C
147	3.16	rbfA.	ribosome-binding factor A
148	3.19	upp.	uracil phosphoribosyltransferase
149	3.20	dnaK.	molecular chaperone DnaK
150	3.21	yqdA.	hypothetical protein L39365
151	3.21	yigC.	hypothetical protein L862989
152	3.21	trxB1.	thioredoxin reductase
153	3.23	glpF1.	glycerol-3-phosphate dehydrogenase
154	3.24	tsf.	elongation factor Ts

13.1.11 DnaB

position	distance	gène	fonction de la protéine
1	1.55	cobQ.	cobyric acid synthase
2	1.67	plpB.	outer membrane lipoprotein precursor
3	1.69	nrdG.	anaerobic ribonucleoside-triphosphate reductase activating protein
4	1.78	ppiB.	peptidyl-prolyl cis-trans isomerase
5	1.83	plpA.	outer membrane lipoprotein precursor
6	2.00	yveC.	hypothetical protein L126998
7	2.24	potA.	spermidine/putrescine ABC transporter ATP-binding protein
8	2.35	ydgC.	amino acid permease
9	2.39	potC.	spermidine/putrescine ABC transporter permease protein
10	2.39	ppiA.	peptidyl-prolyl cis-trans isomerase
11	2.52	acmD.	N-acetylmuramidase
12	2.58	dinG or dinP.	DNA polymerase IV
13	2.63	yeaG.	mevalonate kinase
14	2.72	ytgH.	hypothetical protein L142733
15	2.72	dnaG or dnaE.	DNA primase
16	2.81	ytdD.	hypothetical protein L119456
17	2.82	potB.	spermidine/putrescine ABC transporter permease protein
18	2.85	ymgG.	hypothetical protein L65637
19	2.86	hmcM.	hydroxymethylglutaryl-CoA synthase
20	2.87	yeaH.	diphosphomevalonate decarboxylase
21	2.91	ydgB.	amino acid permease
22	2.95	murB.	UDP-N-acetylenolpyruvoylglucosamine reductase
23	2.97	trxA.	thioredoxin
24	2.99	potD.	spermidine/putrescine ABC transporter substrate binding protein
25	3.01	mycA.	myosin-crossreactive antigen
26	3.04	ysfC.	polysacharide biosynthesis protein
27	3.07	mtsC.	manganese ABC transporter permease protein
28	3.07	ylbD.	hypothetical protein L114717
29	3.13	tkt.	transketolase
30	3.16	yneH.	hypothetical protein L145757
31	3.19	mvaA.	hydroxymethylglutaryl-CoA reductase

13.1.12 DnaD

position	distance	gène	fonction de la protéine
1	0.70	ecsA.	ABC transporter ATP binding protein
2	0.86	yciD.	hypothetical protein L86677
3	0.88	prfA.	peptide chain release factor 1
4	0.89	ruvA.	Holliday junction DNA helicase motor protein
5	0.90	trxB1.	thioredoxin reductase
6	0.90	clpC.	ATP-dependent protease ATP-binding subunit
7	0.95	yqeL.	GTP-binding protein
8	0.99	yqjB.	hypothetical protein L93855
9	1.06	xpt.	exodeoxyribonuclease VII large subunit

10	1.08	yrjA.	hypothetical protein L173313
11	1.08	engC or yuaD.	ribosome-associated GTPase
12	1.08	yuhI.	hypothetical protein L60959
13	1.11	ylgC.	hypothetical protein L161988
14	1.11	trmE or thdF.	tRNA modification GTPase TrmE
15	1.12	yuaA.	hypothetical protein L184033
16	1.16	yjbE.	general stress protein GSP13
17	1.17	def.	peptide deformylase
18	1.19	yedA.	hypothetical protein L33187
19	1.23	rheB.	ATP-dependent RNA helicase
20	1.26	ftsZ.	cell division protein FtsZ
21	1.29	yccH.	hypothetical protein L26998
22	1.31	ftsY.	FtsY
23	1.33	yujA.	hypothetical protein L74738
24	1.36	dnaN.	DNA polymerase III subunit beta
25	1.36	cysM.	cysteine synthase
26	1.36	ywgA or recX.	recombination regulator RecX
27	1.38	yqaC.	hypothetical protein L4747
28	1.41	hslO or yudG.	Hsp33-like chaperonin
29	1.43	ffh.	signal recognition particle protein
30	1.44	ruvB.	Holliday junction DNA helicase B
31	1.44	ygjD.	4-alpha-glucanotransferase
32	1.46	yfjD.	tRNA/rRNA methyltransferase
33	1.46	ygaB.	hypothetical protein L1778
34	1.53	coiA.	CoiA
35	1.54	smpB.	SsrA-binding protein
36	1.56	ydgI.	putative GTP pyrophosphokinase
37	1.56	ywdG.	hypothetical protein L22498
38	1.56	yseF.	hypothetical protein L29491
39	1.57	prfB.	peptide chain release factor 2
40	1.57	yqdA.	hypothetical protein L39365
41	1.59	rnpA.	ribonuclease P
42	1.59	yhhE.	hypothetical protein L173151
43	1.59	ybaF.	hypothetical protein L105256
44	1.61	yeaA.	hypothetical protein L582
45	1.62	rplI.	50S ribosomal protein L9
46	1.63	yogG or cfa.	hypothetical protein L65498
47	1.63	ylxQ.	hypothetical protein L175450
48	1.63	yecE.	putative lipid kinase
49	1.63	yofM.	hypothetical protein L57401
50	1.65	trxH.	thioredoxin H-type
51	1.66	ptpL.	tyrosine phosphatase
52	1.69	gcp.	O-sialoglycoprotein endopeptidase
53	1.71	ysjC.	hypothetical protein L76216
54	1.71	yigC.	hypothetical protein L862989
55	1.72	ytbE.	hypothetical protein L101219
56	1.73	ykiC.	hypothetical protein L84257
57	1.75	yfcI.	hypothetical protein L128550
58	1.76	ykaC.	hypothetical protein L5517
59	1.77	yjiE.	hypothetical protein L188550
60	1.78	ywbA.	hypothetical protein L193121
61	1.78	ysiG.	hypothetical protein L1889726

62	1.78	zitQ.	zinc ABC transporter ATP binding protein
63	1.79	atpA.	ATPase
64	1.79	gidA.	tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA
65	1.79	trmH.	tRNA-guanosine methyltransferase
66	1.79	yhjF.	hypothetical protein L196017
67	1.82	ecsB.	ABC transporter permease protein
68	1.82	thrS.	theronyl-tRNA synthetase
69	1.82	deoB.	phosphopentomutase
70	1.83	sunL.	rRNA methylase
71	1.85	yheB.	hypothetical protein L141547
72	1.85	pyrG.	CTP synthetase
73	1.86	rimM.	16S rRNA-processing protein
74	1.87	yciH.	hypothetical protein L89418
75	1.88	yqjE.	hypothetical protein L100263
76	1.88	yccL.	hypothetical protein L29477
77	1.91	tgt.	queuine tRNA-ribosyltransferase
78	1.92	rnc.	ribonuclease III
79	1.92	clpX.	ATP-dependent protease ATP-binding subunit
80	1.93	yejH.	hypothetical protein L98583
81	1.93	kinD.	sensor protein kinase
82	1.94	yhfD.	hypothetical protein L155396
83	1.94	ybeB.	hypothetical protein L141748
84	1.95	ylfH.	N-acetylglucosamine catabolic protein
85	1.95	mutS or hexA.	MutS
86	1.95	lspA.	lipoprotein signal peptidase
87	1.96	yxbE.	hypothetical protein L103195
88	1.97	rbfA.	ribosome-binding factor A
89	1.97	yciC.	hypothetical protein L84502
90	1.98	radA.	DNA repair protein RadA
91	1.98	yrgF.	hypothetical protein L148945
92	1.98	ykiG.	hypothetical protein L87113
93	2.00	yxaC.	hypothetical protein L86471
94	2.00	obgL.	GTPase ObgE
95	2.01	yljI.	permease
96	2.01	murC.	UDP-N-acetylmuramate--L-alanine ligase
97	2.02	ytdF.	hypothetical protein L120355
98	2.02	murD.	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase
99	2.03	uvrC.	excinuclease ABC subunit C
100	2.05	purA.	adenylosuccinate synthetase
101	2.06	kinA.	sensor protein kinase
102	2.06	pta.	phosphotransacetylase
103	2.07	ysxL or engB.	GTPase EngB
104	2.08	ytfB.	hypothetical protein L133761
105	2.11	yccK.	hypothetical protein L28696
106	2.12	yccJ.	hypothetical protein L28204
107	2.14	comGA.	ComGA
108	2.14	ykjI.	hypothetical protein L98095
109	2.15	cmk.	cytidine monophosphate kinase
110	2.15	comFC.	ComFC
111	2.16	gidB.	glucose-inhibited division protein B
112	2.16	ybbE.	hypothetical protein L114325
113	2.17	prmA.	ribosomal protein L11 methyltransferase

114	2.19	yfdB.	hypothetical protein L131937
115	2.21	ygcC.	oxidoreductase
116	2.23	pheS.	phenylalanyl-tRNA synthetase subunit alpha
117	2.26	yeaD.	DNA replication intiation control protein YabA
118	2.26	lgt.	prolipoprotein diacylglyceryl transferase
119	2.26	metK.	S-adenosylmethionine synthetase
120	2.27	yqgA.	hypothetical protein L61727
121	2.30	phoL.	phosphate starvation inducible protein
122	2.30	secA.	preprotein translocase subunit SecA
123	2.30	yudL.	hypothetical protein L22691
124	2.30	yudI.	oxidoreductase
125	2.31	nusG.	transcription antitermination protein NusG
126	2.34	yteA.	hypothetical protein L123851
127	2.34	yfdE.	hypothetical protein L133367
128	2.35	trmD	tRNA (guanine-N(1)-)-methyltransferase
129	2.36	yljE.	hypothetical protein L193873
130	2.36	ygdA.	hypothetical protein L32389
131	2.37	ytdC.	hypothetical protein L118668
132	2.38	ycfF.	hypothetical protein L56208
133	2.38	nifU.	NifU
134	2.40	ysfB.	ABC transporter ATP-binding protein
135	2.40	rplE.	50S ribosomal protein L5
136	2.40	yhfF.	hypothetical protein L157023
137	2.41	yraC.	DNA polymerase III subunit delta
138	2.41	yhbE.	hypothetical protein L92295
139	2.41	fusA.	elongation factor G
140	2.41	rheA.	ATP-dependent RNA helicase
141	2.42	lplL.	lipoate-protein ligase
142	2.42	ligA.	NAD-dependent DNA ligase LigA
143	2.43	glmS	glucosamine--fructose-6-phosphate aminotransferase
144	2.43	gatB.	aspartyl/glutamyl-tRNA amidotransferase subunit B
145	2.43	cshA.	recombination factor protein RarA
146	2.43	ygiK.	hypothetical protein L87336
147	2.44	cysK.	cysteine synthase
148	2.45	ywdF.	hypothetical protein L20937
149	2.45	trmU.	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase
150	2.46	gltX.	glutamyl-tRNA synthetase
151	2.47	frr.	ribosome recycling factor
152	2.48	hemN.	coproporphyrinogen III oxidase
153	2.48	efp.	elongation factor P
154	2.50	yjaJ.	transcription regulator
155	2.50	ftsH or tma.	FtsH
156	2.50	ybcG.	hypothetical protein L122849
157	2.50	add or acpS.	acyl carrier protein synthase
158	2.51	ytaA.	hypothetical protein L84477
159	2.52	dhaK.	diacylglycerol kinase
160	2.52	recD.	exodeoxyribonuclease V alpha chain
161	2.53	yncB.	hypothetical protein L122632
162	2.53	ycfD.	hypothetical protein L52686
163	2.54	gmk.	guanylate kinase
164	2.54	yuhB.	protease
165	2.54	ydjD.	hypothetical protein L195751

166	2.55	ytgF.	hypothetical protein L150593
167	2.55	llrC.	two-component system regulator
168	2.56	aroC.	chorismate synthase
169	2.57	recJ.	single-stranded DNA specific exonuclease
170	2.57	rgrB.	GntR family transcription regulator
171	2.57	yyaL.	translation-associated GTPase
172	2.58	tig.	trigger factor
173	2.58	yhfB.	hypothetical protein L151062
174	2.60	miaA.	tRNA delta(2)-isopentenylpyrophosphate transferase
175	2.61	yljJ.	hypothetical protein L198787
176	2.61	ribC.	bifunctional riboflavin kinase/FMN adenylyltransferase
177	2.63	parA.	chromosome partitioning protein
178	2.63	dnaJ.	DnaJ
179	2.64	llrD.	two-component system regulator
180	2.64	yseI.	hypothetical protein L32195
181	2.65	trxA.	thioredoxin
182	2.65	lysS.	lysyl-tRNA synthetase
183	2.66	yhhG.	hypothetical protein L175136
184	2.66	rplL.	50S ribosomal protein L7/L12
185	2.67	yfjE.	flavodoxin
186	2.68	rnz or ygcA.	ribonuclease Z
187	2.68	pepA.	glutamyl aminopeptidase
188	2.68	pheT.	phenylalanyl-tRNA synthetase subunit beta
189	2.69	yjjG.	hypothetical protein L196206
190	2.70	secY.	preprotein translocase subunit SecY
191	2.73	dltB.	peptidoglycan biosynthesis protein
192	2.74	truB.	tRNA pseudouridine synthase B
193	2.76	ftsK.	FtsK
194	2.76	rgrA.	GntR family transcription regulator
195	2.77	yudK.	hypothetical protein L21717
196	2.77	purM.	phosphoribosylaminoimidazole synthetase
197	2.78	priA.	primosome assembly protein PriA
198	2.78	yheA.	hypothetical protein L140288
199	2.79	yuiC.	hypothetical protein L68401
200	2.79	yriD.	hypothetical protein L171588
201	2.81	uvrB.	excinuclease ABC subunit B
202	2.82	yjaF.	hypothetical protein L106356
203	2.82	dinF.	DinG
204	2.83	sdaA.	alpha-subuni L-serine dehydratase
205	2.85	rnhB.	ribonuclease HII
206	2.86	rluA.	pseudouridine synthase
207	2.87	ycjB.	hypothetical protein L91807
208	2.87	pyrP.	pyrimidine regulatory protein PyrR
209	2.88	pth.	peptidyl-tRNA hydrolase
210	2.88	yraE.	hypothetical protein L106425
211	2.89	yxfB.	hypothetical protein L141634
212	2.89	dnaE.	DNA polymerase III DnaE
213	2.90	yeaE.	hypothetical protein L6615
214	2.91	ftsQ.	FtsQ
215	2.92	nifZ.	pyridoxal-phosphate dependent aminotransferase
216	2.92	ykhF.	ABC transporter ATP binding protein
217	2.92	llrA.	two-component system regulator

218	2.93	rplA.	50S ribosomal protein L1
219	2.93	yueF.	putative protease
220	2.94	valS.	valyl-tRNA synthetase
221	2.94	yvaB.	hypothetical protein L85091
222	2.95	grpE.	GrpE
223	2.95	ybiD.	hypothetical protein L184159
224	2.96	greA.	transcription elongation factor
225	2.97	msmK.	multiple sugar ABC transporter ATP-binding protein
226	2.98	yudJ.	hypothetical protein L20683
227	2.99	ksgA.	dimethyladenosine transferase
228	3.02	yjbC.	hypothetical protein L112952
229	3.04	gatC.	aspartyl/glutamyl-tRNA amidotransferase subunit C
230	3.05	rplM.	50S ribosomal protein L13
231	3.05	comEA.	ComEA
232	3.06	ymdE.	hypothetical protein L38177
233	3.07	dnaH.	DNA polymerase III subunits gamma and tau
234	3.08	femD.	phosphoglucosamine mutase
235	3.09	nifS.	pyridoxal-phosphate dependent aminotransferase
236	3.09	trxB2.	thioredoxin reductase
237	3.09	yiiB.	hypothetical protein L81441
238	3.11	tkt.	transketolase
239	3.15	upp.	uracil phosphoribosyltransferase
240	3.16	cdsA.	phosphatidate cytidylyltransferase
241	3.17	serS.	seryl-tRNA synthetase
242	3.18	ynbC.	hypothetical protein L110933
243	3.18	atpF.	ATP synthase subunit b
244	3.18	udk.	uridine kinase
245	3.18	fur.	ferric uptake regulator
246	3.18	ezrA.	septation ring formation regulator EzrA
247	3.19	ydgF.	hypothetical protein L164461
248	3.19	yccG.	hypothetical protein L26400
249	3.20	rplJ.	50S ribosomal protein L10
250	3.21	ylqL.	GTP-binding protein
251	3.22	ybiE.	oxidoreductase
252	3.23	plsX	fatty acid/phospholipid synthesis protein
253	3.25	bmpA.	basic membrane protein A

13.1.13 DnaE

position	distance	gène	fonction de la protéine
1	1.35	polC.	DNA polymerase III PolC
2	1.37	atpA.	ATPase
3	1.38	recD.	exodeoxyribonuclease V alpha chain
4	1.69	yeaD.	DNA replication intiation control protein YabA
5	1.82	dnaJ.	DnaJ
6	1.87	priA.	primosome assembly protein PriA
7	1.88	ycjB.	hypothetical protein L91807
8	2.08	ygiK.	hypothetical protein L87336

9	2.22	yjaD.	transcription regulator
10	2.23	yejI.	hypothetical protein L99502
11	2.28	yxfB.	hypothetical protein L141634
12	2.28	pheT.	phenylalanyl-tRNA synthetase subunit beta
13	2.32	ykiC.	hypothetical protein L84257
14	2.33	yraC.	DNA polymerase III subunit delta
15	2.34	accA.	acetyl-CoA carboxylase carboxyl transferase subunit alpha
16	2.46	miaA.	tRNA delta(2)-isopentenylpyrophosphate transferase
17	2.46	infB.	translation initiation factor IF-2
18	2.50	smc.	chromosome segregation SMC protein
19	2.60	ptpL.	tyrosine phosphatase
20	2.62	yhfB.	hypothetical protein L151062
21	2.63	yraE.	hypothetical protein L106425
22	2.64	lacR.	lactose transport regulator
23	2.67	ywdF.	hypothetical protein L20937
24	2.69	yrjA.	hypothetical protein L173313
25	2.70	cca or pacL.	tRNA CCA-pyrophosphorylase
26	2.70	ybiE.	oxidoreductase
27	2.75	mutS or hexA.	MutS
28	2.76	secA.	preprotein translocase subunit SecA
29	2.77	yudL.	hypothetical protein L22691
30	2.78	yjbE.	general stress protein GSP13
31	2.79	yudK.	hypothetical protein L21717
32	2.84	rnz or ygcA.	ribonuclease Z
33	2.85	rgrB.	GntR family transcription regulator
34	2.87	prfB.	peptide chain release factor 2
35	2.87	ybaF.	hypothetical protein L105256
36	2.89	ftsK.	FtsK
37	2.89	dnaD.	DnaD
38	2.90	yciD.	hypothetical protein L86677
39	2.94	sunL.	rRNA methylase
40	2.94	dacB.	D-alanyl-D-alanine carboxypeptidase
41	2.95	yrgF.	hypothetical protein L148945
42	2.95	engC or yuaD.	ribosome-associated GTPase
43	2.95	comFA.	ComFA
44	2.96	yqjB.	hypothetical protein L93855
45	3.00	ysgC.	hypothetical protein L52064
46	3.00	yheB.	hypothetical protein L141547
47	3.01	trxB2.	thioredoxin reductase
48	3.06	clpC.	ATP-dependent protease ATP-binding subunit
49	3.08	mutY.	A/G-specific adenine glycosylase
50	3.12	yudI.	oxidoreductase
51	3.16	ftsY.	FtsY
52	3.17	ecsA.	ABC transporter ATP binding protein
53	3.20	yteB.	hypothetical protein L125707
54	3.24	ylcC.	hypothetical protein L125196
55	3.25	ung.	uracil-DNA glycosylase
56	3.25	ruvB.	Holliday junction DNA helicase B

13.1.14 DnaH

position	distance	gène	fonction de la protéine
1	0.42	dltC.	D-alanine--poly(phosphoribitol) ligase subunit 2
2	0.79	rnhB.	ribonuclease HII
3	0.86	yqgA.	hypothetical protein L61727
4	0.93	murF.	D-Ala-D-Ala adding enzyme
5	0.96	yccK.	hypothetical protein L28696
6	0.98	ykaC.	hypothetical protein L5517
7	1.05	yejH.	hypothetical protein L98583
8	1.11	cshA.	recombination factor protein RarA
9	1.19	yljI.	permease
10	1.25	ylaF.	nicotinate phosphoribosyltransferase
11	1.27	ybbE.	hypothetical protein L114325
12	1.28	yljJ.	hypothetical protein L198787
13	1.36	yciH.	hypothetical protein L89418
14	1.37	ycfD.	hypothetical protein L52686
15	1.42	dltB.	peptidoglycan biosynthesis protein
16	1.49	plsX	fatty acid/phospholipid synthesis protein
17	1.51	gidB.	glucose-inhibited division protein B
18	1.51	ftsH or tma.	FtsH
19	1.53	rheA.	ATP-dependent RNA helicase
20	1.55	yfdB.	hypothetical protein L131937
21	1.57	mraY.	phospho-N-acetylmuramoyl-pentapeptide-transferase
22	1.60	ykhE.	hypothetical protein L72684
23	1.61	ruvB.	Holliday junction DNA helicase B
24	1.61	rluA.	pseudouridine synthase
25	1.62	rheB.	ATP-dependent RNA helicase
26	1.70	yogG or cfa.	hypothetical protein L65498
27	1.75	yqaC.	hypothetical protein L4747
28	1.78	yncB.	hypothetical protein L122632
29	1.80	ygcC.	oxidoreductase
30	1.80	ylfH.	N-acetylglucosamine catabolic protein
31	1.81	ykjI.	hypothetical protein L98095
32	1.84	yseI.	hypothetical protein L32195
33	1.84	rbfA.	ribosome-binding factor A
34	1.85	gatC.	aspartyl/glutamyl-tRNA amidotransferase subunit C
35	1.85	comEA.	ComEA
36	1.86	rnc.	ribonuclease III
37	1.87	gidA.	tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA
38	1.88	ysxL or engB.	GTPase EngB
39	1.88	yuiC.	hypothetical protein L68401
40	1.88	yciC.	hypothetical protein L84502
41	1.88	ecsB.	ABC transporter permease protein
42	1.89	ligA.	NAD-dependent DNA ligase LigA
43	1.92	ffh.	signal recognition particle protein
44	1.92	comFC.	ComFC
45	1.92	engC or yuaD.	ribosome-associated GTPase
46	1.93	yccJ.	hypothetical protein L28204
47	1.95	ribC.	bifunctional riboflavin kinase/FMN adenylyltransferase
48	1.97	ylxQ.	hypothetical protein L175450
49	1.98	trxH.	thioredoxin H-type

50	1.98	pyrG.	CTP synthetase
51	2.01	lysS.	lysyl-tRNA synthetase
52	2.01	ysjC.	hypothetical protein L76216
53	2.05	lspA.	lipoprotein signal peptidase
54	2.06	ykbB.	hypothetical protein L13157
55	2.07	smpB.	SsrA-binding protein
56	2.08	rplL.	50S ribosomal protein L7/L12
57	2.11	ftsY.	FtsY
58	2.12	ywgA or recX.	recombination regulator RecX
59	2.13	gltX.	glutamyl-tRNA synthetase
60	2.14	hemN.	coproporphyrinogen III oxidase
61	2.14	ylqL.	GTP-binding protein
62	2.17	prfA.	peptide chain release factor 1
63	2.18	pcrA.	ATP-dependent helicase PcrA
64	2.19	trmE or thdF.	tRNA modification GTPase TrmE
65	2.19	grpE.	GrpE
66	2.20	phoL.	phosphate starvation inducible protein
67	2.20	yjjG.	hypothetical protein L196206
68	2.22	cysK.	cysteine synthase
69	2.23	yraC.	DNA polymerase III subunit delta
70	2.24	rplI.	50S ribosomal protein L9
71	2.25	ftsQ.	FtsQ
72	2.27	yyaL.	translation-associated GTPase
73	2.30	comGA.	ComGA
74	2.32	pth.	peptidyl-tRNA hydrolase
75	2.32	cysS.	cysteinyl-tRNA synthetase
76	2.34	pepA.	glutamyl aminopeptidase
77	2.37	yccH.	hypothetical protein L26998
78	2.37	ksgA.	dimethyladenosine transferase
79	2.38	ykhF.	ABC transporter ATP binding protein
80	2.38	dinF.	DinG
81	2.39	ybeB.	hypothetical protein L141748
82	2.41	hslO or yudG.	Hsp33-like chaperonin
83	2.42	glmS	glucosamine--fructose-6-phosphate aminotransferase
84	2.42	ywbA.	hypothetical protein L193121
85	2.43	secY.	preprotein translocase subunit SecY
86	2.44	ybcG.	hypothetical protein L122849
87	2.45	ylgC.	hypothetical protein L161988
88	2.45	yxaC.	hypothetical protein L86471
89	2.46	ftsZ.	cell division protein FtsZ
90	2.46	ispA.	farnesyl diphosphate synthase
91	2.46	yvaB.	hypothetical protein L85091
92	2.47	lepA.	GTP-binding protein LepA
93	2.51	yjaF.	hypothetical protein L106356
94	2.52	thrS.	theronyl-tRNA synthetase
95	2.53	kinD.	sensor protein kinase
96	2.53	yqdA.	hypothetical protein L39365
97	2.53	ptpL.	tyrosine phosphatase
98	2.54	ezrA.	septation ring formation regulator EzrA
99	2.56	yseF.	hypothetical protein L29491
100	2.58	dnaN.	DNA polymerase III subunit beta
101	2.60	ykiG.	hypothetical protein L87113

102	2.61	gcp.	O-sialoglycoprotein endopeptidase
103	2.62	yheB.	hypothetical protein L141547
104	2.63	ackA2.	acetate kinase
105	2.66	ecsA.	ABC transporter ATP binding protein
106	2.67	yciD.	hypothetical protein L86677
107	2.69	ysfB.	ABC transporter ATP-binding protein
108	2.70	hflX.	HflX
109	2.72	ykdB.	hypothetical protein L32731
110	2.72	kinC.	sensor protein kinase
111	2.74	purB.	adenylosuccinate lyase
112	2.74	gatB.	aspartyl/glutamyl-tRNA amidotransferase subunit B
113	2.75	radA.	DNA repair protein RadA
114	2.77	hemK.	protoporphyrinogen oxidase
115	2.78	def.	peptide deformylase
116	2.78	rnpA.	ribonuclease P
117	2.79	llrD.	two-component system regulator
118	2.82	yqjB.	hypothetical protein L93855
119	2.82	smc.	chromosome segregation SMC protein
120	2.82	dhaK.	diacylglycerol kinase
121	2.83	ybaF.	hypothetical protein L105256
122	2.83	mutS or hexA.	MutS
123	2.83	ygaB.	hypothetical protein L1778
124	2.83	trxB1.	thioredoxin reductase
125	2.84	ydgF.	hypothetical protein L164461
126	2.85	uvrC.	excinuclease ABC subunit C
127	2.85	parA.	chromosome partitioning protein
128	2.86	recJ.	single-stranded DNA specific exonuclease
129	2.87	yrgF.	hypothetical protein L148945
130	2.87	trmU.	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase
131	2.88	ygjD.	4-alpha-glucanotransferase
132	2.88	optF.	oligopeptide ABC trasporter ATP binding protein
133	2.89	yigC.	hypothetical protein L862989
134	2.90	pfl.	pyruvate-formate lyase
135	2.91	yacB.	hypothetical protein L1001
136	2.91	yuhB.	protease
137	2.92	yuaA.	hypothetical protein L184033
138	2.93	ywdG.	hypothetical protein L22498
139	2.94	yuhI.	hypothetical protein L60959
140	2.95	priA.	primosome assembly protein PriA
141	2.96	yrjA.	hypothetical protein L173313
142	2.97	murC.	UDP-N-acetylmuramate--L-alanine ligase
143	2.97	ruvA.	Holliday junction DNA helicase motor protein
144	2.99	ysfC.	polysacharide biosynthesis protein
145	2.99	ackA1.	acetate kinase
146	3.01	rplJ.	50S ribosomal protein L10
147	3.02	cdsA.	phosphatidate cytidylyltransferase
148	3.03	pheS.	phenylalanyl-tRNA synthetase subunit alpha
149	3.04	purA.	adenylosuccinate synthetase
150	3.04	trmH.	tRNA-guanosine methyltransferase
151	3.06	valS.	valyl-tRNA synthetase
152	3.07	dnaD.	DnaD
153	3.07	yueE.	putative protease

154	3.07	yraB.	hypothetical protein L102051
155	3.07	ybeC.	hypothetical protein L142332
156	3.08	fusA.	elongation factor G
157	3.08	yjaJ.	transcription regulator
158	3.11	yjbE.	general stress protein GSP13
159	3.11	yxbE.	hypothetical protein L103195
160	3.11	ygdA.	hypothetical protein L32389
161	3.12	yuhH.	hypothetical protein L58914
162	3.12	ytbE.	hypothetical protein L101219
163	3.12	ydgI.	putative GTP pyrophosphokinase
164	3.14	yrjF.	hypothetical protein L178172
165	3.14	pta.	phosphotransacetylase
166	3.15	optD.	oligopeptide ABC trasporter ATP binding protein
167	3.15	frr.	ribosome recycling factor
168	3.15	ydjD.	hypothetical protein L195751
169	3.15	lgt.	prolipoprotein diacylglyceryl transferase
170	3.16	femD.	phosphoglucosamine mutase
171	3.16	yhjF.	hypothetical protein L196017
172	3.18	rluB.	pseudouridine synthase
173	3.19	sunL.	rRNA methylase
174	3.20	gmk.	guanylate kinase
175	3.21	infB.	translation initiation factor IF-2
176	3.21	ydaE.	cation transporter
177	3.21	pheT.	phenylalanyl-tRNA synthetase subunit beta
178	3.22	yqeL.	GTP-binding protein
179	3.22	yhhE.	hypothetical protein L173151
180	3.24	ycfF.	hypothetical protein L56208
181	3.24	guaA	bifunctional GMP synthase/glutamine amidotransferase protein
182	3.24	ytdC.	hypothetical protein L118668
183	3.24	lplL.	lipoate-protein ligase
184	3.24	dnaK.	molecular chaperone DnaK

13.1.15 DnaN

position	distance	gène	fonction de la protéine
1	0.59	radA.	DNA repair protein RadA
2	0.60	ecsA.	ABC transporter ATP binding protein
3	0.72	rnpA.	ribonuclease P
4	0.88	ykiG.	hypothetical protein L87113
5	1.04	ylxQ.	hypothetical protein L175450
6	1.09	rimM.	16S rRNA-processing protein
7	1.11	trxB1.	thioredoxin reductase
8	1.11	rplI.	50S ribosomal protein L9
9	1.12	def.	peptide deformylase
10	1.16	recF	DNA replication and repair protein RecF
11	1.19	trxH.	thioredoxin H-type
12	1.20	ftsZ.	cell division protein FtsZ
13	1.22	xpt.	xanthine phosphoribosyltransferase
14	1.23	pyk	pyruvate kinase

15	1.27	yqeL.	GTP-binding protein
16	1.30	yfdE.	hypothetical protein L133367
17	1.32	hslO or yudG.	Hsp33-like chaperonin
18	1.32	rheB.	ATP-dependent RNA helicase
19	1.33	yciH.	hypothetical protein L89418
20	1.36	dnaD.	DnaD
21	1.36	trmE or thdF.	tRNA modification GTPase TrmE
22	1.36	yuhI.	hypothetical protein L60959
23	1.37	smpB.	SsrA-binding protein
24	1.39	yseF.	hypothetical protein L29491
25	1.40	prfA.	peptide chain release factor 1
26	1.41	yqaC.	hypothetical protein L4747
27	1.41	gidB.	glucose-inhibited division protein B
28	1.42	yqjB.	hypothetical protein L93855
29	1.42	yuaA.	hypothetical protein L184033
30	1.42	ysiG.	hypothetical protein L1889726
31	1.44	ylgC.	hypothetical protein L161988
32	1.46	nifU.	NifU
33	1.47	uvrC.	excinuclease ABC subunit C
34	1.47	yujA.	hypothetical protein L74738
35	1.47	femD.	phosphoglucosamine mutase
36	1.48	ftsY.	FtsY
37	1.50	cmk.	cytidine monophosphate kinase
38	1.51	ybeB.	hypothetical protein L141748
39	1.53	cysM.	cysteine synthase
40	1.56	ecsB.	ABC transporter permease protein
41	1.56	ybcG.	hypothetical protein L122849
42	1.56	comFC.	ComFC
43	1.56	ruvA.	Holliday junction DNA helicase motor protein
44	1.59	ruvB.	Holliday junction DNA helicase B
45	1.60	ftsH or tma.	FtsH
46	1.63	engC or yuaD.	ribosome-associated GTPase
47	1.64	ffh.	signal recognition particle protein
48	1.66	gmk.	guanylate kinase
49	1.67	yqdA.	hypothetical protein L39365
50	1.67	ysxL or engB.	GTPase EngB
51	1.68	llrD.	two-component system regulator
52	1.68	ywgA or recX.	recombination regulator RecX
53	1.68	udk.	uridine kinase
54	1.68	yciD.	hypothetical protein L86677
55	1.70	yejH.	hypothetical protein L98583
56	1.71	yccK.	hypothetical protein L28696
57	1.72	yogG or cfa.	hypothetical protein L65498
58	1.72	yfdB.	hypothetical protein L131937
59	1.72	pyrG.	CTP synthetase
60	1.73	yxaC.	hypothetical protein L86471
61	1.74	comGA.	ComGA
62	1.75	purA.	adenylosuccinate synthetase
63	1.76	pheS.	phenylalanyl-tRNA synthetase subunit alpha
64	1.76	ykhF.	ABC transporter ATP binding protein
65	1.76	ygdA.	hypothetical protein L32389
66	1.77	yccH.	hypothetical protein L26998

67	1.77	atpA.	ATPase
68	1.79	ygjD.	4-alpha-glucanotransferase
69	1.79	gcp.	O-sialoglycoprotein endopeptidase
70	1.82	ywbA.	hypothetical protein L193121
71	1.82	yofM.	hypothetical protein L57401
72	1.83	yqgA.	hypothetical protein L61727
73	1.84	ydjD.	hypothetical protein L195751
74	1.85	ytfB.	hypothetical protein L133761
75	1.86	lysS.	lysyl-tRNA synthetase
76	1.86	clpX.	ATP-dependent protease ATP-binding subunit
77	1.86	kinD.	sensor protein kinase
78	1.87	yrgF.	hypothetical protein L148945
79	1.88	cshA.	recombination factor protein RarA
80	1.88	aroC.	chorismate synthase
81	1.88	pth.	peptidyl-tRNA hydrolase
82	1.88	gltX.	glutamyl-tRNA synthetase
83	1.88	yciC.	hypothetical protein L84502
84	1.89	clpC.	ATP-dependent protease ATP-binding subunit
85	1.90	yvaB.	hypothetical protein L85091
86	1.90	yfcI.	hypothetical protein L128550
87	1.91	mutS or hexA.	MutS
88	1.92	yrjA.	hypothetical protein L173313
89	1.93	ygcC.	oxidoreductase
90	1.93	efp.	elongation factor P
91	1.93	mraY.	phospho-N-acetylmuramoyl-pentapeptide-transferase
92	1.96	yeaA.	hypothetical protein L582
93	1.96	ytbE.	hypothetical protein L101219
94	1.96	rnc.	ribonuclease III
95	1.97	gidA.	tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA
96	1.98	yheB.	hypothetical protein L141547
97	1.98	yljJ.	hypothetical protein L198787
98	1.98	obgL.	GTPase ObgE
99	1.99	dltB.	peptidoglycan biosynthesis protein
100	1.99	yhfD.	hypothetical protein L155396
101	1.99	llrC.	two-component system regulator
102	2.00	ybbE.	hypothetical protein L114325
103	2.01	glmS	glucosamine--fructose-6-phosphate aminotransferase
104	2.02	yhhE.	hypothetical protein L173151
105	2.03	lspA.	lipoprotein signal peptidase
106	2.04	nusA.	transcription elongation factor NusA
107	2.04	yjiE.	hypothetical protein L188550
108	2.04	secY.	preprotein translocase subunit SecY
109	2.04	yjaF.	hypothetical protein L106356
110	2.04	yedA.	hypothetical protein L33187
111	2.06	ysfB.	ABC transporter ATP-binding protein
112	2.07	ptpL.	tyrosine phosphatase
113	2.08	thrS.	theronyl-tRNA synthetase
114	2.09	yyaL.	translation-associated GTPase
115	2.09	ycfF.	hypothetical protein L56208
116	2.11	ykjI.	hypothetical protein L98095
117	2.12	trmH.	tRNA-guanosine methyltransferase
118	2.13	phoL.	phosphate starvation inducible protein

119	2.13	rbfA.	ribosome-binding factor A
120	2.14	yigC.	hypothetical protein L862989
121	2.14	yjbE.	general stress protein GSP13
122	2.15	ezrA.	septation ring formation regulator EzrA
123	2.16	ydgI.	putative GTP pyrophosphokinase
124	2.20	ygaB.	hypothetical protein L1778
125	2.20	ytaA.	hypothetical protein L84477
126	2.21	murD.	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase
127	2.22	yncB.	hypothetical protein L122632
128	2.23	tig.	trigger factor
129	2.24	ylaF.	nicotinate phosphoribosyltransferase
130	2.24	yecE.	putative lipid kinase
131	2.24	sdaA.	alpha-subuni L-serine dehydratase
132	2.26	ykhE.	hypothetical protein L72684
133	2.26	deoB.	phosphopentomutase
134	2.27	ybaF.	hypothetical protein L105256
135	2.27	yjjG.	hypothetical protein L196206
136	2.29	rnhB.	ribonuclease HII
137	2.29	hemN.	coproporphyrinogen III oxidase
138	2.29	ylqL.	GTP-binding protein
139	2.29	ksgA.	dimethyladenosine transferase
140	2.32	yseI.	hypothetical protein L32195
141	2.32	gatB.	aspartyl/glutamyl-tRNA amidotransferase subunit B
142	2.34	pepA.	glutamyl aminopeptidase
143	2.36	yccJ.	hypothetical protein L28204
144	2.36	rplA.	50S ribosomal protein L1
145	2.37	ydgF.	hypothetical protein L164461
146	2.37	ywdG.	hypothetical protein L22498
147	2.38	pta.	phosphotransacetylase
148	2.41	rluB.	pseudouridine synthase
149	2.41	murC.	UDP-N-acetylmuramate--L-alanine ligase
150	2.41	ysjC.	hypothetical protein L76216
151	2.42	tgt.	queuine tRNA-ribosyltransferase
152	2.42	yhjF.	hypothetical protein L196017
153	2.43	kinA.	sensor protein kinase
154	2.44	fusA.	elongation factor G
155	2.44	trmU.	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase
156	2.45	ytdC.	hypothetical protein L118668
157	2.46	ycfD.	hypothetical protein L52686
158	2.48	rluA.	pseudouridine synthase
159	2.50	ytdF.	hypothetical protein L120355
160	2.50	rheA.	ATP-dependent RNA helicase
161	2.51	cysK.	cysteine synthase
162	2.52	purB.	adenylosuccinate lyase
163	2.52	dnaA.	chromosomal replication initiation protein
164	2.53	lepA.	GTP-binding protein LepA
165	2.53	frr.	ribosome recycling factor
166	2.53	yccF.	hypothetical protein L26054
167	2.53	rplM.	50S ribosomal protein L13
168	2.54	cdsA.	phosphatidate cytidylyltransferase
169	2.55	purN.	phosphoribosylglycinamide formyltransferase
170	2.58	dnaK.	molecular chaperone DnaK

171	2.58	yueF.	putative protease
172	2.58	prmA.	ribosomal protein L11 methyltransferase
173	2.58	dnaH.	DNA polymerase III subunits gamma and tau
174	2.59	yacB.	hypothetical protein L1001
175	2.60	ribC.	bifunctional riboflavin kinase/FMN adenylyltransferase
176	2.61	ybeC.	hypothetical protein L142332
177	2.63	pcrA.	ATP-dependent helicase PcrA
178	2.64	add or acpS.	acyl carrier protein synthase
179	2.64	rplJ.	50S ribosomal protein L10
180	2.65	nifS.	pyridoxal-phosphate dependent aminotransferase
181	2.65	ycjB.	hypothetical protein L91807
182	2.65	yxbE.	hypothetical protein L103195
183	2.67	guaA	bifunctional GMP synthase/glutamine amidotransferase protein
184	2.67	yfjD.	tRNA/rRNA methyltransferase
185	2.68	dnaC.	replicative DNA helicase
186	2.69	gatC.	aspartyl/glutamyl-tRNA amidotransferase subunit C
187	2.70	secA.	preprotein translocase subunit SecA
188	2.71	nusG.	transcription antitermination protein NusG
189	2.71	yljI.	permease
190	2.72	yueE.	putative protease
191	2.73	llrA.	two-component system regulator
192	2.74	prfB.	peptide chain release factor 2
193	2.74	coiA.	CoiA
194	2.74	ftsQ.	FtsQ
195	2.75	yccG.	hypothetical protein L26400
196	2.78	pg.	glucose-6-phosphate isomerase
197	2.78	tkt.	transketolase
198	2.79	yhfB.	hypothetical protein L151062
199	2.80	hflX.	HflX
200	2.81	yhbE.	hypothetical protein L92295
201	2.83	murB.	UDP-N-acetylenolpyruvoylglucosamine reductase
202	2.84	ligA.	NAD-dependent DNA ligase LigA
203	2.85	grpE.	GrpE
204	2.86	infC.	translation initiation factor IF-3
205	2.86	yuiC.	hypothetical protein L68401
206	2.86	rplE.	50S ribosomal protein L5
207	2.87	yjaJ.	transcription regulator
208	2.89	miaA.	tRNA delta(2)-isopentenylpyrophosphate transferase
209	2.92	zitQ.	zinc ABC transporter ATP binding protein
210	2.92	rplL.	50S ribosomal protein L7/L12
211	2.93	folC.	folylpolyglutamate synthase
212	2.93	ytgF.	hypothetical protein L150593
213	2.93	greA.	transcription elongation factor
214	2.97	pheT.	phenylalanyl-tRNA synthetase subunit beta
215	2.99	yuhB.	protease
216	2.99	comGB.	ComGB
217	2.99	yraB.	hypothetical protein L102051
218	2.99	yjjA.	hypothetical protein L190464
219	3.01	glnR.	glutamine synthetase repressor
220	3.01	comEA.	ComEA
221	3.03	yccL.	hypothetical protein L29477
222	3.09	yrjF.	hypothetical protein L178172

223	3.09	typA.	TypA
224	3.10	ynfG.	hypothetical protein L158193
225	3.11	tuf.	elongation factor Tu
226	3.11	dinF.	DinG
227	3.12	lplL.	lipoate-protein ligase
228	3.14	yhhG.	hypothetical protein L175136
229	3.17	recJ.	single-stranded DNA specific exonuclease
230	3.19	choS.	choline ABC transporter permease and substrate binding protein
231	3.21	yraC.	DNA polymerase III subunit delta
232	3.21	ykiC.	hypothetical protein L84257
233	3.22	hpt.	hypoxantine-guanine phosphorybosyltransferase
234	3.23	mtsC.	manganese ABC transporter permease protein
235	3.23	yheA.	hypothetical protein L140288
236	3.23	ispA.	farnesyl diphosphate synthase
237	3.23	optD.	oligopeptide ABC trasporter ATP binding protein
238	3.25	trmD	tRNA (guanine-N(1)-)-methyltransferase

13.1.16 SigX

position	distance	gène	fonction de la protéine
1	1.14	atpD	ATP synthase alpha subunit
2	1.53	ybdG.	transcription regulator
3	1.58	pi239.	prophage pi2 protein 39
4	1.60	atpG	ATP synthase gamma subunit
5	1.69	pi237.	prophage pi2 protein 37
6	1.75	tpx.	
7	1.82	yfjF.	membrane-bound transport protein
8	1.83	yojB.	hypothetical protein L95697
9	2.03	tra1077B.	IS1077D transposase
10	2.09	ycfA.	transcription regulator
11	2.09	pi240.	prophage pi2 protein 40
12	2.11	thiD1.	phosphomethylpyrimidine kinase
13	2.14	ywhB.	hypothetical protein L56236
14	2.16	pi114.	DNA replication protein
15	2.20	yqeD.	hypothetical protein L42411
16	2.20	yhjC.	hypothetical protein L193718
17	2.22	pi238.	prophage pi2 protein 38
18	2.23	arcD2.	arginine/ornitine antiporter
19	2.26	butB.	2,3-butanediol dehydrogenase
20	2.27	yliC.	hypothetical protein L183246
21	2.30	yjjB.	transcription regulator
22	2.31	yceE.	hypothetical protein L41779
23	2.33	ytgC.	hypothetical protein L145850
24	2.33	yniI.	hypothetical protein L187918
25	2.34	ybdA.	transcription regulator
26	2.36	rbsA.	ribose ABC transporter ATP binding protein
27	2.37	pi328.	prophage pi3 protein 28
28	2.38	pi232.	prophage pi2 protein 32
29	2.40	yucG.	chitin binding protein
30	2.41	yveA.	hypothetical protein L125244

31	2.42	rbsC.	ribose ABC transporter permease protein
32	2.42	yhcG.	hypothetical protein L121994
33	2.46	yfbM.	transcription regulator
34	2.49	pi242.	prophage pi2 protein 42
35	2.50	ybaH.	acetyl transferase
36	2.50	ykbA.	hypothetical protein L9876
37	2.51	ydgH.	hypothetical protein L165247
38	2.51	rlrC.	LysR family transcription regulator
39	2.51	yfcF.	hypothetical protein L123471
40	2.52	yxdF.	hypothetical protein L122569
41	2.53	rlrD.	LysR family transcription regulator
42	2.54	rbsB.	ribose ABC transporter substrate binding protein
43	2.55	yjjC.	ABC transporter ATP-binding protein
44	2.56	yviC.	FMN-binding protein
45	2.57	rmaI.	transcription regulator
46	2.57	yhhA.	hypothetical protein L169709
47	2.59	ytjE.	aminotransferase
48	2.59	xynB.	beta-1,4-xylosidase
49	2.59	ybbB.	hypothetical protein L111003
50	2.61	cpo.	non-heme chloride peroxidase
51	2.61	pdc.	phenolic acid decarboxylase
52	2.61	cydC.	cytochrome D ABC transporter ATP binding and permease protein
53	2.62	ps305.	prophage ps3 protein 05
54	2.63	yhjB.	hypothetical protein L193291
55	2.63	yviA.	hypothetical protein L163025
56	2.65	rmaA.	transcription regulator
57	2.65	yphJ.	hypothetical protein L178933
58	2.65	nah.	Na+/H+ antiporter
59	2.70	rmaH.	transcription regulator
60	2.71	yqeB.	hypothetical protein L40973
61	2.71	pi230.	terminase
62	2.72	floL.	flotillin-like protein
63	2.72	pi236.	prophage pi2 protein 36
64	2.73	yrgH.	hypothetical protein L152002
65	2.73	yndB.	hypothetical protein L130660
66	2.74	ycgA.	ABC transporter ATP binding protein
67	2.74	ybhD.	hypothetical protein L174946
68	2.75	yrgI.	hypothetical protein L152977
69	2.77	ycgH.	amidase
70	2.77	pi326.	terminase large subunit
71	2.80	ycfB.	ABC transporter ATP binding protein
72	2.81	yrjD.	hypothetical protein L176579
73	2.81	yvcC.	hypothetical protein L107379
74	2.85	pnuC2.	nicotinamide mononucleotide transporter
75	2.85	oppD.	oligopeptide ABC trasporter ATP binding protein
76	2.88	yrjC.	iron-binding oxidase subunit
77	2.90	ywiA.	hypothetical protein L63227
78	2.90	ycgB.	ABC transporter ATP binding protein
79	2.90	pi228.	prophage pi2 protein 28
80	2.91	pi135.	prophage pi1 protein 35
81	2.91	yjhE.	hypothetical protein L173848
82	2.91	yvaD.	hypothetical protein L89201

83	2.91	rbsD.	ribose ABC transporter permease protein
84	2.92	ycdH.	transporter
85	2.92	rmeB.	transcription regulator
86	2.93	pi349.	prophage pi3 protein 49
87	2.95	yffB.	hypothetical protein L153086
88	2.96	yddA.	transporter
89	2.97	mleP.	malate transporter
90	2.98	cydA.	cytochrome D ubiquinol oxidase subunit I
91	2.98	pi211.	topoisomerase
92	2.98	rliDB.	transcription regulator
93	2.98	ps216.	prophage ps2 protein 16
94	2.99	ycdF.	transcription regulator
95	2.99	yceG.	hypothetical protein L45062
96	3.00	ps219.	prophage ps2 protein 19
97	3.00	yiaD.	hypothetical protein L3272
98	3.01	arcD1.	arginine/ornitine antiporter
99	3.02	yccE.	hypothetical protein L24277
100	3.02	pi231.	prophage pi2 protein 31
101	3.02	ycfC.	ABC transporter permease protein
102	3.02	yucF.	hypothetical protein L9255
103	3.06	yrhH.	hypothetical protein L161059
104	3.06	rlrA	LysR family transcription regulator
105	3.06	ywdD.	hypothetical protein L19128
106	3.07	yahI.	short-chain type dehydrogenase
107	3.07	yoiC.	hypothetical protein L84260
108	3.09	yddD.	hypothetical protein L133932
109	3.09	napB.	hypothetical protein L106489
110	3.10	ykcE.	hypothetical protein L26878
111	3.10	pacA.	cation-transporting ATPase
112	3.11	dhaL.	dihydroxyacetone kinase
113	3.11	yfhJ.	hypothetical protein L177700
114	3.12	ydiC.	efflux pump antibiotic resistance protein
115	3.14	yjcA.	ABC transporter ATP binding protein
116	3.14	ybeD.	transcription regulator
117	3.15	ybdJ.	hypothetical protein L136552
118	3.15	ysiB.	permease
119	3.15	tenA.	TenA
120	3.16	ywjF.	3-hydroxyisobutyrate dehydrogenase
121	3.16	yveD.	hypothetical protein L127611
122	3.17	cydD.	cytochrome D ABC transporter ATP binding and permease protein
123	3.17	ykjH.	hypothetical protein L95481
124	3.17	yudH.	hypothetical protein L18647
125	3.18	pi360.	integrase
126	3.18	yxcB.	transcription regulator
127	3.20	ytjF.	hypothetical protein L181238
128	3.20	ykjJ.	hypothetical protein L98876
129	3.20	ypaD.	hypothetical protein L104065
130	3.20	yihD.	hypothetical protein L73264
131	3.20	noxC.	NADH oxidase
132	3.20	ykdA.	hypothetical protein L30853
133	3.21	tra981C.	transposase of IS904H
134	3.21	ps215.	prophage ps2 protein 15

135	3.21	pi251.	holin
136	3.22	ysjD.	hypothetical protein L76848
137	3.23	ogt.	6-O-methylguanine-DNA methyltransferase
138	3.23	yjhA.	hypothetical protein L169897
139	3.24	hly.	hemolysin like protein
140	3.24	yweC.	hypothetical protein L29314
141	3.24	pi323.	ATP dependent Clp protease

13.1.17 CcpA

position	distances	gène	fonction de la protéine
1	0.79	proS	prolyl-tRNA synthetase
2	1.10	mutY	A/G-specific adenine glycosylase
3	1.33	secA	preprotein translocase subunit SecA
4	1.48	purM	phosphoribosylaminoimidazole synthetase
5	1.49	uvrB	excinuclease ABC subunit B
6	1.66	gyrB	DNA gyrase subunit B
7	1.72	ftsK	FtsK
8	1.76	nifZ	pyridoxal-phosphate dependent aminotransferase
9	1.87	yjbC	hypothetical protein L112952
10	1.92	pknB	serine/threonine protein kinase
11	1.93	purD	phosphoribosylamine-glycine ligase
12	1.94	glnQ	glutamine ABC transporter ATP-binding protein
13	1.97	yeaE	hypothetical protein L6615
14	1.98	rcfB	transcription regulator
15	2.14	pepM	methionine aminopeptidase
16	2.15	yfjE	flavodoxin
17	2.17	yeaD	DNA replication intiation control protein YabA
18	2.17	ftsA	FtsA
19	2.18	yjaE	hypothetical protein L103741
20	2.20	atpD	F0F1 ATP synthase subunit beta
21	2.21	yraE	hypothetical protein L106425
22	2.22	dacB	D-alanyl-D-alanine carboxypeptidase
23	2.25	trxB2	thioredoxin reductase
24	2.25	folD	cyclohydrolase
25	2.28	nifS	pyridoxal-phosphate dependent aminotransferase
26	2.33	murI	glutamate racemase
27	2.33	yteB	hypothetical protein L125707
28	2.34	mtsA	manganese ABC transporter substrate binding protein
29	2.35	yjiF	hypothetical protein L189428
30	2.35	ftsW1	FtsW1
31	2.42	yccG	hypothetical protein L26400
32	2.42	yiiB	hypothetical protein L81441
33	2.49	yudJ	hypothetical protein L20683
34	2.52	dltD	D-alanine transfer protein
35	2.53	coiA	CoiA
36	2.53	pheT	phenylalanyl-tRNA synthetase subunit beta
37	2.56	pi348	single strand binding helix destabilising protein
38	2.58	rgrB	GntR family transcription regulator

39	2.58	pyrR	pyrimidine regulatory protein PyrR
40	2.59	ygbG	ribonuclease Z
41	2.59	yudL	hypothetical protein L22691
42	2.60	ung	uracil-DNA glycosylase
43	2.62	ybdD	hypothetical protein L132712
44	2.66	yxfB	hypothetical protein L141634
45	2.68	ylfD	hypothetical protein L154225
46	2.70	mtsB	manganese ABC transporter ATP binding protein
47	2.73	zitQ	zinc ABC transporter ATP binding protein
48	2.75	ydjD	hypothetical protein L195751
49	2.77	cdd	cytidine deaminase
50	2.79	bmpA	basic membrane protein A
51	2.82	mfd	transcription-repair coupling factor
52	2.83	serS	seryl-tRNA synthetase
53	2.87	yqjE	hypothetical protein L100263
54	2.95	yngF	sugar ABC transporter permease protein
55	2.97	aroA	3-phosphoshikimate 1-carboxyvinyltransferase
56	2.99	yqgE	transporter
57	3.00	prmA	ribosomal protein L11 methyltransferase
58	3.00	yngE	sugar ABC transporter ATP binding protein
59	3.01	mutS	MutS
60	3.01	pcrA	ATP-dependent helicase PcrA
61	3.01	recD	exodeoxyribonuclease V alpha chain
62	3.02	ytdC	hypothetical protein L118668
63	3.03	comFA	ComFA
64	3.04	ygiI	hypothetical protein L87336
65	3.07	holB	DNA polymerase III subunit delta'
66	3.09	miaA	tRNA delta(2)-isopentenylpyrophosphate transferase
67	3.12	yudK	hypothetical protein L21717
68	3.13	gcp	O-sialoglycoprotein endopeptidase
69	3.14	yheB	hypothetical protein L141547
70	3.17	yhdC	acetyl transferase
71	3.17	ftsE	cell-division ATP-binding protein
72	3.18	yejI	hypothetical protein L99502
73	3.18	priA	primosome assembly protein PriA
74	3.18	aspB	aspartate aminotransferase
75	3.19	ftsQ	FtsQ
76	3.20	ylcC	hypothetical protein L125196
77	3.21	pfs	5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase
78	3.22	aspS	aspartyl-tRNA synthetase
79	3.22	ysgC	hypothetical protein L52064
80	3.23	smc	chromosome segregation SMC protein
81	3.24	yeeE	hypothetical protein L44542
82	3.24	glk	glucose kinase

13.1.18 CopR

position	distances	gène	fontion de la protéine
1	1.23	plpC	outer membrane lipoprotein precursor
2	1.41	ybaI	glycosil transferase

3	1.90	ysbA	hypothetical protein L194765
4	1.90	yedE	hypothetical protein L37338
5	1.93	xerS	site-specific tyrosine recombinase XerS
6	1.94	ybiC	hypothetical protein L183216
7	2.08	yjcF	hypothetical protein L126819
8	2.08	yrbD	hypothetical protein L113400
9	2.10	pstF	phosphate ABC transporter substrate binding protein
10	2.10	dexB	alpha 1-6-glucosidase
11	2.14	ybfB	hypothetical protein L159400
12	2.19	yhdC	acetyl transferase
13	2.21	yidB	cellobiose-specific PTS system IIC component
14	2.21	pbp2A	penicillin-binding protein 2a
15	2.29	rgrB	GntR family transcription regulator
16	2.30	ylfI	hypothetical protein L157321
17	2.33	ypiB	transporter
18	2.36	pstE	phosphate ABC transporter substrate binding protein
19	2.36	yhgC	transcription regulator
20	2.36	yieH	hypothetical protein L48341
21	2.37	ysiD	hypothetical protein L68066
22	2.41	dacA	D-alanyl-D-alanine carboxypeptidase
23	2.42	yveI	hypothetical protein L131806
24	2.43	ykiD	hypothetical protein L84937
25	2.44	ytcE	hypothetical protein L112263
26	2.45	yfjB	transposon-related protein
27	2.46	yrbI	transcription regulator
28	2.48	parE	DNA topoisomerase IV subunit B
29	2.50	yqaD	hypothetical protein L5610
30	2.50	ylbA	ABC transporter ATP-binding protein
31	2.50	yngB	fibronectin-binding protein
32	2.51	xpt	xanthine phosphoribosyltransferase
33	2.52	comEC	ComEC
34	2.53	yxdC	cation-transporting ATPase
35	2.54	yvdA	hypothetical protein L114054
36	2.54	yfjE	flavodoxin
37	2.56	ywaF	glycosyltransferase
38	2.57	aroK	shikimate kinase
39	2.58	nagA	N-acetylglucosamine-6-phosphate deacetylase
40	2.60	ptcC	cellobiose-specific PTS system IIC component
41	2.60	yrgE	hypothetical protein L148513
42	2.60	yuaB	hypothetical protein L185224
43	2.61	plpD	outer membrane lipoprotein precursor
44	2.61	ypcB	hypothetical protein L121071
45	2.62	xseB	exonuclease VII small subunit
46	2.62	yhbF	hypothetical protein L115789
47	2.63	ysgC	hypothetical protein L52064
48	2.64	ung	uracil-DNA glycosylase
49	2.65	pi360	integrase
50	2.67	fhuR	fhu operon transcription regulator
51	2.72	pbp1B	penicillin-binding protein 1B
52	2.72	ywaG	lipopolysaccharide biosynthesis protein
53	2.72	rgpB	rhamnosyltransferase
54	2.75	yheD	hypothetical protein L143879

55	2.76	ps316	integrase
56	2.77	xylH	4-oxalocrotonate tautomerase
57	2.77	ylbB	ABC transporter permease protein
58	2.79	groEL	chaperonin GroEL
59	2.79	ylcC	hypothetical protein L125196
60	2.79	recA	recombinase A
61	2.80	hasC	UTP-glucose-1-phosphate uridylyltransferase
62	2.81	yjaB	hypothetical protein L102093
63	2.82	htrA	exported serine protease
64	2.83	parC	DNA topoisomerase IV subunit A
65	2.84	ywiI	transcription regulator
66	2.86	yliF	hypothetical protein L184880
67	2.86	polA	DNA polymerase I
68	2.87	yxdB	hypothetical protein L117205
69	2.88	yeiE	hypothetical protein L85854
70	2.89	noxE	NADH oxidase
71	2.89	yfjC	acylphosphate phosphohydrolase
72	2.89	yrjE	transport permease
73	2.90	ydbF	transcription regulator
74	2.90	xerD	site-specific tyrosine recombinase XerD-like protein
75	2.91	ssbA	single-strand DNA-binding protein
76	2.91	pfs	5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase
77	2.92	uvrB	excinuclease ABC subunit B
78	2.94	yqjD	hypothetical protein L98109
79	2.95	recN	DNA repair protein RecN
80	2.95	ynbB	hypothetical protein L110441
81	2.96	pepDA	dipeptidase
82	2.97	ysfD	regulatory protein
83	2.98	yhcB	hypothetical protein L122299
84	2.99	ylfD	hypothetical protein L154225
85	3.00	ygbG	ribonuclease Z
86	3.00	pepDB	dipeptidase
87	3.01	malE	maltose ABC transporter substrate binding protein
88	3.02	murE	UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase
89	3.02	yfgL	hypothetical membrane protein
90	3.04	tag	DNA-3-methyladenine glycosidase I
91	3.04	yidC	beta-glucosidase
92	3.05	yliA	positive transcription regulator
93	3.08	yufA	hypothetical protein L34138
94	3.10	yabA	transcription regulator
95	3.10	yqgE	transporter
96	3.10	yfjA	hypothetical protein L189883
97	3.13	ywdE	transcription regulator
98	3.13	yddD	hypothetical protein L133932
99	3.14	ytcB	hypothetical protein L104285
100	3.14	yceE	hypothetical protein L41779
101	3.16	yvhA	hypothetical protein L154438
102	3.17	yvdF	amino acid ABC transporter substrate binding protein
103	3.22	ybaA	hypothetical protein L101209
104	3.24	rliC	transcription regulator