Développement d'une base de données sur la résistance aux antibiotiques et son utilisation en génomique.

Mémoire

Maxime Déraspe

Maîtrise en biochimie
Maître ès sciences (M.Sc.)

Québec, Canada

Résumé

Le projet de maîtrise consistait à développer une base de données (BD) sur la résistance bactérienne aux antibiotiques et de l'utiliser dans les analyses bio-informatiques de deux projets de génomiques. La BD MERGEM (« Mobile Elements and Resistance Genes Enhanced for Metagenomics ») mettait l'emphase sur la bonne nomenclature des gènes et la fiabilité de l'annotation de leurs séquences, qui s'avère un réel problème dans les BD publiques en biologie. La BD MERGEM mit aussi de l'avant l'utilisation de technologies du Web sémantique et de développement Web pour enrichir et publier son contenu. De plus, un pipeline bio-informatique d'annotations fonctionnelles fut réalisé dans le but de correctement identifier les éléments de MERGEM et leur contexte génomique dans deux projets de séquençages importants : 264 métagénomes du microbiote intestinale et 390 génomes de *Pseudomonas aeruginosa*. Les résultats démontrent l'utilité de développer des BD spécialisées en génomique.

Abstract

The current Master's project consist of the development of a database (DB) on bacterial antibiotic resistance and its use in bioinformatic analyses for two major genomic projects. The DB is called MERGEM (Mobile Elements and Resistance Genes Enhanced for Metagenomics) and puts a particular emphasis on a good genes nomenclature and the reliability of the annotation of their sequences, which is a real problem in biological public databases. The MERGEM database also adopts technologies of the Semantic Web and utilizes Web development to enrich and publish its content. Furthermore, a bioinformatic annotation pipeline was developed in order to correctly identify MERGEMs' genes and their contexts in two important sequencing projects : one with 264 metagenomes from the human gut microbiome and another one consisting of 390 *Pseudomonas aeruginosa* genomes. The results of this project proves the usefulness of specialized databases in genomic studies.

Table des matières

Liste des tableaux

Liste des figures

Je dédie ce mémoire à mes adorables enfants et très chers parents.
En espérant que cet ouvrage puisse survivre à autant de générations.

Remerciements

Je voudrais tout d'abord remercier mes parents qui m'ont toujours supportés dans mes études ainsi que la réalisation de mes projets, comme celui d'écrire ce mémoire. Je remercie mes enfants qui sont tout simplement le soleil de ma vie et sans qui je n'aurais jamais même entamé l'écriture de ce mémoire.

Je remercie aussi mes directeurs, les Dr Jacques Corbeil et Paul H. Roy, pour leur compréhension, leur soutien et les nombreuses connaissances qu'ils ont su m'inculquer durant ces dernières années. Je remercie aussi leur équipe de recherches, les différentes équipes de bio-informatique du CHUL, tous ceux qui agrémentent mon quotidien là-bas et mes amis en général.

Introduction

La découverte des antibiotiques est très certainement l'une des plus fructueuses sur la santé humaine au 20^e siècle. Elle a permis de guérir un bon nombre de maladies infectieuses ayant pu être mortelles dans l'ère préantibiotique et a ainsi augmenté de plusieurs années l'espérance de vie. On attribut la découverte du premier antibiotique à Alexander Fleming qui en 1928, par inadvertance, s'aperçut de l'effet létal de la pénicilline sur des bactéries en culture. S'en suivit l'âge d'or (∼1940-1970) des antibiotiques, période pendant laquelle il y eut le plus grand nombre de découvertes des différentes classes d'antibiotiques connues à ce jour. Certains voyaient déjà la bataille contre les maladies infectieuses réglée pour de bon, comme en témoignent les propos attribués au Dr William H. Stewart dans les années 1965-1969 : « *It is time to close the book on infectious diseases, and declare the war against pestilence won* » (Spellberg, 2008). Le Dr Stewart était alors le « US Surgeon General », celui qui doit rendre des comptes au gouvernement américain sur les avancées scientifiques dans le domaine de la santé et sur la prévention des risques liés aux maladies.

Malheureusement, il n'aurait pu s'avérer plus erroné que d'affirmer la bataille déjà gagnée contre les infections causées par les micro-organismes. Ces derniers ont l'avantage d'évoluer beaucoup plus rapidement que nous, les humains, et cela depuis bien plus longtemps. Sans compter leur capacité à transmettre du matériel génétique entre eux, tels des gènes de résistances aux antibiotiques. Jamais ce ne fut une question de si les bactéries allaient développer une résistance aux antibiotiques, mais plutôt à savoir quand cela se produirait. Déjà Fleming était conscient que l'utilisation impropre des antibiotiques pourrait engendrer de la résistance chez les bactéries (Levy, 1999).

Aujourd'hui, la résistance aux antibiotiques est reconnue à son juste titre, comme une réelle menace à notre capacité de se défendre contre ces microbes nuisibles à la santé humaine. Beaucoup d'organisations dont l'OMS [1], le CDC [2], la CARA [3], l'APUA [4] et plusieurs autres, se préoccupent maintenant de ce fléau et sensibilisent les différents intervenants (gouvernements, hôpitaux, médecins, scientifiques, etc.) à appréhender une période postantibiotique.

1. Organisation Mondiale de la Santé
2. *Centers for Disease Control and Prevention*
3. *Canadian Antimicrobial Resistance Alliance*
4. *Alliance for the Prudent Use of Antibiotics*

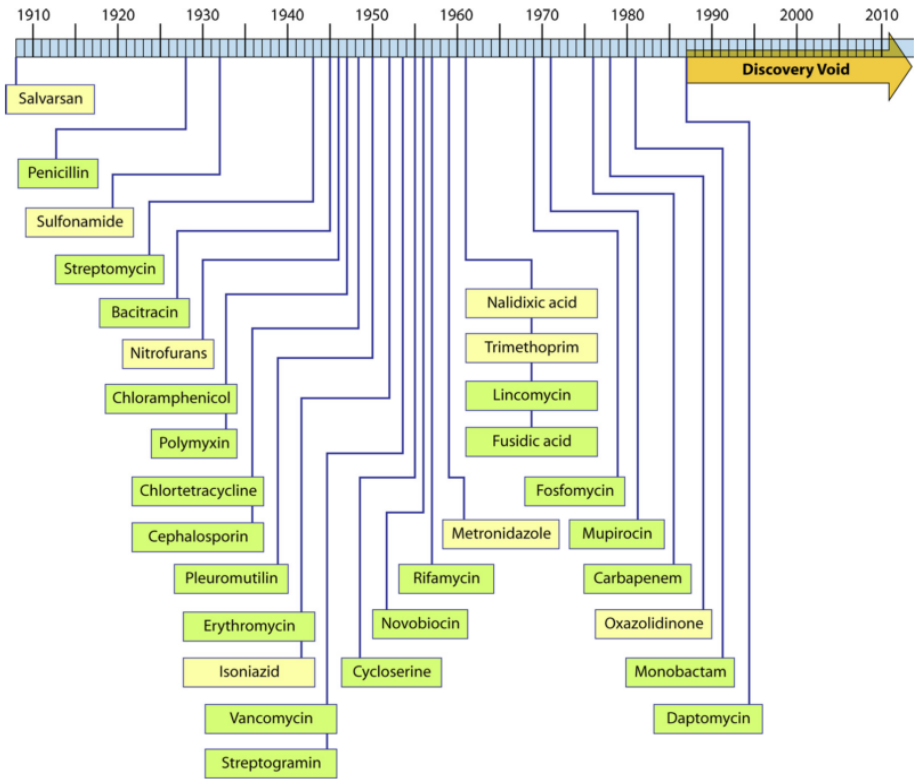


Figure 0.1 – Chronologie de la découverte des antibiotiques, basée sur la date de parution du brevet ou de la découverte initiale de la formulation du médicament. Tirée de Silver (2011) avec permission.

Le dernier quart de siècle fut plutôt pauvre en découverte de nouveaux antimicrobiens, comme le démontre la ligne du temps de la figure 0.1. L'une des principales raisons qui a mené les compagnies pharmaceutiques à délaisser la recherche de nouveaux antimicrobiens est que le coût et la complexité à amener une nouvelle molécule sur le marché sont très élevés et peut-être qu'il s'avère tout simplement plus lucratif de soigner les maladies chroniques que de simples infections passagères (Projan, 2003). De plus, un nouvel antibiotique qui arriverait sur le marché ne serait certainement pas prescrit en masse question de préserver son efficacité en évitant de sélectionner de la résistance à celui-ci trop rapidement. Malgré tout, les maladies infectieuses demeurent la seconde cause de mortalité dans le monde et les organismes en santé publique, autant en Europe qu'en Amérique du Nord, ouvrent maintenant la voie aux compagnies et aux universités en investissant dans la recherche et le développement de nouvelles molécules (Plumridge, 2014). Les antibiotiques et leurs résistances seront vue plus en détails au chapitre 1.

L'une des plus importantes avancées technologiques du 21e siècle pour la recherche sur les maladies infectieuses est sans aucun doute les nouvelles générations de séquençages (NGS) (Fauci, 2001). Il est maintenant possible et abordable de séquencer le génome complet d'une bactérie ou même d'une communauté entière de microbes (microbiome) dans des temps rai-

sonnables. L'étude génomique de ces organismes a permis d'approfondir nos connaissances sur les mécanismes de virulence et de résistance qu'abritent les pathogènes. Les technologies d'informations ont aussi beaucoup évolué et permettent des analyses informatiques intensives pour pallier les données massives produites par les NGS. L'accumulation de ces données génomiques qui découlent des projets de séquençages d'à travers le monde a grandement surpassé nos connaissances biologiques et biochimiques fondamentales de tous ces organismes séquencés. Il en résulte un grand nombre d'éléments fonctionnels (protéines, ARN, etc.) inconnus qui sont stockés dans les bases de données (BD) publiques. Beaucoup de ces projets comptent sur des annotations automatiques qui répliquent les identifications géniques provenant de projets antérieurs. Malheureusement, un nombre considérable d'erreurs d'annotations subsistent dans ces bases de données publiques ce qui rend leur fiabilité parfois médiocre (Schnoes *et al.*, 2009). C'est pourquoi un lot de bases de données spécialisées sont régulièrement créées et publiées par différents groupes de recherches pour mûrir les connaissances sur certains processus biologiques d'intérêts. L'établissement de nomenclatures standardisées et acceptées par tous est une approche viable et désirable dans un milieu complexe telle la génomique. Le cas de la résistance aux antibiotiques ne fait pas exception à cette règle. C'est au chapitre 2 que sera introduite la génomique et les BD biologiques en général.

Au début du projet de recherche (2012), il existait principalement une seule base de données qui couvrait un maximum de gènes de résistances aux antibiotiques et qui était disponible au téléchargement sur le Web : « Antibiotic Resistance Database » (ARDB) (Liu et Pop, 2009). ARDB, publiée en 2009, fut utilisée dans plusieurs projets de génomique bactérienne, pour avoir été citée à 163 reprises. Elle répondait donc à un besoin qu'était l'identification de gènes de résistances aux antibiotiques dans les génomes bactériens. Malheureusement, le développement de la BD s'arrêta en 2009 et elle ne fut plus mise à jour depuis. L'opportunité était donc là pour développer une nouvelle BD sur la résistance aux antibiotiques et c'est ainsi que débuta le projet de MERGEM - « Mobile Elements and Resistance Genes Enhanced for Metagenomics ». Plus récemment (juillet 2013), « The Comprehensive Antibiotic Resistance Database » (CARD) publiait leur BD avec un contenu plus à jour que ARDB en incluant beaucoup de gènes caractérisés depuis 2009, mais contenait à la date de leur publication moins de gènes au total que la BD MERGEM. Cependant, l'un des avancements intéressants réalisés par CARD fut la construction d'une ontologie sur la résistance - « Antibiotic Resistance Ontology » (ARO). La beauté des ontologies est leur capacité à modéliser un ensemble de connaissances sous forme de graphe. L'une des ontologies des plus connues et utilisées en pratique vient du projet « Gene Ontology » (GO), qui standardise le vocabulaire pour l'annotation fonctionnelle des gènes de beaucoup d'organismes. L'avantage de travailler avec des ontologies est que leur sémantique peut aussi bien être comprise par un humain qu'un ordinateur. Les ontologies sont d'autant plus attrayantes, puisqu'elles nous mènent dans le merveilleux monde du Web sémantique. Le Web sémantique est un concept mené par le W3C [5] et qui fut inventé par le

5. *World Wide Web Consortium*

créateur même du Web, Sir Timothy John "Tim" Berners-Lee. Pour reprendre ses propos : « *The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.* » (Berners-Lee *et al.*, 2001). La dernière section du chapitre 2 traitera plus en profondeur des particularités du Web sémantique et des technologies s'y rattachant. Il y sera aussi discuté de la popularité croissante du Web sémantique chez les grandes institutions de fournisseurs de données biologiques et de son utilité en génomique.

Le chapitre 3 se penchera sur le développement et l'implémentation de la base de données MERGEM et des spécificités techniques s'y rattachant. Le chapitre 4, quant à lui, se penchera sur les différents cas d'utilisations de la BD et des différents projets connexes ayant bénéficié de celle-ci ; entre autres le projet CQDM [6] portant sur la sélectomique pour suivre et prédire l'émergence de résistances aux antibiotiques, ainsi que deux projets portant sur des souches de *Pseudomonas aeruginosa* résistantes aux antibiotiques.

6. Consortium Québécois sur la Découverte du Médicament

Chapitre 1

Résistance aux antibiotiques

1.1 Les antibiotiques

1.1.1 Origines

Les antibiotiques sont connus pour avoir été découverts accidentellement par le Dr Alexander Fleming, en 1928. En effet, il s'aperçut après un retour de vacances qu'une culture du champignon, *Penicillium notatum*, avait envahi l'une de ses boîtes de Petri et inhibait la croissance de sa culture bactérienne de staphylocoques (Greek, 1999). L'ingrédient actif produit par le champignon était en fait la pénicilline, un antibiotique de type β-lactamine. Ce fut Howard Florey, Ernst Chain (récipiendaire du prix Nobel avec Fleming en 1945) et leurs collègues de l'Université d'Oxford qui firent de la pénicilline, en 1939, un réel médicament grâce à leurs techniques d'extraction et de purification de la molécule. En 1940, Florey utilisait avec succès la pénicilline chez des souris en les guérissant d'une infection aux streptocoques et c'est en 1941 qu'un premier test fut réalisé chez l'homme. Le patient fut rétabli le temps qu'il était sous l'effet de l'antibiotique, mais décéda durant les jours qui suivirent la fin du traitement dû à une malheureuse rupture de stock. À l'été 1941, Florey et un collègue voyagèrent aux États-Unis pour courtiser les pharmaceutiques, question de produire de la pénicilline en plus grande quantité. De fil en aiguille, plusieurs compagnies embarquèrent (Merck, Pfizer, Squibb) pour améliorer les techniques de fermentation en vue d'une production massive de l'antibiotique. De plus, en 1943, le gouvernement américain supporta aussi cette production massive dans le cadre de l'effort de guerre et désignait 21 compagnies pour participer au projet. Cet ainsi que des quantités industrielles de pénicillines furent produites en plein temps de guerre, et ce même antibiotique sauva assurément des millions de vies depuis.

Il serait aussi important de bien définir ce qu'est un antibiotique autrement qu'avec l'historique de sa découverte. L'étymologie du mot antibiotique révèle son origine dans le mot antibiose qui décrit l'effet antagoniste que peuvent avoir les micro-organismes entre eux. La définition la plus couramment utilisée est cependant celle du Dr Selman A. Waksman établit en 1947

(Waksman, 1947; Bentley et Bennett, 2003) :
« *An antibiotic is a chemical substance, produced by micro-organisms, which has the capacity to inhibit the growth of and even to destroy bacteria and other micro-organisms. The action of an antibiotic against micro-organisms is selective in nature, some organisms being affected and others not at all or only to a limited degree ; each antibiotic is thus characterized by a specific antimicrobial spectrum. The selective action of an antibiotic is also manifested against microbial vs. host cells. Antibiotics vary greatly in their physical and chemical properties and in their toxicity to animals. Because of these characteristics, some antibiotics have remarkable chemotherapeutic potentialities and can be used for the control of various microbial infections in man and animals.* »
On pourrait simplifier la définition en trois points essentiels :

1. un antibiotique est une molécule produite par un micro-organisme qui a la capacité de tuer où de prévenir la croissance d'une bactérie ;
2. le spectre d'un antibiotique est sélectif, il n'agira donc pas de la même façon face à différentes bactéries ;
3. il existe une grande diversité de molécules avec des propriétés antibiotiques et leur toxicité diffère lorsqu'administrée chez les animaux.

Le terme antibiotique est souvent utilisé pour des molécules synthétiques dérivées d'une autre produite par un micro-organisme. Pour désambiguïsation, ces dernières devraient plutôt être désignées comme agents antibactériens ou antimicrobiens, mais seront volontairement nommées antibiotique dans ce manuscrit.

1.1.2 Action des antibiotiques

Les antibiotiques peuvent être généralement classés en deux catégories selon leur potentiel d'action, soit bactéricide ou bactériostatique. De par leur définition, bactéricide implique que l'agent tuera la bactérie alors que bactériostatique signifie que l'agent préviendra la croissance de celle-ci. Cependant, certains antibactériens pourraient, semble-t-il, avoir les deux modes d'action et cela dépendrait simplement de la concentration administrée pour avoir un mode ou l'autre. On pourrait avoir tendance à penser qu'un antibiotique bactéricide a une efficacité supérieure, mais aucune évidence clinique ne supporte cela. Un traitement avec un effet bactériostatique est souvent désirable particulièrement pour les infections dues aux bactéries à Gram négatif. En effet, bien que certaines situations cliniques nécessitent un antibiotique bactéricide (endocardites, méningites, ostéomyélite, neutropénie), des études ont démontré des effets indésirables d'un tel potentiel d'action et recommandent même des antibiotiques avec un mode bactériostatique pour certains traitements (Pankey et Sabath, 2004).

Malgré la diversité des molécules antibiotiques qui existent sur le marché, leurs mécanismes d'actions sont loin d'être autant diversifiés. On pourrait classifier les mécanismes en 4 catégories distinctes :

1. synthèse et fonctionnement de la membrane ;
2. synthèse des protéines ;
3. réplication et transcription de l'ADN ;
4. métabolisme de l'acide folique ;

1. Synthèse et fonctionnement de la membrane

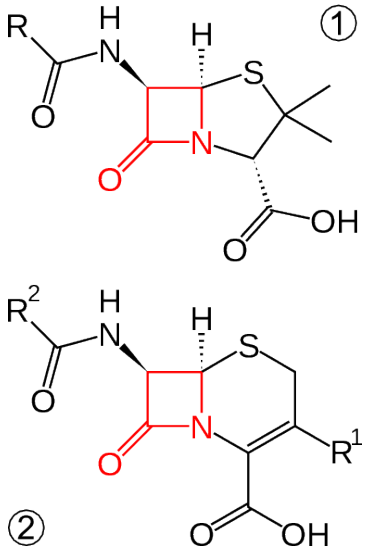


Figure 1.1 – Noyau β-lactame. 1-Penicilline. 2-Cephalosporine. Source : Wikipedia

Cette première catégorie comprend tous les antibiotiques de la classe des **β-lactamines** (voir figure annexe A.1) qui ont comme cible la « Penicillin-Binding-Protein » (PBP). La PBP est impliquée dans la synthèse du peptidoglycane - une composante essentielle à la structure de la paroi cellulaire pour la plupart des bactéries (Vollmer *et al.*, 2008). Elle agit entre autres en catalysant la réticulation (« *cross-linking* ») de deux chaînes de glycanes adjacentes nécessaire au maintien d'une bonne pression osmotique pour la cellule (Sauvage *et al.*, 2008). L'inhibition de la PBP par l'antibiotique devient alors létale pour la bactérie en provoquant la lyse de la cellule. La classe antibiotiques des β-lactamines est de loin celle qui comprend le plus de molécules différentes sur le marché, et qui possèdent tous un noyau β-lactame (rouge en figure 1.1). Ils sont globalement les antibiotiques les plus utilisés comme traitement chez l'humain (Leiros *et al.*, 2012). Selon les différents groupements chimiques qui entourent le noyau β-lactame, les β-lactamines sont généralement subdivisées en 5 catégories : les pénicillines, les céphalosporines, les monobactames, les carbapénems et finalement les inhibiteurs de β-lactamases. Il existe aussi des combinaisons de ces molécules, particulièrement une β-lactamine administrée avec un inhibiteur de β-lactamase pour contrer la résistance et préserver l'activité de l'antibiotique. Les 5 classes ont des spectres d'activités différents, ils ne sont donc pas tous actifs contre les mêmes bactéries et conséquemment n'ont pas la même utilisation en clinique.

Une autre classe d'intérêt qui vise la membrane cellulaire est celle des glycopeptides, principalement avec son antibiotique la **vancomycine**. Son mécanisme d'action est aussi relié à la synthèse du peptidoglycane, mais interfère plutôt directement avec celle-ci en se liant aux résidus terminaux de la chaîne de glycanes (D-alanine ou D-Ala) empêchant ainsi leurs réticulations par la PBP. La vancomycine est principalement active et utilisée contre les bactéries à Gram positif, comme les *Staphylococcus aureus* ou les *Clostridium difficile*.

D'autres classes d'intérêts cliniques qui visent la structure et le fonctionnement de la membrane sont les **polymyxines B** et E (**colistine**) ainsi que la **daptomycine**.

Les polymyxines sont des polypeptides qui interagissent avec la couche de lipopolysaccharide (LPS) de la membrane externe des bactéries à Gram négatif. La colistine par exemple, déplace les ions magnésium et calcium de la couche LPS et dérange ainsi son électrostatique ce qui cause une augmentation de la perméabilité de la membrane externe et la fuite de produits cellulaires jusqu'à provoquer la lyse de la cellule (Falagas et Kasiakou, 2005). Malgré la néphrotoxicité que les polymyxines peuvent causer chez un patient, elles ont récemment regagné en intérêt à cause de la multirésistance chez certaines bactéries à Gram négatif, principalement les *Acinetobacter baumannii*, *Pseudomonas aeruginosa* et *Klebsiella pneumoniae* (Zavascki *et al.*, 2007). Elles sont devenues les antibiotiques de tout dernier recours lors d'une infection par ces genres de bactéries.
La daptomycine, qui fut découverte dans les années 1980 (figure 0.1) et seulement approuvée en 2003 par la FDA [1], est maintenant commercialisée sous le nom de Cubicin par la compagnie pharmaceutique Cubist. C'est un lipopeptide cyclique actif contre une majorité des pathogènes à Gram positif. Elle agirait en insérant sa queue lipophilique à l'intérieur de la membrane, induisant sa dépolarisation et un important flux d'ions potassium vers l'extérieur de la cellule causant ensuite l'arrêt de la synthèse de l'ADN, de l'ARN et des protéines, et conséquemment la mort de la bactérie (Steenbergen *et al.*, 2005).

2. Synthèse des protéines

Une deuxième cible de choix pour un bon nombre d'antibiotiques est le ribosome - un complexe ribonucléoprotéique qui sert à la synthèse des protéines. Le ribosome traduit les ARN [2] messagers (ARNm) en séquence protéique ; il agit en décodant les codons [3] des ARNm en acides aminés et crée ainsi une chaîne peptidique qui se repliera dans une conformation tridimensionnelle pour former une protéine active. Le ribosome est formé d'au moins deux sous-unités : la plus grande nommée 50S (bleu figure 1.2) et la plus petite nommée 30S (vert figure 1.2) qui forment à leur deux le 70S [4]. La sous-unité 50S est composée des ARN ribosomaux (ARNr) 5S et 23S, alors que la sous-unité 30S comprend seulement l'ARNr 16S. Les deux sous-unités sont cependant aussi composées de plusieurs protéines qui aident à la formation et au maintien de leur structure. À noter aussi qu'on distingue trois sites de liaisons des ARN de transferts (ARNt) qui sont porteurs des

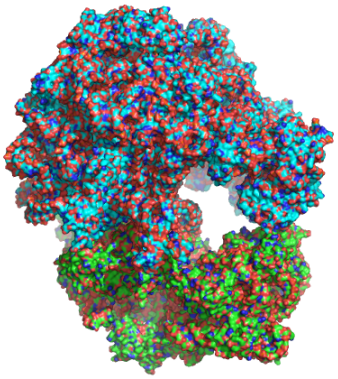


Figure 1.2 – Structures du ribosome de *Escherichia coli*. Bleu = sous-unité 50S (PDB : 3OFC). Vert = sous-unité 30S (PDB : 3OFA) (Dunkle *et al.*, 2010).

1. *Food and Drug Administration* des États-Unis
2. Acide ribonucléique
3. Codon : trois nucléotides subséquents qui codent pour un acide aminé.
4. Le S est un coefficient de sédimentation : vitesse à laquelle une particule sédimente avec une accélération appliquée.

acides aminés destinés à construire la chaîne polypeptidique :
le site-A où se fait la reconnaissance du bon aminoacyl-ARNt
avec l'anticodon, le site-P - ou peptidyl-ARNt - où se produit la formation de lien peptidique avec le prochain acide aminé et finalement le site-E où l'ARNt libre procède à sa sortie du ribosome.

Il existe plusieurs classes d'antibiotiques qui inhibent le fonctionnement de la sous-unité 50S du ribosome. Une première « métaclasse » sont les **MLSKO** qui regroupent les classes suivantes : les **macrolides**, **lincosamides**, **streptogramines**, **kelotides** et **oxazolidinones** (Roberts, 2008). Malgré leur différente structure, les quatre premières classes (MLSK) se lient à des sites actifs qui se chevauchent au site-P du ribosome (Wilson, 2009), et qui sont donc tous affectées par un même mécanisme de résistance qui visent la méthylation d'un ARN ribosomal par les gènes *erm* (voir 1.2.1). Les oxazolidinones ne sont cependant pas affectées par ce mécanisme de résistance, puisqu'elles se lient à un site différent sur le ribosome, plus près du site-A. Il existe tout de même au moins 2 mécanismes de résistances face à celles-ci connus à ce jour (voir 1.2.1). Les oxazolidinones agissent en prévenant la formation du complexe d'initiation formé du N-formylméthionine-ARNt, du ribosome et de l'ARNm (Swaney *et al.*, 1998). Une autre classe qui se lie au site-A du ribosome est celle des phénicols. Le principal antibiotique de cette classe a être utilisé en clinique est le **chloramphénicol**. Il existe aussi le florfénicol, mais ce dernier est plutôt utilisé en médecine vétérinaire. Les chloramphénicols, contrairement aux oxazolidinones, agissent plutôt comme inhibiteurs de l'élongation en bloquant le ribosome sur l'ARNm et en empêchant l'hydrolyse du peptidyl-ARNt (Wilson, 2009).

La sous-unité 30S, quant à elle, est inhibée par les deux classes d'antibiotiques suivantes : les **tétracyclines** et les **aminoglycosides**. Les antibiotiques de la classe des tétracyclines se lient à l'ARN ribosomal 16S et empêchent la liaison des aminoacyl-ARNt avec leur facteur d'élongation (Ef-Tu+GTP) au site-A et bloquent ainsi la synthèse protéique. Les aminoglycosides, quant à eux, se lient près du site de reconnaissance codon-anticodon et auraient la capacité d'induire un mauvais appareillement entraînant ainsi l'incorporation d'acide aminé non légitime dans la chaîne polypeptidique. Il en résulterait donc des protéines non fonctionnelles affectant évidemment le bon fonctionnement de la cellule.

Enfin, il existe deux antibiotiques qui visent plus directement les ARNt : la **puromycine** et la **mupirocine**. Le premier - la puromycine - agit de façon similaire à un ARNt en s'insérant dans le site-A du ribosome et termine ainsi abruptement la synthèse de la chaîne polypeptidique (Hong *et al.*, 2014). Le second - la mupirocine - inhibe l'enzyme iso-leucyl-ARNt synthétase (IleRS) empêchant ainsi le chargement des ARNt isoleucine. La bactérie perd donc sa capacité à ajouter l'acide aminé isoleucine durant la synthèse des protéines ce qui lui est fatale. La mupirocine est considérée comme un médicament essentiel dans la liste du World Health Organization (WHO) (2013) alors que la puromycine est plutôt utilisée comme un outil d'étude de la synthèse des protéines et non comme un antibiotique en clinique.

3. Réplication et transcription de l'ADN

Les principales cibles des antibiotiques liés à ce mécanisme cellulaire sont l'ADN gyrase et la topoisomérase IV. Ces enzymes rétablissent le surenroulement de l'ADN qui est créé lors de sa transcription et de sa réplication. En effet, elles sont capables de couper les deux brins d'ADN, de leur faire faire un supertour négatif et de les recoller par la suite. À noter que ce mécanisme est essentiel à la survie d'une cellule bactérienne.

Figure 1.3 – ADN Gyrase en complexe avec de l'ADN et l'antibiotique moxifloxaxin en rouge (PDB : 3FOE (Laponogov *et al.*, 2009)).

Deux classes d'antibiotiques agissent sur ces cibles, les quinolones (particulièrement les **fluoroquinolones**) et les **aminocoumarins**. La classe des fluoroquinolones est celle des deux qui est la plus utilisée en clinique avec la ciprofloxacine, la norfloxacine et la lévofloxacine (Collin *et al.*, 2011). Les fluoroquinolones se lient à une région bien précise sur la gyrase (*gyrA* et *gyrB*) et la topoisomérase IV (*parC* et *parE*) conventionnellement nommée QRDR (« Quinolone-Resistance-Determining-Region ») et qui est d'ailleurs impliquée dans la résistance (voir 1.2.1). Il a été démontré que les molécules interagissent avec l'ADN en plus des enzymes (voir figure 1.3) pour stabiliser le complexe de réaction de coupure des deux brins ce qui perturbe le reste de la réaction et provoque une accumulation du complexe dans la cellule (Aldred *et al.*, 2014).

Les aminocoumarins, comme la novobiocine, se lieraient encore plus fortement que les fluoroquinolones à la gyrase B en compétitionnant au site de l'ATP [5] - c'est l'hydrolyse de l'ATP qui fournit l'énergie au fonctionnement de la gyrase. Malgré sa forte affinité avec la gyrase, cette classe n'a pas connu le même succès en clinique que les quinolones, principalement à cause de leur piètre efficacité contre les bactéries à Gram négatif et leur cytotoxicité chez les mammifères (Collin *et al.*, 2011).

4. Métabolisme de l'acide folique

La synthèse de l'acide folique (folate ou vitamine B_9) est la cible de deux classes d'antibiotiques : les **triméthoprimes** et les **sulfonamides**. Chacune de ces classes inhibe une enzyme différente dans le sentier métabolique du folate (voir figure annexe A.2) - un précurseur essentiel à la synthèse des purines, acides nucléiques de l'ADN et de l'ARN. L'enzyme inhibée par les triméthoprimes est la dihydrofolate réductase (DHFR). Celle-ci apparaît tardivement dans le sentier métabolique et est responsable de la réduction du 5,6-dihydrofolate en 5,6,7,8-tetrahydrofolate et fait sa catalyse à l'aide du cofacteur NADPH [6] (Heaslet *et al.*, 2009). La

5. Adénosine triphosphate.
6. Forme réduite du *Nicotinamide Adenine Dinucleotide Phosphate* (NADP+)

deuxième cible qui est inhibée par les sulfonamides (ou sulfamides) est la dihydroptéroate synthase (DHPS), une enzyme qui intervient un peu avant la DHFR dans la synthèse du folate en créant une molécule intermédiaire, le 7,8-dihydroptéroate (Yun *et al.*, 2012).

Malgré le fait que chacune des deux classes ont au moins un représentant dans la liste des médicaments essentiels de l'OMS (World Health Organization (WHO), 2013), leur principale utilisation est une formule antibiotique qui combine les deux classes, le **triméthoprime/sulfamethoxazole** qui est d'ailleurs recommandée comme traitement dans le guide (The Johns Hopkins Hospital, 2014). La formule antibiotique a un large spectre antibactérien et s'avère efficace contre les pathogènes des infections urinaires (*Escherichia coli* ou autres entérobactéries), des infections respiratoires (*Streptococcus pneumoniae*, *Haemophilus influenzae*), des infections de la peau (*Staphylococcus aureus*) ou des infections entériques (*Escherichia coli* ou *Shigella*) (Huovinen *et al.*, 1995).

Perspectives pour les antibiotiques

En rétrospective, on s'aperçoit que les principaux antibiotiques utilisés en clinique n'opèrent pas sur une grande diversité de cibles. Il y aurait environ 200 protéines essentielles conservées chez les bactéries, mais les antibiotiques ayant eu le plus de succès viseraient principalement trois cibles : le ribosome, la synthèse de la paroi cellulaire et l'ADN gyrase/topoisomérase (Lewis, 2013). Ces cibles sont tous impliquées dans des fonctions cellulaires fondamentales et sont présentes chez une panoplie de bactéries commensales. La prise d'un antibiotique implique donc un certain bouleversement dans la diversité microbienne du microbiome d'un patient. Des impacts néfastes furent d'ailleurs constatés dans différentes études, comme l'augmentation de la susceptibilité à diverses pathogènes ou l'augmentation du risque pour plusieurs maladies telles la polyarthrite rhumatoïde, les maladies inflammatoires de l'intestin ou l'obésité (Keeney *et al.*, 2014). Une approche pour contrer ce genre de problème pourrait être d'augmenter la spécificité des antibiotiques en visant directement les bactéries pathogènes. Par exemple, des antibiotiques qui visent directement les facteurs de virulence des bactéries (Clatworthy *et al.*, 2007) ou même des thérapies phagiques (Reardon, 2014) sont des solutions ayant déjà été envisagées.

1.2 La résistance

Il va sans dire qu'une période postantibiotique aurait le potentiel de réduire de façon drastique notre espérance de vie, et une infection banale pourrait se développer en complication majeure. Différentes pratiques médicales, comme les chirurgies, deviendraient beaucoup plus à risque en l'absence de traitements antimicrobiens préventifs. Bref, il existe une multitude de raisons pour lesquelles la perte de notre ultime défense face aux micro-organismes - les antibiotiques - aurait des répercussions troublantes sur la santé humaine et pourrait vraisemblablement causer une régression de notre espérance de vie. La résistance aux antibiotiques, quoi que

très présente dans nos milieux cliniques, trouverait origine bien avant la commercialisation du premier antibiotique. En effet, un large éventail de gènes de résistances ont été identifiés dans des sédiments isolés du pergélisol en Béringie qui dateraient de plus de 30 000 ans (D'Costa *et al.*, 2011).

1.2.1 Mécanismes de résistance

Si les micro-organismes ont développé des outils, comme les antibiotiques, pour s'attaquer les uns les autres, ils ont bien sûr aussi développé des mécanismes pour se protéger de ces attaques. Suite à l'élaboration des différents modes d'action des antibiotiques dans la précédente section, procédons maintenant avec les mécanismes de résistances qui existent pour contrer notre arsenal anti-infectieux. Les différents mécanismes de résistances peuvent être regroupés en quatre grandes catégories :

1. désactivation de l'antibiotique ;
2. acheminement de l'antibiotique vers la cible ;
3. altération ou protection de la cible ;
4. acquisition de sentier métabolique secondaire.

1. Désactivation de l'antibiotique

La désactivation d'un antibiotique se réalise généralement par une enzyme qui vise directement la molécule antibiotique, soit pour la briser ou encore la modifier et ainsi annuler son bon fonctionnement.

Les **β-lactamases** (figure 1.4) sont un bel exemple d'enzyme qui brise les antibiotiques auxquels elle confère une résistance. En effet, l'enzyme réussit à ouvrir le noyau β-lactame (voir figure 1.1) par hydrolyse désactivant ainsi l'antibiotique puisque son pouvoir inhibiteur contre la PBP réside justement dans ce noyau. Il existe une grande diversité de β-lactamases et leur nomenclature est somme toute très élaborée. On peut voir un parallèle entre la diversité de leur nomenclature et la diversité des antibiotiques de type β-lactamine qui sont utilisés en clinique. Ils diffèrent habituellement dans leur affinité et leur spécificité face aux différentes β-lactamines. Ils existe d'ailleurs des β-lactamases à large spectre, nommées ESBL [7], qui peuvent hydrolyser les pénicillines, les céphalosporines de première, deuxième et troisième générations en plus de l'aztréoname (Paterson et Bonomo, 2005).

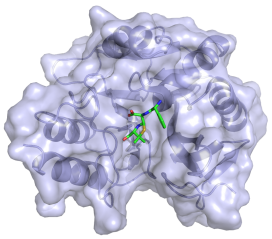


Figure 1.4 – Beta-lactamase NDM-1 avec ampicilline hydrolysé (PDB : 3Q6X (Zhang et Hao, 2011)).

7. *Extended Spectrum β-Lactamases*

On compte au moins deux nomenclatures connues pour ces gènes de résistances, soit celle de Ambler qui sont les classes A, B, C, D (Ambler, 1980) et celle de George A. Jacoby et Karen Bush (Bush et Jacoby, 2010), (Bush *et al.*, 1995). La première (Ambler) est basée principalement sur la divergence entre les séquences protéiques. Les classes A, C et D sont toutes des sérines β-lactamases et auraient un même ancêtre commun alors que les classes B sont plutôt des métallos-β-lactamases qui nécessitent la présence d'un ion métallique - de zinc - pour faire leur catalyse. À noter que la classe B est souvent subdivisée en trois sous-classes : B1, B2 et B3. Une légère modification de la classification de Ambler a été proposée par Hall et Barlow (2005), mais demeure très similaire à la notation originale de Ambler. La classification de Jacoby et Bush propose plutôt une classification basée sur le spectre d'activité des β-lactamases. Elle possède quatre groupes différents : 1- céphalosporinases (classe C de Ambler), 2- sérines β-lactamases (classes A et D de Ambler), 3- métallos-β-lactamases (classe B de Ambler). Différentes lettres ont aussi été désignées pour subdiviser les classes ; la deuxième classe qui est très diverse possède les lettres a (pénicillinase), b (« *broad-spectrum* »), c (carbénicillinase), d (oxacillinase), e (« *extended-spectrum* »), f (carbénicillinase) et r (résistant aux inhibiteurs de β-lactamases). Les lettres peuvent être conjuguées pour additionner les phénotypes et une β-lactamase TEM, par exemple, pourrait être soit de type 2b, 2be ou 2ber avec une différence de seulement quelques mutations ponctuelles entre elles.

Une autre catégorie de gènes de résistances assez élaborée et faisant partie de ce groupe de mécanisme sont les enzymes d'inactivation des **aminoglycosides**. Il en existe au moins trois types qui confèrent la résistance à cette classe, les AAC (« *Aminoglycoside Acetyltransferases* »), les ANT (« *Aminoglycoside Nucleotidyltransferases* ») et les APH (« *Aminoglycoside Phosphotransferases* ») (Shaw *et al.*, 1993). Il s'agit évidemment d'une classification basée sur le type de l'enzyme et de la réaction qu'elle catalyse sur la molécule antibiotique. Encore une fois, les trois types d'enzymes sont subdivisés en plusieurs catégories selon le groupement chimique visé pour la modification sur la molécule d'aminoglycoside. Par exemples, les AAC(3), AAC(6') et AAC(2') visent respectivement les groupements chimiques 3, 6' et 2' des aminoglycosides. À noter qu'il existe aussi un gène (*aac(6')-Ib-cr*) capable de conférer autant de la résistance aux aminoglycosides qu'à l'antibiotique ciprofloxacine - une **fluoroquinolone** (Robicsek *et al.*, 2006).

Les MLS, sous-classe des **MLSKO**, peuvent aussi être désactivés directement par certaines enzymes. Les macrolides sont entre autres la cible des phosphorylases (Mph) et estérases (Ere), les lincosamides par les transférases (Lnu), les streptogramines A par les transférases (Vat) et les streptogramines B par les lyases (Vgb) (Roberts, 2008), (Roberts *et al.*, 1999).

Les **chloramphénicols** sont principalement la cible des acétyltransférases de chloramphénicols (enzymes CAT) (Schwarz *et al.*, 2004), mais peuvent aussi être phosphorylés par une chloramphénicol 3'-O-phosphotransférase reportée dans Mosher *et al.* (1995).

Les **tétracyclines** pourraient aussi être la cible d'une enzyme (Tet(X)) qui désactive la molécule, mais celle-ci a été reportée dans quelques cas isolés et n'est pas le mécanisme de résistance le plus important face à cette classe d'antibiotique (Chopra et Roberts, 2001).

On s'aperçoit donc que la majorité des classes d'antibiotiques peuvent être victime d'une enzyme capable d'altérer leurs molécules et les rendre inactives. Ces enzymes de résistances sont aussi d'une grande importance clinique puisqu'une majorité d'entre elles se retrouvent sur des éléments mobiles ce qui mène à leur dissémination (van Hoek *et al.*, 2011). L'étude plus approfondie des mécanismes moléculaires de ces enzymes est d'intérêt et a déjà permis la conception de plusieurs antibiotiques synthétiques capables de les confronter, soit avec le design de nouvelles molécules qui évitent leurs sites actifs ou encore d'inhibiteurs pour les désamorcer (Wright, 2005).

2. Acheminement de l'antibiotique vers la cible

La présente section sur l'acheminement d'un antibiotique vers sa cible combine deux mécanismes de résistances, le premier englobe tous les changements ayant un impact sur la perméabilité cellulaire face à la molécule antibiotique et le deuxième vise l'exportation de la molécule hors de la cellule par les pompes d'efflux.

Un manque de perméabilité de la membrane face à un antibiotique bloque ou diminue l'importation de la molécule à l'intérieur de la cellule bactérienne et conséquemment son acheminement vers la cible. Le phénomène a été observé pour plusieurs classes d'antibiotiques avec une évidence particulière pour les **β-lactamines**, les **chloramphénicols** et les **fluoroquinolones**. Chez les *Pseudomonas* - des bactéries à Gram négatif - la porine OprD (figure 3.1) de la membrane externe qui est normalement liée à l'importation de l'acide aminé arginine contribuait à la résistance à certaines carbapénems lorsque non fonctionnelle - du à des mutations délétères ou autres (Huang et Hancock, 1996), (Kos *et al.*, 2014). Le même phénomène fut aussi proposé comme mécanisme de résistance aux chloramphénicols chez *Haemophilus influenzae* (Burns *et al.*, 1985), *Pseudomonas cepacia* (Burns *et al.*, 1989) et *Salmonella typhi* avec l'absence de leur porine OmpF (Toro *et al.*, 1990). La même porine OmpF, chez *Escherichia coli*, serait l'une des portes d'entrée aux quinolones, aux tétracyclines et à céfoxitine (une β-lactamine) et sa défaillance diminuerait la sensibilité à ces antibiotiques (Bryan et Bedard, 1991).

Les pompes d'efflux qui acheminent les molécules antibiotiques hors de la cellule couvrent encore un plus grand nombre de classes d'antibiotiques (**β-lactamines**, **aminoglycosides**, **tétracyclines**, **chloramphénicols**, **MLSKO**, **quinolones**) et plusieurs pompes confèrent même la résistance à plus d'une classe. Les systèmes d'efflux sont généralement classés en cinq catégories distinctes, dont les quatres premières utilisent principalement la force protomotrice (dépolarisation de la membrane) pour l'extrusion des molécules antibiotiques alors que la dernière utilise plutôt l'ATP comme source d'énergie (Putman *et al.*, 2000). Le lecteur

est prié de se référer à l'article de revue Poole (2005) pour un listage plus détaillé des différents groupes de systèmes d'efflux et de leurs implications dans la résistance aux antibiotiques. Voici quand même un bref sommaire de chacune des familles.

1. **MFS** « *Major Facilitor Superfamily* » ou superfamille de facilitateurs majeurs.
   Les transporteurs MFS se retrouvent autant chez les bactéries, archéobactéries ou eucaryotes et sont associés à plusieurs fonctions cellulaires essentielles. Elles sont d'ailleurs présentes chez beaucoup de bactéries et induisent une résistance à divers antibiotiques. Il en existe une multitude chez les Gram positifs (staphylocoques, streptocoques, entérocoques, bacilles et lactobacilles) qui provoquent une résistance aux fluoroquinolones (NorAB, PmrA, EmeA, Bmr, Lde). Chez les Gram négatifs, E. coli possède le système MdfA aussi caractérisé pour conférer une multirésistance à divers degré face aux chloramphénicols, macrolides, tétracyclines, aminoglycosides et fluoroquinolones. De plus, un bon nombre de gènes de résistance *tet* (voir Chopra et Roberts (2001)) sont des MFS qui protègent contre les tétracyclines, et cela autant chez les Gram positifs que les Gram négatifs. Enfin, il existe d'autres systèmes MFS qui furent attribués à la résistance aux chloramphénicols (Cml, Flo) et à différentes MLS (Mef(A), MdeA, LmrB, Cme).

2. **SMR** « *Small Multidrug Resistance family* ».
   Les transporteurs SMR sont les plus petites pompes d'efflux qui existent, avec une longueur de séquence d'une centaine d'acides aminés seulement, et seraient fort probablement fonctionnels sous forme de complexe homodimère (Chen *et al.*, 2007). Chez E. coli, le transporteur EmrE serait impliqué dans la résistance aux tétracyclines et légèrement aux aminoglycosides. Les gènes QAC sont aussi des transporteurs SMR et visent l'extrusion des composés d'ammonium quaternaire, lesquels sont plutôt utilisés comme désinfectants. À noter que *qacE*, surtout la version tronquer du gène (*qacE*Δ*1*), est souvent associé aux intégrons de type I répandus chez les bactéries à Gram négatif. Le système Mmr de *Mycobacterium tuberculosis* serait, quant à lui, lié à une faible résistance face aux fluoroquinolones. Il existe aussi des transporteurs SMR formés de deux gènes distincts, tels les opérons EbrAB, YkkCD et YvdRS trouvés chez *Bacillus subtilis*, et ceux-ci formeraient plutôt des complexes hétérodimères (Masaoka *et al.*, 2000).

3. **RND** « *Resistance-Nodulation-Cell Division family* ».
   Les transporteurs de type RND se retrouvent dans tout le règne de la vie, mais sont surtout présents chez les Gram négatifs en ce qui concerne les bactéries. Chez ces dernières, cela implique une composante qui traverse la membrane interne, une autre qui traverse la membrane externe et une troisième qui intervient dans l'espace périplasmique. Il en existe une multitude qui confère la résistance à un ou plusieurs antibiotiques comme les β-lactamines, les chloramphénicols, les macrolides, les tétracyclines, les aminoglycosides et les oxazolidinones. Seulement chez *Pseudomonas aeruginosa* au moins quatres systèmes de cette famille ont été caractérisés : MexAB-OprM, MexCD-OprJ, MexEF-OprN

et MexXY-OprM (Aeschlimann, 2003). Deux autres systèmes RND multirésistants ont aussi été caractérisés chez *Escherichia coli* : AcrAB-TolC et AcrEF-TolC. Il est intéressant de remarquer que la même protéine transmembranaire située dans la membrane externe (TolC et OprM) peut être réutilisée dans deux complexes RND différents.

4. **MATE** « *Multidrug And Toxic Compound Extrusion family* ».
   Cette famille fut la dernière des cinq à avoir été proposée et caractérisée par Brown *et al.* (1999). Il se basa sur la divergence des séquences avec les autres familles de pompes d'efflux pour justifier sa création. Des représentants de la famille apparaîtraient aussi dans chacun des règnes de la vie. Deux protéines d'intérêts dans cette famille sont NorM de *Vibrio parahaemolyticus* et YdhE de *Escherichia coli* auxquelles on associe la résistance à plusieurs antibiotiques des classes fluoroquinolones et aminoglycosides.

5. **ABC** « *ATP-Binding Cassette* ».
   La famille de transporteurs ABC se distingue des autres familles de par la source d'énergie qu'elle utilise pour permettre l'extrusion des antibiotiques, soit l'hydrolyse de l'ATP. Ces transporteurs sont généralement composés de quatre domaines : deux domaines membranaires et deux domaines de liaison nucléotidique (Chang, 2003). Génétiquement parlant le système peut apparaître sous plusieurs formes, soit avec des gènes séparés pour les différents domaines ou encore avec une protéine unique qui peut former un complexe homodimère. Une majorité des transporteurs ABC confèrent une résistance plutôt spécifique à certains antibiotiques : les Msr(A), Msr(C) et Msr(D) pourraient exporter certains antibiotiques des MLSK, Vga(A/B) viseraient les streptogramines de type A, Lsa(B) la clindamycine, etc. Il existe tout de même le système LmrA, chez *Lactococcus lactis*, qui augmenterait la résistance à une multitude de classes d'antibiotiques, comme les aminoglycosides, MLS, quinolones et tétracyclines (Poelarends *et al.*, 2002).

3. Altération et protection de la cible

Ce mécanisme de résistance est certes celui qui touche le plus de classes antibiotiques, sinon la totalité. Il s'agit souvent de mutations ponctuelles dans les gènes ciblés par l'antibiotique, mais on compte aussi des gènes de résistance capables de protéger la cible en la modifiant et qui peuvent se retrouver sur des éléments mobiles. Divisons donc la présente section par classe d'antibiotique, par souci de clarté.

Les **β-lactamines** ont comme cible la PBP, il est prévisible que des mutations dans celle-ci affectent son affinité avec les antibiotiques de cette classe. Un bel exemple est les gènes *mecA* chez les staphylocoques qui sont en fait des variants de la PBP (PBP2a ou PBP2') et ont la faculté d'être transmise par un élément mobile, une cassette chromosomique recombinase (*ccr*) qui forme l'élément SCC*mec* (« *Staphylococcal Cassette Chromosome mec* ») (IWG-SCC, 2009). Chez les pneumocoques une faible résistance fut constatée avec les variants PBP2x et PBP2b de l'enzyme et une plus forte résistance avec la PBP1a mutée. Les entérocoques, eux,

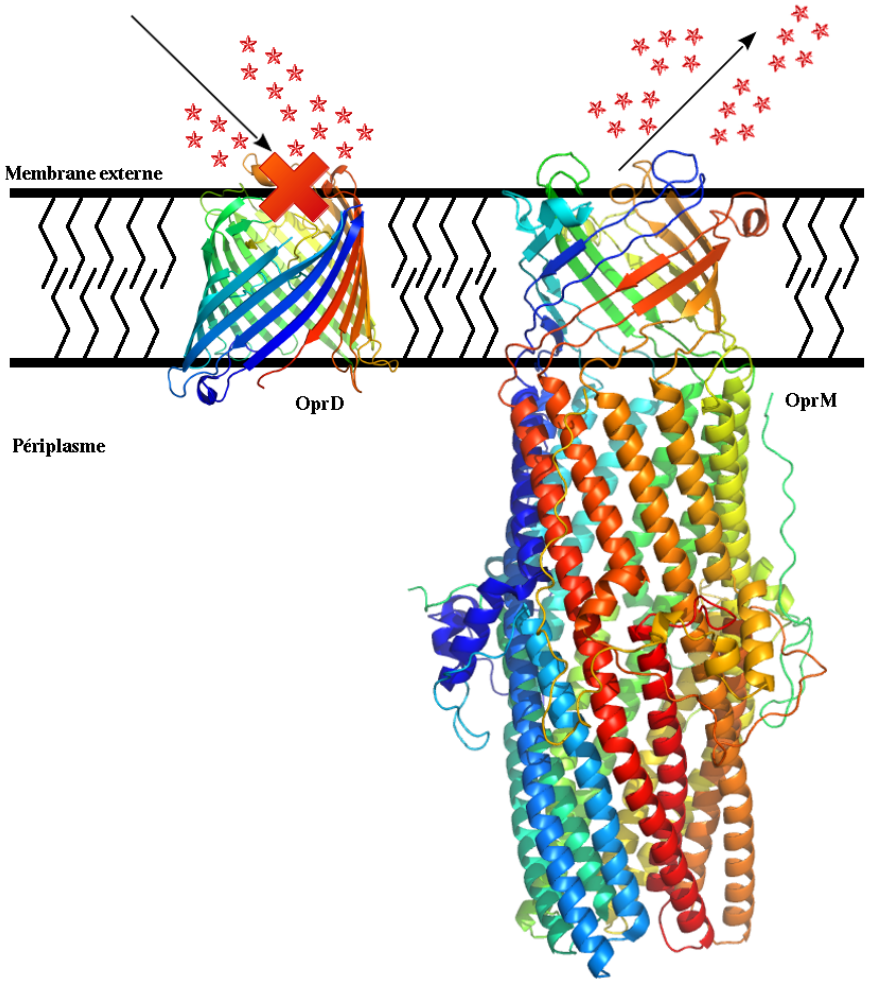


Figure 1.5 – Illustration de la membrane externe des bactéries à Gram négatif avec la porine OprD (PDB : 3SY7 (Eren *et al.*, 2012)) et la porine à efflux OprM (PDB : 1WP1 (Akama *et al.*, 2004)) de *Pseudomonas aeruginosa*.

peuvent avoir la PBP5fm qui lorsque fortement exprimée confère un haut niveau de résistance à la pénicilline et ses analogues (Macheboeuf *et al.*, 2006).

La cible des **aminoglycosides**, l'ARNr 16S de la sous-unité 30S du ribosome, peut être victime de modifications post-traductionnelles par les méthyltransférases de l'ARNr 16S (ArmA, RmtA-D) (Doi et Arakawa, 2007). Ces méthyltransférases ont tous été trouvées sur des éléments mobiles (transposons) impliqués dans leur propagation d'où leur importance clinique. Des mutations ponctuelles peuvent aussi affecter la sensibilité de l'ARNr 16S à cette classe, avec entre autres les positions 1406A et 1408G qui sont associées à une forte résistance aux aminoglycosides 4,6-désubstitués, telles la kanamycine et la gentamicine (Recht et Puglisi, 2001).

Les **tétracyclines**, comme les aminoglycosides, ciblent l'ARNr 16S et sont aussi victimes d'enzymes qui préviennent leur liaison à la cible. Au moins sept gènes *tet* (et *otr*(A)) ont été identifiés pour agir comme protection ribosomale, les autres gènes *tet* étant des pompes d'efflux ou des enzymes qui ciblent la molécule directement (Chopra et Roberts, 2001). Pour ce qui est des mutations ponctuelles, la position 1058 de l'ARNr 16S lorsque mutée d'une cytosine pour une guanine conférerait de la résistance à cette classe, chez *Escherichia coli* (Ross *et al.*, 1998).

Les mutations reconnues pour conférer une résistance aux **chloramphénicols** sont rares et ne sont pas, a priori, la première cause de résistance à cette classe (Schwarz *et al.*, 2004). L'une des raisons est que les changements structuraux au site-P, où se lie l'antibiotique, ont un trop grand impact sur le fonctionnement du ribosome et ne seraient donc pas viable pour la bactérie (Fischetti et for Microbiology, 2006). Il existe néanmoins une mutation dans l'ARNr 23S chez *Escherichia coli*, une guanine pour une adénine à la position 2057, qui fut caractérisée pour la résistance aux chloramphénicols et à l'érythromycine (Ettayebi *et al.*, 1985).

Les ARNr méthyltransférases forment le plus grand groupe de gènes de résistances acquis qui confère la résistance aux **MLSKO**. Ils agissent en ajoutant un ou deux groupements méthyles sur l'ARNr 23S, plus spécifiquement à la position 2058 chez *Escherichia coli* (Roberts, 2008). Les principaux gènes de ce mécanisme sont les *erm* et leur nomenclature est maintenue par Marilyn C. Roberts à l'Université de Washington (http ://faculty.washington.edu/marilynr/). En plus des mutations connues dans l'ARNr 23S, deux protéines du ribosome (L4 et L22) lorsque mutées sont aussi responsables de la résistance à tous les antibiotiques de la métaclasse des MLSKO.

Le plus souvent, la résistance aux **quinolones** est dues à des mutations dans la région QRDR sur les cibles de l'antibiotique, soit la gyrase (les gènes *gyrA* et *gyrB*) et la topoisomérase IV (les gènes *parC* et *parE*) (Aldred *et al.*, 2014). Un autre mécanisme de protection de la cible apparaît avec les gènes *qnr* et leur nomenclature est maintenue par le groupe de la Lahey Clinic (http ://www.lahey.org/qnrstudies/) (Jacoby *et al.*, 2008). Les gènes *qnr* sont de la famille des protéines « *pentapeptides repeat* » et agiraient en se liant à la gyrase et la topoisomérase IV. Ils réduiraient le nombre de complexes gyrase-ADN créés et accessibles à l'antibiotique (les quinolones) et entraveraient aussi à la bonne fixation de l'antibiotique au site actif de ce même complexe gyrase-ADN (Tran *et al.*, 2005).

Les antibiotiques **triméthoprimes** visent la DHFR et des mutations dans cette dernière ou une surexpression de celle-ci provoque la résistance. À ce jour, au moins 20 variants résistants du gène *dhfr* ont été caractérisés sur des éléments mobiles leur donnant donc la possibilité d'être mobilisés et disséminés (Huovinen, 2001). Les **sulfonamides** visent une deuxième enzyme dans le métabolisme du folate, la DHPS. Le gène *folP* qui code pour cet enzyme a été attribué à de la résistance chez *Neisseria meningitidis* lorsque muté (Fiebelkorn *et al.*, 2005).

4. Acquisition de sentier métabolique secondaire

Le meilleur exemple d'acquisition de sentier métabolique secondaire est celui des opérons de gènes de résistances *van* qui modifient les terminaisons D-Ala D-Ala des précurseurs du peptidoglycane, soit par une chaîne D-Ala D-Lac (avec les opérons VanA, VanB et VanD) ou encore une autre D-Ala D-Ser (VanC, VanE et VanG) (van Hoek *et al.*, 2011). Seuls les opérons VanA et VanB ont été retrouvés sur des plasmides alors que les autres furent plutôt identifiés comme chromosomiques. Le type VanB se retrouve entre autres sur un transposon conjugatif (Tn*1549*) qui a mené à sa dissémination chez les *Enterococcus* (Garnier *et al.*, 2000). Le type VanA se retrouve aussi sur un transposon conjugatif (Tn*1546*) normalement présent chez les entérocoques (VRE), mais sa présence sur un plasmide (pLW1043) fut aussi caractérisée chez un *Staphylococcus aureus* résistant à la vancomycine (VRSA) (Weigel *et al.*, 2003). Un opéron de type VanG est aussi présent chez une multitude de *Clostridium difficile*, mais malgré son expression, les résidus terminaux des chaînes du peptidoglycane (D-Ala D-Ala) ne seraient pas affectés et les souches demeureraient sensibles à la vancomycine (Peltier *et al.*, 2013).

1.2.2 Dissémination de la résistance

Le transfert horizontal de gènes (THG) est bien connu pour être un facteur très important qui contribue à la propagation des gènes de résistances aux antibiotiques, mais aussi des gènes de virulences chez les bactéries pathogènes (Juhas, 2013). Il existe au moins trois mécanismes connus qui sont capables de faire l'échange d'ADN d'un microbe à un autre : la transformation, la conjugaison et la transduction (Frost *et al.*, 2005). La transformation est le processus de relargage d'ADN dans un milieu et de sa capture par une autre cellule, qui l'intégrera ensuite dans son chromosome pour qu'il devienne fonctionnel. Un tel évènement ne peut cependant pas être prévu par de simples analyses génomiques et bio-informatiques. Un deuxième mécanisme est la transduction qui est la conséquence d'une intégration de l'ADN d'une bactérie hôte dans un phage (virus de bactérie) qui pourra ensuite être injecté dans une seconde bactérie. Finalement, la conjugaison est le transfert direct de matériels génétiques d'une bactérie à une autre via un système conjugatif qui sert à créer un pont formé généralement de pili et nécessaire au transfert. Les plasmides conjugatifs font parti des éléments capables d'opérer par conjugaison en plus des éléments conjugatifs et intégratifs (ICE). Les plasmides et les ICE sont très certainement les éléments mobiles et transférables ayant été les plus étudiés pour leur impact sur la résistance aux antibiotiques.

Les éléments génétiques mobiles (MGE) sont habituellement caractérisés une fois présents chez une bactérie avec l'aide de la génomique. Dans un premier temps, les intégrons sont des éléments capables d'intégrer, à partir d'une intégrase, à leur site attI un gène, qu'on nomme cassette, et qui possède un site attC. La recombinaison entre les deux sites provoque l'intégration de la cassette au sein de l'intégron. Les intégrons possèdent souvent une série de cassettes (gènes) de résistances qui sont rangées l'une à la suite de l'autre. Ces cassettes

sont toutes exprimées à partir d'un seul promoteur. L'intégrase d'intégron qui est responsable de l'insertion des cassettes serait exprimée avec la réponse SOS, réponse face à un stress comme la présence d'un antibiotique dans la cellule (Da Re et Ploy, 2012). Elle pourrait même réarranger l'ordre des cassettes en les excisant et les réintégrant au début de l'intégron, plus près du promoteur, ce qui augmenterait donc leur expression. Les intégrons peuvent aussi être mobilisés en entier lorsqu'ils sont situés sur un transposon fonctionnel. Les transposons nécessitent au moins une transposase et parfois aussi une résolvase pour permettre leur excision et leur intégration à un site donné sur l'ADN qui peut être soit spécifique ou multiple tout dépendamment du type de transposon. Il existe en effet différents types de transposons avec différents mécanismes de transposition et leur nomenclature, depuis Tn*6000*, est maintenue par le Dr Adam Roberts au University College London (UCL) (Roberts *et al.*, 2008). Les séquences d'insertions (IS) sont les plus petits éléments transposables puisqu'elles peuvent agir seules et se mobiliser de manière indépendante. Elles peuvent aussi former des transposons composés en bornant le transposon aux deux extrémités. Les transposons composés peuvent contenir d'autres gènes entre les IS, comme ceux de résistance aux antibiotiques. Un autre mécanisme de résistance relié aux IS est lorsqu'elles s'insèrent en amont d'un gène de résistance et renforcent une partie de son promoteur (la région -35) ce qui augmente son expression (Allmansberger *et al.* (1985), Goussard *et al.* (1991)) . Une classification des diverses familles d'IS a été introduite dans l'étude de Siguier *et al.* (2006b) et sera revue au chapitre 3. D'autres parts, les plasmides et prophages, lorsqu'ils sont intégrés au sein du chromosome, deviennent des îlots génomiques et font alors partie du génome accessoire de la bactérie hôte. On peut les distinguer par génomique comparative avec des génomes d'une même espèce. Par exemple, chez *Pseudomonas aeruginosa* une nomenclature de ces îlots, nommés régions de plasticités du génome (RGP), fut réalisée dans Mathee *et al.* (2008) et utilisée comme référence dans le projet de la souche PA96 au chapitre 4.2.2.

Un exemple typique qui combine les éléments tout juste présentés serait la présence d'un intégron, avec plusieurs gènes de résistances, sur un transposon qui est situé sur un plasmide. C'est le cas du plasmide pOZ176 qui sera présenté au chapitre 4.2.1. Si en plus le plasmide est conjugatif, alors il pourra être transmis à une autre bactérie et où le transposon pourra s'intégrer au chromosome ou sur un autre plasmide, etc. Quoi qu'il en soit, la compréhension des mécanismes qui mènent à la dissémination des gènes de résistances aux antibiotiques est essentielle pour espérer un jour garder un certain contrôle sur l'éclosion de bactéries multirésistantes.

1.2.3 Combattre la résistance

Pour chiffrer l'importance de la résistance aux antibiotiques, il a été reporté qu'aux États-Unis seulement il y aurait plus de 2 millions d'infections sur une base annuelle causées par une bactérie résistante à l'antibiotique conseillé comme premier traitement (CDC, 2013). Sur ces 2 millions d'infections annuelles, au moins 23000 causeraient la mort du patient suite à l'échec

du traitement antibiotique dû à la résistance. Les coûts engendrés, autant à la santé publique qu'à la société en général, se chiffrent en milliards de dollars et pourraient être atténués avec un meilleur contrôle préventif des infections par des bactéries résistantes aux traitements.

Le rapport du CDC met aussi en garde contre certains groupes de bactéries résistantes qui sont globalement un danger à la santé publique. Il les classifie en trois échelons d'importances selon le niveau de préoccupation associé aux dites bactéries. Voici un résumé en ordre croissant d'importance :

— Inquiétant : *Staphylococcus aureus* résistant à la vancomycin (VRSA), streptocoques du groupe A résistant à l'érythromycine et du groupe B résistant à la clindamycine ;

— Sérieux : *Pseudomonas aeruginosa* et *Acinetobacter* multirésistants, *Staphylococcus aureus* résistant à la méthicilline (MRSA), entérocoques résistant à la vancomycine (VRE), entérocoques avec une ESBL et plusieurs autres microbes résistants comme les *Campylobacter*, *Salmonella*, *Shigella*, *Streptococcus pneumoniae*, *Mycobacterium tuberculosis* et *Candida* (un fungi).

— Urgent : *Clostridium difficile*, entérobactéries résistantes aux carbapénems (CRE), *Neisseria gonorrhoeae* multirésistant.

De nos jours, la menace engendrée par la résistance aux antibiotiques est bien connue et beaucoup d'intervenants tentent d'apporter différentes solutions à ce fléau. À titre d'exemple, le premier ministre de la Grande-Bretagne, David Cameron, fit une apparition dans les médias pour faire part de l'importance alarmante de la résistance aux antibiotiques. Il affirma qu'il fallait agir maintenant pour contrer le problème, puisqu'une période postantibiotique nous retournerait dans « l'âge noir » de la médecine où une infection banale pouvait s'avérer mortelle (Walsh, 2014). Suite à ces propos, le sujet de la résistance aux antibiotiques se valut le prix Longitude 2014 (http://www.longitudeprize.org/challenge/antibiotics), pour une somme de 10 millions de livres sterling attribuable d'ici 2019 à l'équipe qui remplira la première tous les critères du concours pour investiguer d'éventuelles solutions au problème.

Une autre action importante fut celle du président américain, Barack Obama, qui émit un décret présidentiel pour combattre les bactéries résistantes aux antibiotiques (Obama, 2014). En effet, cet ordre exécutif propose la mise en place de mesures organisées et coordonnées par trois ministères (défense, agriculture, santé et services sociaux), pour assurer la surveillance, le contrôle et la prévention des bactéries résistantes. Plusieurs aspects intéressants ressortent du décret, mais plus particulièrement pour le sujet à l'étude la section 6 qui mentionne entre autres les bases de données génomiques comme une technologie centrale aux efforts de surveillance nationale.

Chapitre 2

Base de données en génomique

2.1 Génomique

La génomique a pleinement pris son essor avec l'arrivée du séquençage de nouvelle génération. Ces nouvelles générations sont ainsi nommées puisqu'elles donnent suite à la première génération de séquençage basé sur la méthode de Sanger qui était devenue trop chère et fastidieuse pour la demande (Metzker, 2005). L'amélioration des technologies de séquençage à donc fait baisser énormément les coûts pour séquencer un génome, passant de plus de 5000$ pour un million de paires de bases en 2001 à 0.05$ en 2014 (Wetterstrand, 2014). Il est donc devenu très abordable de séquencer des génomes bactériens, avec des coûts qui se chiffrent dans les centaines de dollars, ou même des communautés entières de microbes (microbiomes) et cela dans des temps raisonnables. Cette accessibilité croissante au séquençage a entraîné un déluge de données génomiques dans les bases de données publiques, tel GenBank du NCBI. En effet, la publication d'un article scientifique concernant le génome d'un organisme vient avec la responsabilité de déposer celui-ci dans une base de données publique pour la vérification des affirmations en lien avec le génome. Le principe UPSIDE (« *Uniform Principle for Sharing Integral Data and materials Expeditiously* ») que promeut le NCBI décrit bien cette responsabilité, soit que toutes données essentielles à un article scientifique doivent être rendues disponibles pour permettre la vérification et la réplication par les pairs (Cozzarelli, 2004). C'est d'ailleurs le cas pour les séquences et les annotations de génomes bactériens qui sont soumises à une base de données publique pour supporter les informations des articles associées. Les projets sur *Pseudomonas aeruginosa* PA96 et son plasmide pOZ176, au chapitre 4.2, sont de bons exemples de soumissions de fichiers d'annotations GenBank nécessaires à la publication de leur article respectif.

À noter que pour rester en lien avec le sujet de cette dissertation, les éléments de génomiques présentés concernent plutôt la génomique microbienne et non la génomique humaine (où des eucaryotes supérieurs). Ce n'est pas non plus une représentation exhaustive de toutes les techniques existantes, mais plutôt une introduction aux approches réalisées durant le projet.

L'une des premières analyses bio-informatiques réalisées suite à l'obtention des séquences d'ADN (lectures), depuis un séquenceur à haut débit, est l'assemblage (*de novo* ou avec référence) de ces lectures en *contigs*[1]. Les séquenceurs de type Illumina - qui dominent le marché des séquenceurs (McPherson, 2014) - créent de courtes lectures appariées (100-150 paires de bases) qui ne se chevauchent pas et ont un espacement variable entre elles lorsqu'alignées sur l'ADN original. Tous les projets connexes du chapitre 4 ont utilisé des séquenceurs de type Illumina, les autres types de séquenceurs ne seront donc pas couverts au courant de ce chapitre. Les algorithmes de choix pour assembler les lectures appariées en *contigs*, se basent sur le graphe de Bruijn, puisque la méthode permet de mieux découper et paralléliser le problème comparativement aux algorithmes qui se basent uniquement sur l'alignement et le chevauchement de séquences (Illumina, 2010). La méthode utilisée avec le graphe de Bruijn consiste à découper les lectures d'ADN en plus courts segments nommés K-mers. Chaque K-mer devient alors un sommet dans le graphe et une arête est dirigée d'un sommet vers un autre lorsque leur k-mer se chevauche de N-1 nucléotides.

Il existe plusieurs assembleurs qui utilisent le graphe de Bruijn : Velvet (Zerbino et Birney, 2008), SOAPdeNovo (Li *et al.*, 2009), Abyss (Simpson *et al.*, 2009), SPAdes (Bankevich *et al.*, 2012) et Ray (Boisvert *et al.*, 2010). Ray permet aussi de faire du profilage à base de K-mers avec son extension RayMeta (Boisvert *et al.*, 2012). Le profilage (ou coloriage) sert à identifier le contenu en K-mers du graphe de Bruijn d'un génome ou d'un métagénome en se basant sur des ensembles de données de séquences qu'on lui donne en entrée. Pour un métagénome de microbes, par exemple, cela est très intéressant puisque RayMeta permet de quantifier l'abondance en K-mers des différents taxons et niveaux taxonomiques bactériens - famille, genre, espèce, etc. Puisque le coloriage ne tolère pas les non-identités (« *mismatchs* ») entre K-mers, comme le ferait un alignement, il faut donc avoir un ensemble de données représentatif de toute la diversité des gènes qu'on veut quantifier ou identifier dans l'échantillon. C'est d'ailleurs l'un des aspects ayant été pris en compte dans la fabrication de la base de données MERGEM, qui sera présentée au chapitre 3. L'étape de l'assemblage s'avère aussi nécessaire pour un projet d'annotations. Avec un assemblage, on peut ensuite faire l'identification des cadres de lectures ouverts (CLO) qui indiquent la présence potentielle de gènes dans les *contigs* assemblés. Cette étape peut aussi se réaliser *de novo*, c'est-à-dire qu'aucun ensemble de gènes n'est utilisé pour leur identification. Habituellement, les identificateurs de gènes se basent principalement sur l'utilisation préférentielle de codon, comme GeneMark (Besemer et Borodovsky, 2005), Glimmer (Delcher *et al.*, 1999) et Prodigal (Hyatt *et al.*, 2010). Prodigal a été conçu spécifiquement pour les génomes bactériens et son algorithme se sert en plus des séquences Shine-Dalgarno[2] ce qui augmente sa précision par rapport à ces prédécesseurs. Par la suite, les produits protéiques des gènes identifiés *de novo* sont alignés avec des ensembles de données

1. Un contig est une suite contiguë d'une région d'ADN recréée à partir de plus courtes lectures d'ADN qui se chevauchent.
2. Les séquences de Shine-Dalgarno sont des séquences consensus de liaison du ribosome chez les bactéries.

de protéines, les plus fiables possible, pour en ressortir des annotations fonctionnelles précises et conformes. L'avantage de travailler avec les protéines versus les gènes est la précision des identifications tout en évitant les mutations synonymes dans l'ADN codant. Les noms des gènes et des produits protéiques identifiés avec les alignements sont alors reportés dans des fichiers d'annotations qui pourront être soumis à une base de données publique par la suite. Un pipeline d'annotations a d'ailleurs été développé dans le cadre du projet CQDM et vous sera présenté en début de chapitre 4. Peu importe la technique utilisée pour identifier le contenu génomique d'un échantillon, une chose est sûr c'est qu'elle dépend grandement de la qualité de la BD qui sert à l'identification des gènes.

2.2 Bases de données biologiques

Lorsqu'on parle des bases de données publiques on fait référence à l'une des trois qui fait partie de l'INSDC [3] : la DDBJ [4], le EMBL-EBI [5] et le NCBI [6]. En effet, les trois institutions collaborent pour synchroniser toutes leurs données sur les séquences génomiques qui leur sont soumises. Donc peu importe la source utilisée, les données seront les mêmes pour un même génome, lorsqu'extrait de l'une de ces BD. Cependant, tel que reporté dans Schnoes *et al.* (2009), les annotations fonctionnelles dans ces BD publiques d'envergures sont souvent de piètres qualités. Pour 10 des 37 familles d'enzymes testées dans l'article, le taux d'erreurs d'annotations dépassait les 80% dans l'une ou plusieurs des bases de données publiques suivantes : NCBI GenBank, UniprotKB/SwissProt, UniprotKB/TrEMBL ou KEGG. La BD avec le plus faible taux d'erreurs était UniprotKB/SwissProt et se rapprochait de 0% pour la plupart des familles d'enzymes. La BD SwissProt est l'un des projets ayant eu le plus d'ampleur sur l'édition de contenu manuelle des annotations de protéines (Bairoch et Apweiler, 1999) (Bairoch et Apweiler, 2000) (Boeckmann *et al.*, 2003). En effet, un important effort fut réalisé par le SIB [7] pour extraire manuellement les connaissances fonctionnelles de milliers de protéines dans les articles scientifiques leur étant associés. SwissProt est maintenant affilié à la base de données Uniprot et fait partie du UniprotKB (« *Uniprot KnowledgeBase* ») en compagnie de TrEMBL, une base de données de traduction de tous les gènes de EMBL en séquences protéiques.

Une pratique courante dans les pipelines d'annotations en bio-informatique est de transposer automatiquement l'annotation d'une séquence homologue identifiée à partir d'alignements de séquences. Beaucoup d'erreurs d'annotations se sont propagées avec le temps à cause de cette façon de procéder. Si une erreur d'annotation est introduite, elle risque d'être répliquée rapidement dans d'autres génomes et, qui plus est, à force de copier l'annotation de séquences homologues on finit par perdre l'homologie et diverger de la séquence originale. La nomen-

3. *International Nucleotide Sequence Database Collaboration*
4. *DNA DataBank of Japan*
5. *European Molecular and Biological Laboratory - European Bioinformatics Institute*
6. *National Center for Biotechnology Information*
7. *Swiss Institute of Bioinformatics*

clature des gènes est très importante puisqu'elle établit un langage avec lequel les biologistes peuvent décrire le contenu génomique d'un organisme. Une nomenclature élaborée permet aussi de différencier le phénotype des variants d'un même gène type. À titre d'exemple, les β-lactamases TEM ont plus de 200 variants déjà assignés selon leur séquence protéique. Leur nomenclature élargie est justifiée puisque ces enzymes n'ont pas tous le même spectre d'activité, ils n'hydrolysent pas exactement les mêmes antibiotiques avec la même efficacité et une simple mutation ponctuelle permet à changer le spectre (cf. classification de Jacoby et Bush au chapitre 1). Il existe donc des chercheurs de renom qui se portent volontaires pour faire le suivi de la nomenclature de certains gènes ou systèmes d'intérêts. C'est assurément une pratique viable, à condition que la majorité des chercheurs qui caractérisent un nouveau variant s'assurent de passer par ces responsables pour se faire attribuer un nom unique à leur gène. Bien sûr, des critères préalablement définis doivent être respectés pour l'attribution d'un nouveau nom à l'élément qu'on veut caractériser. Des exemples intéressants de critères sont les algorithmes de décision pour la caractérisation d'un nouveau transposon (voir Tn Number Registry at UCL par le Dr Adam Roberts) et d'un nouvel intégron (voir Schema de Nomenclature de INTEGRALL (Moura *et al.*, 2009)).

Il subsiste aussi un bon nombre de bases de données spécialisées (cf. Nucleic Acids Research - Database issue 2014) qui se confinent sur des processus biologiques particuliers et qui sont la plupart du temps maintenues par des groupes de recherches. Leur développement est parfois arrêté par faute de financement et ce fut malheureusement le cas pour ARDB. Certaines BD réussissent toutefois à publier plusieurs articles concernant leurs mises à jour et cette pratique peut se présenter comme un bon moyen de justifier la continuité d'une BD biologique dans le domaine de la recherche. Un bel exemple est la « *Virulence Factors Database* » (VFDB) sur les facteurs de virulences bactériens qui publia trois articles (Chen *et al.*, 2005), (Yang *et al.*, 2008) et (Chen *et al.*, 2012b). Chaque BD spécialisée devrait aussi supporter activement la nomenclature de ses éléments pour éviter les éventuels conflits d'appellation de gène. Elle devrait aussi apporter suffisamment de plus-values par rapport aux BD publiques pour justifier leur existence, c'est-à-dire de créer du nouveau contenu pertinent et de partager celui-ci à la communauté visée. Enfin, interconnecter ce contenu avec d'autres bases de données pour l'enrichir est une bonne tactique que facilite le Web sémantique, sujet de la prochaine section.

2.3 Web sémantique

2.3.1 Origine

Le Web sémantique, tel que déjà introduit, est un concept inventé par Tim Berners-Lee et mené par le W3C. Le Web sémantique s'insert dans les méthodes du « *Linked Data* » tel que bien présenté dans l'article de Bizer *et al.* (2009). Le « *Linked Data* » se réfère aux bonnes pratiques pour publier et connecter des données structurées sur le Web. En résumé, le concept

vise à utiliser le Web pour connecter différentes sources de données entre elles de manière compréhensible autant par un humain qu'un ordinateur. La connexion des données doit être structurée de façon à désambiguïser les éléments (objets) de chacune des sources en utilisant un langage commun et une désignation formelle pour leurs éléments en communs. Étant basé sur le Web, la désignation d'une ressource se doit d'être un IRI [8] qui est un concept général englobant les URI [9] et les URL [10], tel que recommandé par le W3C (W3C, 2014a) et le IETF [11] (Dürst et Suignard, 2005).

En 2006, Tim-Berner Lee rédiga un document HTML (Bizer *et al.* (2009), Lee (2006)) qui décrivait 4 règles fondamentales pour publier du « *Linked Data* » sur le Web (URI ↔ IRI) :

1. utiliser des URIs pour nommer les ressources ;
2. utiliser des URIs HTTP valides pour que les gens puissent consulter ces ressources avec un navigateur Web ;
3. fournir de l'information utile sur la ressource avec des technologies du Web sémantique (RDF [12], SPARQL [13]) ;
4. inclure d'autres liens (URIs) dans leur description pour permettre de découvrir encore plus de ressources.

Ces règles, quoique simples, illustrent bien l'essentiel du Web sémantique. Tim-Berner Lee partageait avec celles-ci sa vision d'un Web beaucoup mieux connecté en voulant l'élever vers un espace global de données, tel que le suggère le titre du manuscrit de Heath et Bizer (2011) : « *Linked Data : Evolving the Web into a Global Data Space* ». La standardisation des procédures pour le faire se veut maintenant gérée par le W3C dont Tim Berners-Lee est le directeur. La prochaine sous-section présente les différentes technologies du Web sémantique et leur aspect central dans le développement de solution du « *Linked Data* ».

2.3.2 Technologies

RDF/RDFS/OWL

Le modèle de données RDF est assurément la fondation technique du Web sémantique. Il permet de modéliser de l'information sous forme de graphe sans avoir de schéma fixe au préalable et et fut conçu pour standardiser la publication et les échanges d'informations via le Web. Il existe bien sûr des couches de plus haut niveau d'abstraction bâties sur le RDF, tels le RDFS (*Resource Description Framework Schema*) et le OWL (*Ontology Web Langage*),

8. *Internationalized Resource Identifier*
9. *Uniform Resource Identifier*
10. *Unique Resource Locator*
11. *Internet Engineering Task Force*
12. *Resource Description Framework*
13. Acronyme récursif pour *SPARQL Protocol and RDF Query Language*

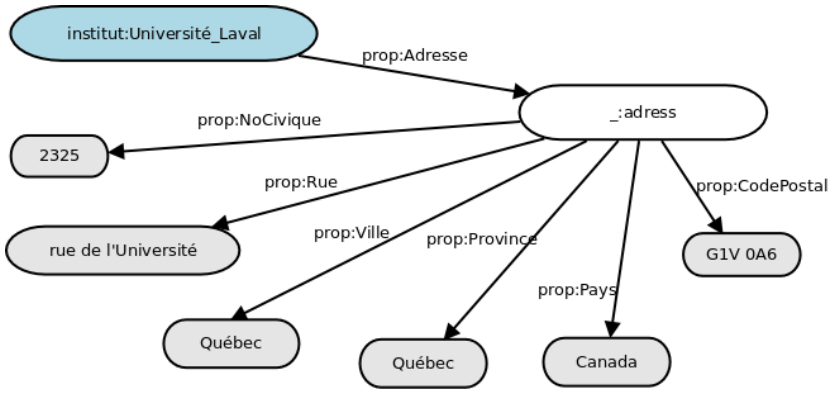


Figure 2.2 – Exemple d'un noeud anonyme (en blanc) dans un graphe RDF qui schématise l'adresse de l'Université Laval.

ainsi que plusieurs formats de sérialisations possibles sur lesquels nous nous attarderons un peu plus tard. Débutons d'abord avec une légère introspection du RDF.

Une notion fondamentale pour la modélisation de données en RDF est celle de triplet. Un triplet est essentiellement un tuple de trois éléments qui sert à désigner une relation entre le premier et le dernier avec une liaison qui pourrait être vue comme un verbe de transition. On peut faire l'analogie avec une phrase élémentaire qui pourrait être composée respectivement d'un sujet, un prédicat et un objet, tel qu'illustré à la figure 2.1. Un triplet est donc défini par convention comme un tuple sujet-prédicat-objet. Le sujet et le prédicat sont toujours des IRI, alors que l'objet peut être un IRI ou un littéral, comme un nombre entier ou une chaîne de caractères. Les littéraux sont des feuilles dans le graphe, c'est-à-dire qu'ils ne peuvent pas être le sujet ou le prédicat d'un triplet. Cependant, lorsque l'objet est un IRI il peut devenir le sujet d'un autre triplet. Un autre concept qu'il serait important d'introduire est celui de noeud anonyme. Le sujet ou l'objet d'un triplet peut être un noeud anonyme qui n'aura pas la forme d'un IRI, mais sera représenté avec un tiret bas et un double point (_ :a1). Ces noeuds anonymes ont plusieurs usages dont cinq qui ont été décrits dans Chen *et al.* (2012a). L'un des plus importants est la capacité de bâtir des relations plus complexes, comme par exemple avoir une propriété (un prédicat) avec plusieurs attributs. Un cas typique d'utilisation d'un noeud anonyme est la modélisation d'une adresse, telle qu'illustrée à la figure 2.2.

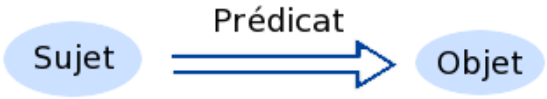


Figure 2.1 – Schématisation d'un graphe RDF avec un seul triplet.

Le RDFS (RDF-*Schema*) est une extension sémantique au vocabulaire du RDF (W3C, 2014b). Il définit plusieurs prédicats comme standards pour modéliser des données en RDF. Par convention on utilise le préfixe « rdfs » (dans un graphe au format Turtle ou dans une requête

SPARQL) pour remplacer l'IRI de l'espace de nommage : http ://www.w3.org/2000/01/rdf-schema#. Voici quelques prédicats du RDFS souvent utilisés en pratique :

— **rdfs :type** : définir le type d'une ressource ;

— **rdfs :subClassOf** : indiquer la classe parente d'une ressource ;

— **rdfs :label** : indiquer le label (étiquette) de la ressource ;

— **rdfs :comments** : indiquer un commentaire ou une courte description de la ressource.

Le OWL[14] est un langage ontologique qui a été conçu afin de permettre à un ordinateur de traiter un ensemble de données RDF et d'en extraire de nouvelles connaissances par inférence à l'aide d'un raisonneur (W3C Recommendation (2004), W3C Recommendation (2012)). Le OWL se sert d'éléments du RDF et du RDFS, mais introduit aussi plusieurs autres termes qui permettent l'abstraction de concepts et rendent possible leur raisonnement. Par exemple, il définit un type classe (owl :Class) pour regrouper des éléments communs ensemble, des termes d'égalités et d'inégalités (owl :equivalentClass, owl :equivalentProperty, owl :sameAs, etc.), des termes caractéristiques pour les propriétés (owl :inverseOf, owl :TransitiveProperty, owl :SymmetricProperty, etc.) ainsi que des restrictions et cardinalités pour les propriétés (owl :allValuesFrom, owl :someValuesFrom, owl :minCardinality, owl :maxCardinality, etc.). Dans le cadre du projet les termes utilisés sont : **owl :sameAs** pour désigner l'équivalence entre deux ressources, **owl :Class** pour définir la classe d'une ressource ainsi que la restriction **owl :Restriction** sur certaines propriétés pour lesquelles les valeurs sont ensuite liées avec les prédicats **owl :allValuesFrom** ou **owl :someValuesFrom**. Un cas d'utilisation sera présenté au prochain chapitre qui traite de la création de la BD MERGEM.

Langage de requête SPARQL

Le langage de requête pour interroger un graphe RDF est le SPARQL. La version actuelle du langage est 1.1 et ses spécifications, tout comme celle du RDF, sont maintenues par le W3C (W3C Recommendation (2008), W3C Recommendation (2013)). Le langage permet d'interroger un graphe RDF en construisant un patron basique de graphe qui est constitué de plusieurs triplets qui contiennent des variables. Chacune des variables du triplet peut être liée à une ressource exacte, lorsqu'explicitement définie, ou encore être libre (non liée) dans le but d'être résolue par l'engin de requête SPARQL. Les variables non liées sont toujours précédées d'un point d'interrogation (?a1) ou encore d'un signe de dollar ($a1). Cependant pour les noeuds anonymes la variable peut aussi être représentée avec un tiret bas et un double point (_ :a1) dans la requête. On compte 4 formes de requêtes SPARQL qui sont définies par les mots clés suivants : « DESCRIBE », « ASK », « CONSTRUCT » et « SELECT » (W3C Recommendation, 2008). Une requête de type « DESCRIBE » retournera une description de la ressource interrogée. Dépendamment de son implémentation dans l'engin de requête et de la structure du graphe, le « DESCRIBE » n'est pas assuré de retourner toute l'information pertinente sur

14. *Web Ontology Language*

la ressource, puisque les noeuds anonymes ne seront pas résolus automatiquement. Il devient alors plus pratique d'utiliser un « SELECT » pour aller chercher spécifiquement l'information dont on a besoin considérant le cas où on connaît déjà la structure du graphe RDF. Une requête « ASK » sert seulement a vérifier l'existence d'une ressource ou de certains triplets dans le graphe. Un « CONSTRUCT » permet de bâtir une requête qui retournera un graphe RDF en résultat tout en résolvant les variables non liées qui seront remplacées par leurs valeurs réelles. Un « SELECT » résoudra aussi les variables non liées par les valeurs présentes dans le graphe, mais il retournera plutôt un tableau où chacune des colonnes sera associée aux résultats des variables sélectionnées dans la requête. Quelques autres mots clés importants sont le « WHERE » qui sert à imbriquer les clauses de la requête (obligatoire pour un « SELECT »), le « FILTER » qui sert à filtrer les triplets obtenus en résultat, le « LIMIT » et le « OFFSET » qui déterminent le nombre d'éléments retournés et l'index du premier élément, le « GRAPH » qui sert à spécifier dans quel graphe exécuter la requête, le « UNION » pour joindre plusieurs patrons de triplets indépendants et le « ORDER BY » pour ordonner les résultats en sortie. Des exemples de requêtes SPARQL vous seront présentés au chapitre 3.3 qui porte sur la création du graphe RDF de MERGEM.

Format de sérialisation du RDF

Il existe plusieurs formats de fichiers (ou de sérialisation) pour le RDF qui ont chacun leurs particularités. Le tableau annexe A.2 dresse un exemple pour chacun des formats. Le RDF/XML fut le premier standard de fichier pour le RDF, il est basé sur le format XML [15] et n'est donc pas très convivial à lire pour l'utilisateur étant donnée sa syntaxe (W3C Recommendation, 2014). Un autre format, le N-triples, est très explicite puisque chaque ligne du fichier comprend un seul triplet et se termine par un point. Sa simplicité le rend trivial à décrypter par les programmes informatiques. Le N-Quads est identique au N-triples sinon avec un quatrième élément qui spécifie le graphe auquel le triplet appartient. Il permet donc de définir plusieurs graphes à l'intérieur d'un seul fichier. Le Turtle est une forme compacte du N-triples en ce sens qu'il permet la réutilisation du sujet (avec le symbole ' ;') ou du prédicat (avec le symbole ',') sans pour autant avoir à les répéter. Les symboles délimiteur '[ ]' servent à introduire un noeud anonyme dans ce format. Finalement, le JSON-LD [16] propose une adaptation du RDF au JSON qui est le format de sérialisation standard d'un objet dans une application Javascript. Il comprend aussi différentes formes de représentations, comme la forme étendue ou compacte. Au même titre que les formats Microdata et RDFa, le JSON-LD peut être inséré dans une page HTML pour permettre à un moteur de recherches d'indexer intelligemment le contenu de la page. Pour ce faire, une ontologie bien précise doit être utilisée question de partager un langage commun entre développeurs et machines. L'ontologie standard pour les moteurs de recherches (Bing, Google, Yahoo!, Yandex et autres) est définie à l'adresse

15. *Extensible Markup Language*
16. *JavaScript Object Notation for Linked Data*

suivante : http://www.schema.org.

Ontologies

Si schema.org standardise une ontologie pour les moteurs de recherches, les ontologies biomédicales se raccordent autour du format de fichier OBO (*Open Biomedical Ontologies*). Les spécifications de ce format de fichier sont maintenues par le principal investigateur du mouvement, le projet Gene Ontology (Ashburner *et al.* (2003), Day-Richter (2006)). Un consortium a aussi été créé, « OBO Foundry » (obofoundry.org), pour assurer l'interopérabilité entre toutes les ontologies biologiques et biomédicales développées par la communauté scientifique (Smith *et al.*, 2007). L'un des principaux dépôts publics des ontologies OBO est BioPortal, officiellement supporté par le NCBO [17] (Noy *et al.*, 2009). L'ontologie ARO, développée par CARD (McArthur *et al.*, 2013), fut d'ailleurs conçue avec le format de fichier OBO.

Un des points intéressants des ontologies au format OBO est qu'elles s'amalgament bien avec le mouvement du Web sémantique. En effet, il est possible de convertir n'importe quelle ontologie OBO en un format du « Linked Data », comme le OWL. La librairie utilisée pour la conversion d'un format vers un autre dans le cadre du projet se nomme ONTO-PERL et est disponible dans le dépôt officiel de paquets du langage de programmation PERL (http://search.cpan.org/dist/ONTO-PERL/) (Antezana *et al.*, 2008). Elle respecte les spécifications 1.2 décrites par le projet Gene Ontology pour la conversion d'un fichier OBO (Day-Richter, 2006). Les technologies du Web sémantique augmentent radicalement les possibilités qu'offriraient un simple fichier OBO versus un graphe au format OWL. Le OWL permet de créer littéralement une base de données de graphe en chargeant le contenu du fichier dans un magasin de triplets. Il est ensuite possible d'interroger le graphe avec le langage de requête SPARQL, de le manipuler, de l'enrichir et même de le connecter avec d'autres BD du Web sémantique.

Bases de données biologiques du Web sémantique

Le Web sémantique gagne aussi en popularité au sein des grandes institutions qui fournissent les bases de données publiques, tels le EBI ou le NCBI. Le EBI possède six services SPARQL (BioModels, BioSamples, ChEMBL, Expression Atlas, Reactome et Uniprot) qui sont tous disponibles à partir de l'adresse suivante : http ://www.ebi.ac.uk/rdf/ (Jupp *et al.*, 2014). La plate-forme RDF du EBI facilite aussi l'intégration des données entre ses différentes BD en utilisant un vocabulaire standardisé. Les annotations sur leurs données permettent aussi d'intégrer d'autres ressources comme Gene Ontology. La première des six à avoir vu le jour est Uniprot (Redaschi *et al.*, 2009) et comprend un impressionnant nombre de triplets, dépassant les 17 milliards, et qui sont distribués dans 16 graphes distincts. Le NCBI semble aussi vouloir faire partie du mouvement avec leur récente publication de PubChem au format RDF (Bolton

17. *National Center for Biomedical Ontology*

et al., 2013). Il serait cependant intéressant de voir d'autres de leurs BD voir le jour au format RDF, comme PubMed ou encore Genbank. Néanmoins, le projet Bio2RDF ayant vu le jour à l'Université Laval en 2008 s'occupe de convertir et rendre disponible plusieurs BD biologiques au format RDF (Belleau *et al.*, 2008). La version actuelle (version 3) compte 36 services, dont PubMed. Finalement, la quantité croissante de bases de données biologiques rendues publiques au format RDF allait guider les choix technologiques de la création de la BD MERGEM, sujet du prochain chapitre.

Chapitre 3

Base de données MERGEM

La base de données sur les gènes de résistance et éléments mobiles MERGEM - « Mobile Elements and Resistance Genes Enhanced for Metagenomics » - fut développée pour pallier l'absence d'un ensemble de données fiable et qui représentait la nomenclature actuelle d'un maximum de gènes connus et caractérisés. Cet ensemble de données était nécessaire aux analyses génomiques en rapport avec la résistance aux antibiotiques, plus particulièrement pour le projet CQDM en lien avec la métagénomique et qui sera présenté au chapitre 4. En effet, les projets d'annotations fonctionnelles doivent se baser sur des ensembles de gènes bien annotés qui servent de base comparative pour les nouveaux génomes issus du séquençage à haut débit. C'est pourquoi la création d'une nouvelle base de données sur les gènes de résistances aux antibiotiques était justifiée. De plus, le contexte de métagénomique et de sélectomique du projet CQDM amenait le besoin d'avoir un aspect « éléments mobiles » pour pouvoir aiguiller

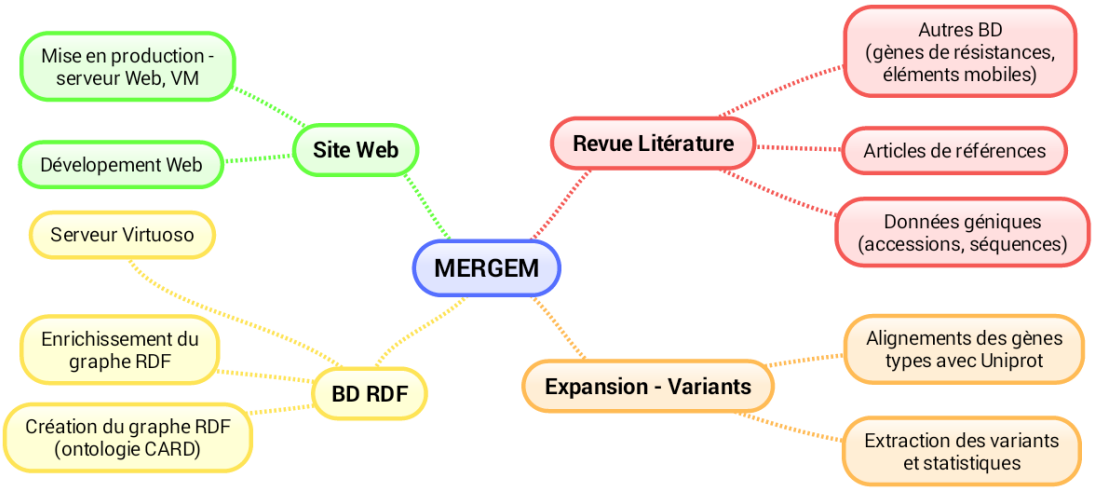


Figure 3.1 – Étapes importantes de la création de la BD MERGEM.

sur le potentiel de transmission des gènes de résistances entre différentes bactéries.

3.1 Revue de la littérature

3.1.1 Gènes de résistances

La base de données MERGEM débuta avec une compilation de gènes de résistances, extraits depuis la littérature ou autres bases de données, pour tenter du mieux possible de se représenter une nomenclature généralement acceptée dans le monde de la résistance aux antibiotiques. Un des articles pour résumer les débuts de sa création serait celui de van Hoek *et al.* (2011) qui assura une bonne terminologie au départ avec un nombre important de gènes de résistances et les identifiants des séquences associées. Les gènes de l'article avaient aussi l'avantage d'être associés à la mobilité bactérienne, soit la faculté d'être acquis ou transmis entre bactéries. Une autre ressource notable fut « The Antibiotic Resistance Database » (ARDB) (Liu et Pop, 2009), l'un des piliers des bases de données sur la résistance aux antibiotiques et qui fut citée en date d'aujourd'hui 113 fois dans la littérature. Il est à se douter qu'elle fut utilisée dans un grand nombre d'analyses génomiques sur le sujet. Les gènes qui n'apparaissaient pas dans la liste de van Hoek furent évidemment rajoutés à la collection de gènes que représentait alors MERGEM. La base de données ARDB n'était cependant plus maintenue à jour depuis 2009, rendant l'article de van Hoeuk plus à jour que celle-ci. Heureusement, il arrive que des chercheurs de renom se portent volontaires pour récolter et collectionner les gènes d'un intérêt particulier dans le but d'assurer une bonne organisation de leur nomenclature. Généralement, la nomenclature est en lien avec un antibiotique spécifique pour lequel les gènes confèrent une résistance. L'un des cas typiques est celui d'un grand nombre de β-lactamases maintenues par Jacoby et Bush étant associées à la Lahey Clinic (http://www.lahey.org/Studies/) (Bush *et al.* (1995), Bush et Jacoby (2010)). De plus, les gènes *qnr* qui confèrent une résistance aux quinolones y sont aussi hébergés et leur nomenclature préservée par Jacoby (Jacoby *et al.*, 2008) (http://www.lahey.org/qnrStudies/). À noter qu'au cours des 2 dernières années, les collections de Jacoby et Bush furent mises à jour à plusieurs reprises. Les dernières collections de leurs β-lactamases et des gènes *qnr* incorporées dans MERGEM datent de février 2014. Dans la même optique que la Lahey Clinic, l'Institut Pasteur entrepris aussi (http://www.pasteur.fr/[...]/beta-lactamase-enzyme-variants) la conservation de la nomenclature de 3 groupes de β-lactamases, soit les types OXY, OKP et LEN. Ces dernières furent aussi introduites dans MERGEM, en décembre 2012, ce qui représente toujours en date d'aujourd'hui, leur dernière mise à jour. Les β-lactamases sont très diversifiées en terme d'enzymes et sont généralement séparées, comme présentées au chapitre 1, en 4 classes distinctes (A, B, C et D). Un autre projet de BD sur les β-lactamases, mais spécifique aux classes B, est « The Metallo-Beta-Lactamase Engineering Database » (MBLED) (Widmann *et al.*, 2012). Elle permit encore une fois d'étendre la nomenclature des β-lactamases au sein de MERGEM, plus particulièrement pour le sous-groupe des métallos-β-lactamases.

Une autre source de β-lactamases provient d'une étude métagénomique du microbiote intestinal qui caractérise celui-ci comme un réservoir de gènes de résistances aux antibiotiques (Sommer *et al.*, 2009). Étant donnée sa ressemblance avec le projet CQDM, l'incorporation de leurs gènes dans MERGEM allait de soi pour être en accord avec leur nouvelle nomenclature de β-lactamases (blaHG[A-I] et blaHOA). Une certaine vigilance doit toutefois être de mise avec celles-ci puisqu'aucune d'elles n'ont été caractérisées de façon biochimique. Un autre cas notoire de détenteur d'une nomenclature, supporté par MERGEM, est celui de Marilyn C. Roberts (http://faculty.washington.edu/marilynr/) qui compile et garde une collection des gènes de résistances aux macrolides, lincosamides, streptogramines, kelotides et oxazolidinones (MLSKO) (Roberts *et al.* (1999), Roberts (2008)). Elle y maintient aussi la nomenclature des gènes *tet* - associés à la résistance aux tétracyclines - en partenariat avec le Dr Stuart B. Levy (Levy *et al.* (1999), Chopra et Roberts (2001)). Une autre compilation très intéressante est celle des enzymes de modification des aminoglycosides reportées dans l'article de revue de Shaw *et al.* (1993) et qui furent tous incorporées dans MERGEM.

Question d'agrandir l'étendue de la collection de gènes de résistance de MERGEM, sans pour autant sacrifier leurs nomenclatures, une expansion fut entreprise avec la portion SwissProt de UniprotKB. Puisque les protéines de SwissProt sont manuellement révisées par des scientifiques, son contenu peut être considéré suffisamment fiable pour accomplir cette tâche. Toutes les séquences protéiques de SwissProt furent donc alignées avec le logiciel d'alignement de séquences Fasta36 (Pearson, 2000), similaire au populaire BLAST (Altschul *et al.*, 1990), mais qui possède en plus une implémentation MPI [1] le rendant ainsi exécutable sur plusieurs noeuds de calculs en parallèle. Fasta36 se voulait donc un outil de choix pour faire des alignements de séquences protéiques en masse, distribués sur différents superordinateurs que nous avions à notre disposition. Deux superordinateurs du consortium Calcul Québec (http://www.calculquebec.ca/) furent spécifiquement utilisés durant le projet : « Colosse » de l'Université Laval et « Mammouth » de l'Université de Sherbrooke. Les variants des gènes types de MERGEM ainsi trouvés dans SwissProt furent manuellement inspectés. De manière générale, le seuil de <80% d'identités pour définir un nouveau gène type fut appliqué. Ce pourcentage d'identités tire référence des articles précédemment mentionnés de Stuart B. Levy et Marilyn C. Roberts. Tous les variants identifiés au dessous du seuil de 80% et qui possédaient une bonne similarité avec le gène type (>60% d'identités protéiques), une annotation fonctionnelle reliée à la résistance aux antibiotiques et une nomenclature différente des gènes déjà rapportés dans la BD MERGEM étaient donc rajoutés à cette dernière. Il existe cependant certains cas d'exceptions, comme les β-lactamases, où une seule mutation d'un acide aminé peut entraîner un changement dans l'activité catalytique de l'enzyme et alors changer son spectre. La nomenclature de ces dernières fut donc laissée telle que déjà récoltée dans MERGEM puisque le seuil de 80% d'identités ne s'appliquait dans leur cas et que leur nomenclature était déjà très élaborée depuis l'incorporation des autres sources de données, spécialisées à leur

1. *Message Passing Interface*

sujet. Plus récemment, la base de données CARD donnait suite au projet ARDB en distribuant une importante collection de gènes de résistances sur le Web (McArthur *et al.*, 2013). Ils réalisèrent aussi une ontologie, au format OBO, pour modéliser les relations entre les gènes de résistances, les antibiotiques et leurs cibles. Cette ontologie allait être utilisée par la BD MERGEM et plus de détails techniques vous seront fournis dans la section sur la création du graphe RDF. Une seule expansion de MERGEM avec les gènes types de CARD fut réalisée en 2013. À cette date, la BD MERGEM contenait déjà beaucoup plus de gènes de résistances que CARD. Cependant, comme le démontre le tableau 3.1, la plus récente version de CARD dépasse maintenant MERGEM en terme de gènes types (pour la résistance), mais ont toujours cette lacune pour les variants qui étaient un besoin prépondérant dans le développement de MERGEM.

La figure 3.2 illustre l'évolution quantifiée de la BD MERGEM en nombre de gènes types reliés à la résistance aux antibiotiques. Les principales sources incorporées aux dates de parution des différentes versions, telles qu'illustrées sur la figure, sont résumées ci-dessous :

1. 2012-12
   - article de référence de van Hoek *et al.* (2011) ;
   - les enzymes d'inactivations des aminoglycosides de Shaw *et al.* (1993) ;
   - ARDB ;
   - les gènes de résistances aux tétracyclines et MLS de Marilyn C. Roberts et Stuart B. Levy ;
   - β-lactamases et gènes qnr de la Lahey Clinic ;
   - β-lactamases de l'Institut Pasteur.
2. 2013-02
   - β-lactamases de MBLED.
   - expansion avec SwissProt.
3. 2013-04
   - expansion avec CARD.
4. 2014-02
   - mise à jour des gènes de la Lahey Clinic ;
5. 2014-08
   - β-lactamases identifiées dans le projet métagénomique de Sommer *et al.* (2009).

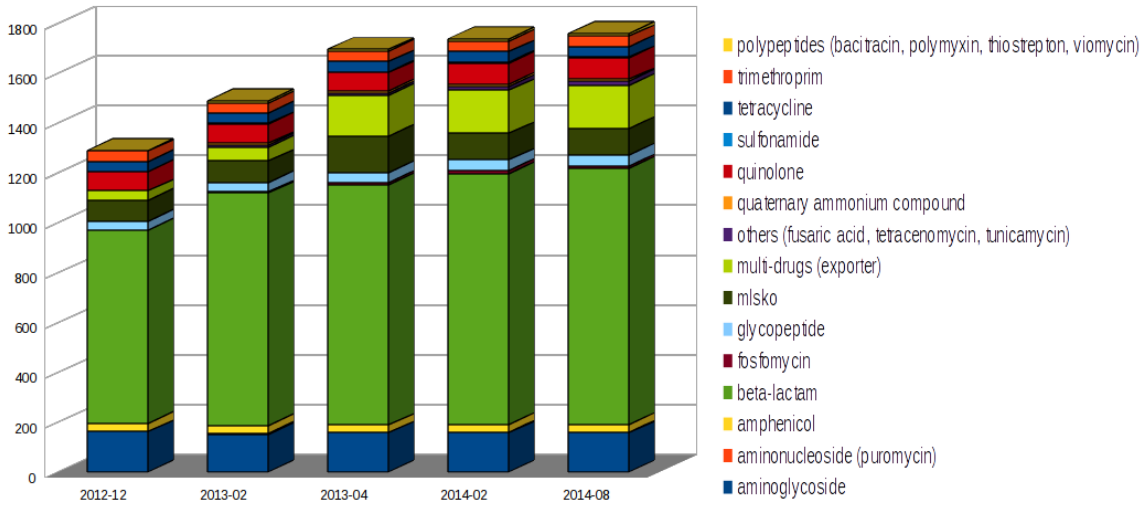


Figure 3.2 – Évolution quantitative de la BD MERGEM en nombre de gènes de résistances par classe d'antibiotique.

Table 3.1 – Comparaisons des bases de données de gènes de résistances.

	ARDB (2009)	**CARD (2014/08)**	**MERGEM (2014)**
Gènes types	380	1870	1760
+ Variants	23137	2972	101611

3.1.2 Gènes de mobilités bactériennes

Comme introduits au chapitre 1.2, les éléments génétiques mobiles (MGE) représentent un facteur important dans l'ampleur actuelle de la résistance aux antibiotiques. Il a été proposé que le microbiote intestinal humain agirait comme un réservoir de gènes de résistances aux antibiotiques (Salyers *et al.*, 2004). Par réservoir on sous-entend que le microbiote a le potentiel de transmettre ses gènes de résistances à d'autres bactéries, comme des pathogènes qui pourraient y transiter. Malgré les difficultés techniques que représente une étude qui évaluerait les impacts des transferts horizontaux au sein du microbiote intestinal, plusieurs évidences supportent le fait qu'ils se produisent (Huddleston, 2014). Il était donc important de considérer la mobilité dans le projet CQDM du chapitre 4.1. L'identification des MGE par génomique et bio-informatique présente toutefois un réel défi (Frost *et al.*, 2005). Les identificateurs de gènes se basent principalement sur l'utilisation de codons des séquences données en entrée, et les éléments mobiles étant de l'ADN étranger diffèrent habituellement pour leur utilisation de codons et leur contenu G+C ce qui biaise leur identification. Heureusement, l'outil Prodigal se base aussi sur les séquences de Shine-Dalgarno ce qui augmente la précision de leur identification. Un autre point relevé dans l'article de Frost *et al.* (2005) était le manque de standard pour les nomenclatures des MGE. Cela semble toujours le cas pour l'assignation des noms de plasmides par exemple, mais un important effort de classification des séquences d'insertions fut réalisé par Siguier *et al.* (2006b). Ils créèrent aussi une base de données, pour le maintien

de cette nomenclature, nommée ISfinder (https://www-is.biotoul.fr/) (Siguier *et al.*, 2006a). Comme cette collection représente la meilleure nomenclature des IS, mais aussi des transposases en général, elle fut donc incluse dans MERGEM. Le tableau 3.1.2 démontre l'étendue du nombre de gènes types d'IS transposases incorporés dans la BD MERGEM.

Famille IS	**Gènes types**	**Famille IS**	**Gènes types**
IS110	173	IS6	99
IS1595	73	IS630	108
IS1380	99	IS66	113
IS1182	101	IS607	29
IS1634	38	IS701	44
IS1	49	IS982	45
IS21	110	IS91	20
IS200/IS605	226	ISAs1	65
IS256	100	ISAzo13	12
IS3	408	ISH3	24
IS30	77	ISL3	85
IS481	101	ISLre2	10
IS4	172	ISNCY	62
IS5	431	Tn3	20
		Total=	**2894**

3.2 Expansion de MERGEM

Dans le but d'avoir une représentation plus exhaustive des gènes de résistances dans les bases de données publiques, un travail d'identification des variants à partir de la BD complète de Uniprot fut entrepris avec les gènes types de MERGEM. Plusieurs avantages ressortent d'une telle analyse. Tout d'abord, cela évite de se fier aux annotations possiblement erronées des gènes de résistances dans les BD publiques. Ensuite, cela donne un compte rendu de toute la diversité des gènes de résistances ainsi que leur répartition chez les différentes espèces bactériennes. Il était aussi nécessaire de le faire pour être en mesure de réaliser un profilage de qualité avec RayMeta dans le projet de métagénomique CQDM. Un autre exemple d'avantage pouvant être exploité d'une BD riche en variants serait le design d'amorce PCR [2] plus sensible à certains types de gènes.

Les mêmes outils que ceux présentés pour l'expansion avec SwissProt furent utilisés, Fasta36 pour les alignements en plus des scripts de programmations spécialement développés pour extraire les statistiques des alignements ainsi que les séquences des variants identifiés. Toutefois, contrairement à l'expansion avec SwissProt, qui avait pour but d'identifier de nouveaux gènes

2. *Polymerase Chain Reaction*

types, aucune validation manuelle des variants ne fut réalisée et le seuil de 80% d'identités protéiques et une couverture d'au moins 50% de l'alignement sur les deux séquences impliquées étaient considérés valide dans tous les cas. Toujours dans le but d'améliorer la qualité du profilage avec RayMeta, les gènes types ayant une numérotation développée furent regroupés ensemble. La raison est que le profilage se base sur de courts segments d'ADN (K-mers), beaucoup moins grand que la longueur d'un gène, et que les gènes qui sont numérotés diffèrent souvent entre eux que de quelques mutations ponctuelles, par exemple les β-lactamases TEM. Dans leur cas, le souhait était de quantifier les β-lactamases TEM dans leur ensemble et non pas diviser le profilage entre les différents TEM, qui ont plus de 200 représentants (TEM-1 à TEM-221). Le regroupement est donc composé de 485 classes de gènes avec un total de 101611 gènes au total incluant les gènes types et leurs variants (voir tableau 3.1). À notre connaissance, cela représente le plus grand nombre de variants jamais rapporté dans une BD de gènes de résistances.

3.3 Création du graphe RDF

Le projet de la base de données CARD avait d'innovateur la création d'une ontologie OBO nommée ARO. En effet, l'ontologie dessinait bien la classification des gènes de résistances, des antibiotiques et de leurs cibles. En plus des relations entre les éléments, des descriptions étaient disponibles pour une majorité des gènes de résistances avec les références aux articles scientifiques. Il était donc naturel dans le développement de MERGEM d'utiliser une telle ontologie pour projeter la BD vers le Web sémantique. Le fichier OBO de l'ontologie ARO fut donc converti en OWL pour être utilisé comme fondation sémantique à MERGEM. La librairie ONTO-PERL (Antezana *et al.*, 2008) fut utilisée pour la conversion de l'ontologie. Les données d'intérêts obtenus étaient donc les descriptions des entités liées à la résistance, le lien entre ces entités et leurs classes ainsi que leurs références dans la littérature. Les informations sur les séquences associées aux gènes déjà collectées dans MERGEM furent ensuite rajoutées aux graphes RDF. Pour manipuler la nouvelle ontologie et le graphe de MERGEM, plusieurs scripts furent développés avec le langage de programmation Ruby et sa librairie pour le RDF (https://github.com/ruby-rdf). De nouveaux identifiants furent donc créés pour chacun des gènes dans l'ontologie en s'assurant de bien mentionner les références et plus particulièrement celle de ARO avec le prédicat owl :sameAs qui sert à désigner l'équivalence des IRI pour une même ressource.

Il existe plusieurs engins capables de traiter, héberger et interroger des données RDF. On les appelle habituellement magasin de triplets (« *triplestore* »). Celui utilisé pour la BD RDF de MERGEM est la version libre du logiciel Virtuoso (http://virtuoso.openlinksw.com/). Virtuoso est aussi utilisé dans d'autres projets d'envergures de base de données RDF, comme la version RDF de Wikipedia - DBpedia (http://dbpedia.org/) - ou encore la version RDF de Uniprot (http://sparql.uniprot.org). Ils possèdent beaucoup de fonctionnalités, mais essentiel-

lement celles qui sont utilisées dans le projet de MERGEM sont l'engin de requête SPARQL et l'engin de facettes qui permet d'indexer les labels du graphe RDF. L'indexage des chaînes de caractères des labels permet la recherche de mots clés, comme le nom des gènes, dans toute la BD. De plus l'engin de facettes de vrituoso offre la saisie semi-automatique qui est d'ailleurs exploitée dans l'application Web (voir section 3.4).

Listage 3.1 – *Describe* de l'élément MGM_3000535, un exportateur ABC, macB

```
@prefix mgm_ontology: <http://mergem.org/ontology/> .
@prefix mgm_property: <http://mergem.org/property/> .
@prefix mgm_reference: <http://mergem.org/reference/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .

mgm_ontology:MGM_3000535 rdf:type 'owl':Class;
   rdfs:label "macB"@en;
   mgm_property:hasDefinition [ rdf:type mgm_property:Definition;
     rdfs:label "MacB is an ATP-binding cassette (ABC) transporter that
         exports macrolides with 14- or 15- membered lactones. It forms an
         antibiotic efflux complex with MacA and TolC."@en ];
   mgm_property:hasOBONamespace "antibiotic_resistance";
   mgm_property:hasReference
     mgm_reference:89f6786b3d7e0e1ea031a163ac522d6e5750a8c1,
     mgm_reference:31c5f077b7f32ee7eb46a2602aab3edbc2bec52a,
     mgm_reference:1e32f05cbc99fad3f5fbcded99ce65f5d18dd509;
   rdfs:subClassOf mgm_ontology:MGM_3000748,
     [ rdf:type owl:Restriction;
     owl:onProperty mgm_property:confers_resistance_to_drug;
     owl:someValuesFrom mgm_ontology:MGM_0000006 ],
     [ rdf:type owl:Restriction;
     owl:onProperty mgm_property:part_of;
     owl:someValuesFrom mgm_ontology:MGM_3000545];
     mgm_property:hasRGene [
       mgm_property:genbankEntry [ mgm_property:proteinAccession "AAC73966.1";
         mgm_property:seqAccession "U00096";
         mgm_property:seqBegin "920347";
         mgm_property:seqEnd "922293";
         mgm_property:seqStrand "1" ];
       mgm_property:nucleotideSequence "
           atgacgcctttgctcgaattaaaggatattcgtcgcagctatcctgccggtgatgagc[..]" ;
       mgm_property:proteinSequence "
           MTPLLELKDIRRSYPAGDEQVEVLKGISLDIYAGEMVAIVGASGSGKSTLMNILGCLDKA[..]";
       mgm_property:sequenceHeader ">U00096|920347|922293|1|AAC73966.1|macB";
       mgm_property:uniprotEntry <http://purl.uniprot.org/uniprot/P75831> ];
   owl:sameAs "ARO_3000535" .
```

Le listage 3.1 est un exemple d'une partie du graphe RDF de MERGEM au format Turtle et qui concerne le gène de résistance *macB* - un transporteur ABC qui confére de la résistance à l'érythromycine. L'identifiant unique du gène est le sujet ou la racine qui donne accès aux éléments d'informations stockés sur le gène en question. Le format de fichier utilisé pour l'exemple est le Turtle étant donné sa concision, sa convivialité pour la lecture et sa proximité avec le langage de requête SPARQL. Tout d'abord, dans l'en-tête sont définis des préfixes (en rouge) qui sont réutilisés dans le reste du graphe pour éviter de répéter l'IRI complet chaque fois. Après les préfixes, apparaît le triplet complet qui définit le type de l'élément mgm_ontology :MGM_3000535 comme étant une classe OWL (owl :Class). Le nom des gènes et les définitions sont spécifiés par le prédicat standard rdfs :label. À noter que la définition est liée par un noeud anonyme à la propriété mgm_property :hasDefinition de MERGEM, avec le symbole '[' qui sert à introduire un noeud anonyme dans le format Turtle. La définition possède donc deux attributs soit un type (rdf :type) et un label qui est la chaîne de caractères associée à la définition. La propriété mgm_property :hasOBONamespace est l'espace de nommage de l'élément (mgm_ontology :MGM_3000535) et qui provient de l'ontologie ARO ('antibiotic_resistance'). Ensuite apparaissent les références aux articles scientifiques reliées au gène *macB*. Les références pointent vers des noeuds avec une empreinte numérique ayant été calculée avec la fonction de hachage SHA-1 [3] à partir de l'identifiant Pubmed de l'article. On utilise une empreinte numérique au lieu d'un noeud anonyme dans le cas des références pour pouvoir se référer ultérieurement à un même noeud pour un même article. Cela évite de dupliquer l'information sur les articles en les stockant qu'une seule fois dans le graphe, puisque plusieurs éléments de MERGEM peuvent se référer à un même article. Les informations concernant les articles n'apparaissent pas dans le graphe en exemple (listage 3.1) pour brièveté, mais il suffit de résoudre l'IRI à un second niveau avec une requête SPARQL pour alors les obtenir. La requête SPARQL 1 dans le listage 3.2 ainsi que son tableau de résultats (3.2) démontrent bien comment procéder pour sélectionner quelques informations pertinentes à propos des articles références du gène *macB*. Le deuxième exemple de requête est une requête fédérée qui démontre toute la puissance du Web sémantique. En effet, les requêtes fédérées permettent d'interroger un deuxième service SPARQL (avec le mot clé « SERVICE ») et d'imbriquer les résultats à l'intérieur de la première requête. L'exemple utilise le service SPARQL de Uniprot (http ://beta.sparql.uniprot.org/sparql) pour extraire l'information sur la constante enzymatique K_m [4] de la β-lactamase NDM-1 pour différents antibiotiques β-lactamines. Le tableau résultats 3.3 possède donc une colonne qui provient de la BD MERGEM (celle de gauche) et une autre qui provient de la BD Uniprot (celle de droite).

3. *Secure Hash Algorithm*
4. Constante de Michaelis

Listage 3.2 – Requête SPARQL pour extraire l'information sur le gène *macB*.

```
prefix mgm_ontology: <http://mergem.org/ontology/>
prefix mgm_property: <http://mergem.org/property/>
prefix mgm_reference: <http://mergem.org/reference/>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix dc: <http://purl.org/dc/elements/1.1/>

# Requête SPARQL 1
SELECT ?author ?title ?date ?journal
WHERE {
  ?iri rdfs:label ?label .
  ?iri mgm_property:hasReference ?refID .
  ?refID dc:creator ?author .
  ?refID dc:title ?title .
  ?refID dc:publisher ?journal .
  ?refID dc:date ?date .
  FILTER (?label = ''macB'' @en)
}

# Requête SPARQL 2 - (fédéré sur Uniprot)
prefix up: <http://purl.uniprot.org/core/>

SELECT ?label ?KM
WHERE {
  ?iri rdfs:label ?label .
  ?iri mgm_property:hasRGene ?rGene .
  ?rGene mgm_property:uniprotEntry ?uniprotID .
  SERVICE <http://beta.sparql.uniprot.org/sparql>
  {
    ?uniprotID up:annotation ?annotation .
    ?annotation up:measuredAffinity ?KM
  }
  FILTER (?label = ''NDM-1'' @en)
}
```

Table 3.2 – Résultat de la requête SPARQL 1.

author	title	date	journal
"Xu, Y."	"Crystal structure of the periplasmic region of MacB, a noncanonic ABC transporter."	"2009 Jun 16"	"Biochemistry"
"Kobayashi, N."	"Membrane topology of ABC-type macrolide antibiotic exporter MacB in Escherichia coli."	"2003 Jul 10"	"FEBS Lett"
"Nishino, K."	"Virulence and drug resistance roles of multidrug efflux systems of Salmonella enterica serovar Typhimurium."	"2006 Jan"	"Mol Microbiol"

Table 3.3 – Résultat de la requête SPARQL 2.

label	KM
"NDM-1"@en	"10 uM for cefotaxime"
"NDM-1"@en	"10 uM for cephalothin"
"NDM-1"@en	"12 uM for piperacillin"
"NDM-1"@en	"16 uM for penicillin G"
"NDM-1"@en	"181 uM for ceftazidime"
"NDM-1"@en	"22 uM for ampicillin"
"NDM-1"@en	"49 uM for cefoxitin"
"NDM-1"@en	"49 uM for meropenem"
"NDM-1"@en	"77 uM for cefepime"
"NDM-1"@en	"8 uM for cefuroxime"
"NDM-1"@en	"94 uM for imipenem"

3.4 Création du site Web

Le site Web pour explorer les gènes de résistances contenus dans la BD MERGEM est basé sur le graphe RDF qui est hébergé dans un serveur Virtuoso. Le langage de programmation utilisé pour concevoir l'application est le Ruby [5] (https://www.ruby-lang.org/) et la principale librairie pour le développement Web est Sinatra http://www.sinatrarb.com/. Ruby Sinatra est un DSL (Domain-Specific Language) et à l'avantage de pouvoir être utilisé simplement pour bâtir rapidement des services Web, en résolvant programmatiquement les URL. En effet, des routines sont lancées selon la résolution de l'URL lors d'une requête HTTP à l'application. À titre d'exemple, la routine principale qui sert à afficher la page pour un gène de résistance - ou n'importe quelle ressource de la base de données - est appelée avec le mot clé « describe ». Par exemple, l'URL http ://mergem.genome.ulaval.ca/describe/MGM_3000535 sert à décrire l'élément avec l'identifiant MGM_3000535. En arrière-plan sont alors appelées plusieurs requêtes SPARQL qui résoudront les différents éléments d'informations sur la ressource. La première requête du listage 3.2 est un exemple de requête pour obtenir seulement l'information sur les articles références. Le site Web gère aussi d'autres requêtes comme celles pour obtenir l'information sur les séquences du gène, sur l'antibiotique auquel il confère une résistance ou encore sur l'opéron dont il fait partie. La figure 3.3 illustre un « describe » du gène *macB* dans l'application Web de MERGEM. On y aperçoit le menu à gauche qui sert à naviguer dans les différentes sections de la page, chacune des sections qui peut être affichée où cachée avec un clic sur son en-tête (la section des séquences est cachée dans la figure), une barre de navigation pour accéder aux différentes sections de l'application Web ainsi qu'une barre de recherche pour faire une recherche d'un gène ou autres concepts de la BD. Les différentes sections navigables à partir du menu principal servent principalement à lister les éléments d'intérêts d'une manière hiérarchique. L'exemple à la figure 3.4 liste les gènes de résistances selon la classe d'antibiotique à laquelle ils confèrent une résistance.

Le design de l'application Web se base sur le patron de conception Modèle-Vue-Controlleur (MVC) et est implémenté avec la librairie Sinatra. La décomposition vise essentiellement à séparer les données des interactions de l'utilisateur. Les modèles servent à gérer les données, dans notre cas principalement la communication avec le serveur Virtuoso pour interroger le graphe RDF avec l'aide de requêtes SPARQL. Les vues sont les différentes pages HTML générées par l'application et le contrôleur s'assure de bien diriger les routes (URL) entre les pages à partir des liens cliqués par l'utilisateur. Essentiellement, les modèles sont soit les « describe » des éléments de la BD comme les gènes de résistances (figure 3.3) ou encore le listage hiérarchique des éléments (figure 3.4).

5. Ruby est un langage de programmation interprété (au même titre que Python ou Perl) qui possède de bons outils pour le développement Web.

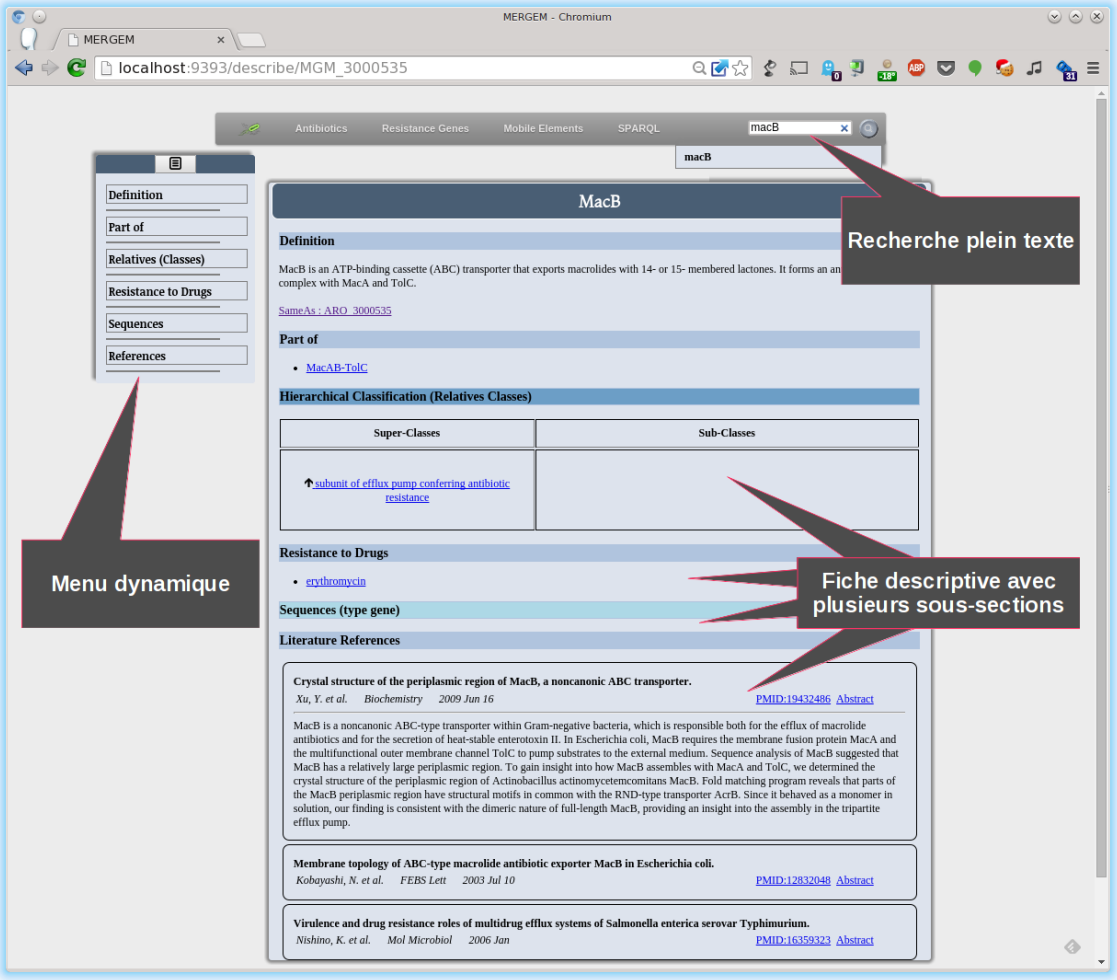


Figure 3.3 – Capture d'écran de l'application Web de MERGEM : « Describe » du gène *macB*.

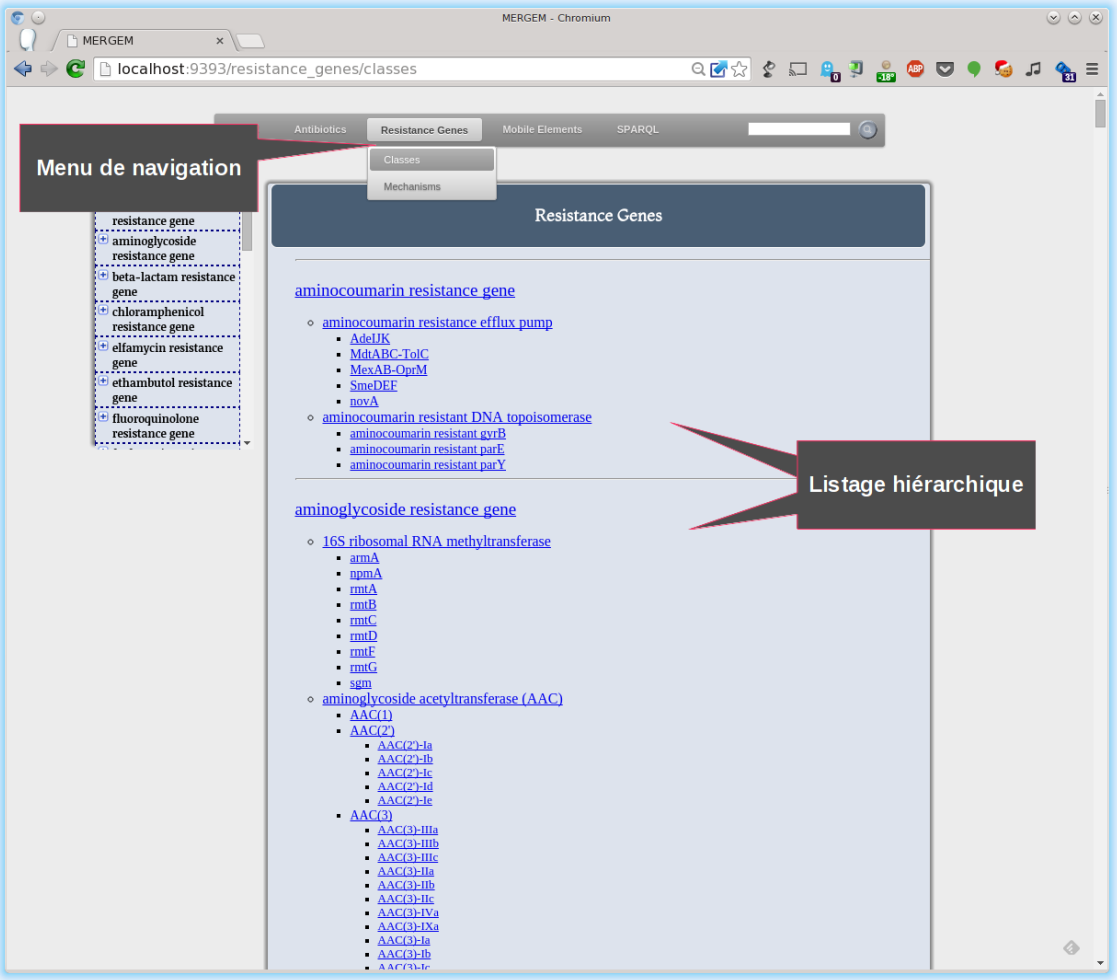


Figure 3.4 – Capture d'écran de l'application Web de MERGEM : listage des gènes de résistances par classe d'antibiotique.

Chapitre 4

Projets connexes et cas d'utilisation de MERGEM

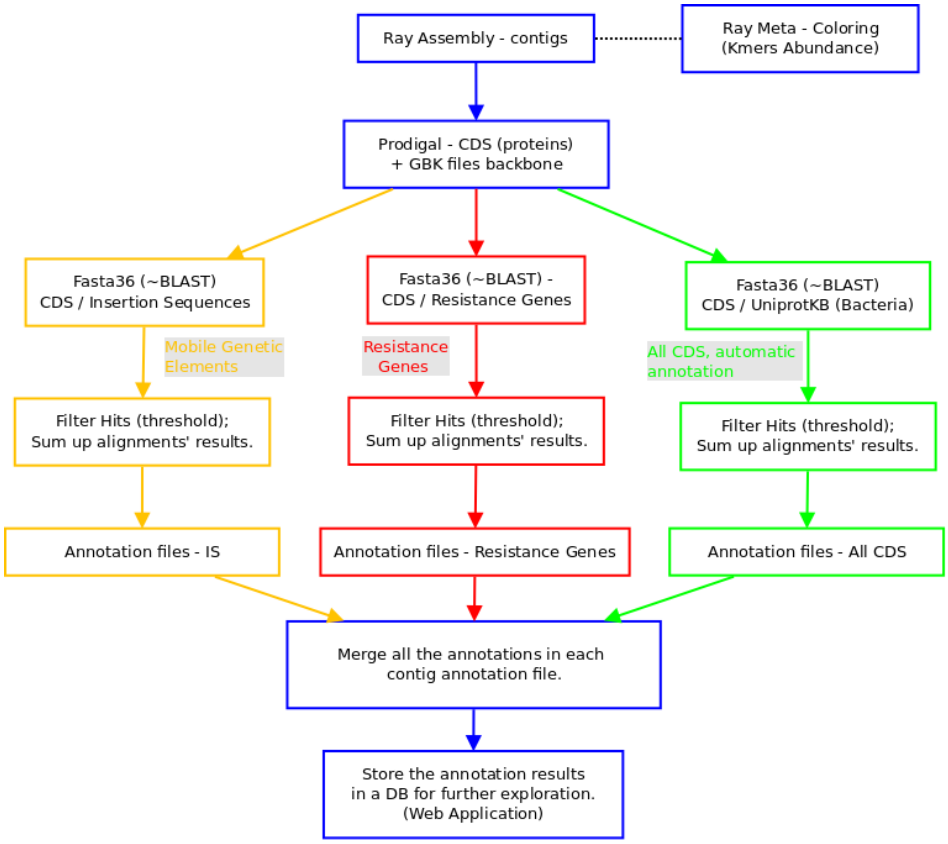
4.1 Projet CQDM - Prédire l'émergence de résistances aux antibiotiques

L'un des buts du projet CQDM était de pouvoir prédire l'émergence de gènes de résistances aux antibiotiques qui pouvaient avoir le potentiel d'être transféré entre bactéries, par des éléments génétiques mobiles (MGE). Les outils développés devaient pouvoir évaluer le potentiel de nouveaux composés antibactériens à sélectionner de la résistance (CQDM, 2011). Plusieurs groupes de recherches associés au centre de recherche en infectiologie de l'Université Laval (CRI) collaboraient sur ce projet. L'objectif principal consistait en l'administration d'un antibiotique à des patients sains et normaux selon des critères préalablement établis et révisés par un comité éthique. Le microbiote intestinal humain est bien connu pour héberger une grande quantité diverse de bactéries avec des fonctions essentielles qui contribuent à la santé humaine. Un antibiotique, même administré par voie orale, aura toutes les chances d'avoir un impact sur la communauté microbienne du microbiote, surtout pour les antibiotiques à large spectre. Le projet était donc de séquencer l'ADN provenant des fèces de patients ayant subi un traitement antibiotique (au cefprozil) et cela à 3 temps différents, soit au jour 0 (avant la prise de l'antibiotique) et aux jours 7 et 90 après le traitement. Les métagénomes, contenus dans les fèces étaient séquencés en plus des sélectomes - sélection de bactéries avec différents antibiotiques pour identifier de la résistance à d'autres classes antibiotiques que celui offert dans le traitement. L'appareil utilisé pour le séquençage était un Illumina HiSeq 1000. Suite aux séquençages, les séquences obtenues étaient analysées par bio-informatique pour ressortir les informations pertinentes sur l'actuel contenu génomique des échantillons.

D'un point de vue technique, le pipeline d'annotation est ici présenté tel que réalisé pour le projet CQDM. La figure 4.1 présente les différents logiciels utilisés pour la réalisation du

pipeline ainsi que les données qui y transitent. En plus des logiciels externes incorporés au pipeline, des scripts ont été développés spécifiquement pour celui-ci afin d'assurer le flux des données et la création de fichiers transitoires parfois nécessaires à d'autres analyses. La structure des dossiers et fichiers créée par le pipeline devait être facilement compréhensible pour que leurs contenus puissent être incorporés dans une base de données. Une application Web allait aussi être conçue afin d'offrir une interface pour faciliter l'exploration et l'accès aux résultats des analyses.

Figure 4.1 – Diagramme représentant les étapes du pipeline d'annotation.



Les différents logiciels utilisés sont donc énumérés ci-dessous selon leur ordre d'arrivée dans le flux des données. Les logiciels utilitaires développés spécifiquement pour les analyses servent en quelque sorte de colle à l'exécution du pipeline et gèrent entre autres le bon formatage des données pour leur exécution, leur stockage et leur distribution dans l'application Web.

- **Ray Assembler**
  - **Utilité** : réaliser les assemblages des séquences produites par le séquençage des métagénomes et sélectomes. Sert aussi à réaliser le profilage taxonomique des échantillons.
  - **Articles référence** : (Boisvert *et al.*, 2010), (Boisvert *et al.*, 2012)
  - **URL pour téléchargement** : https://github.com/sebhtml/ray

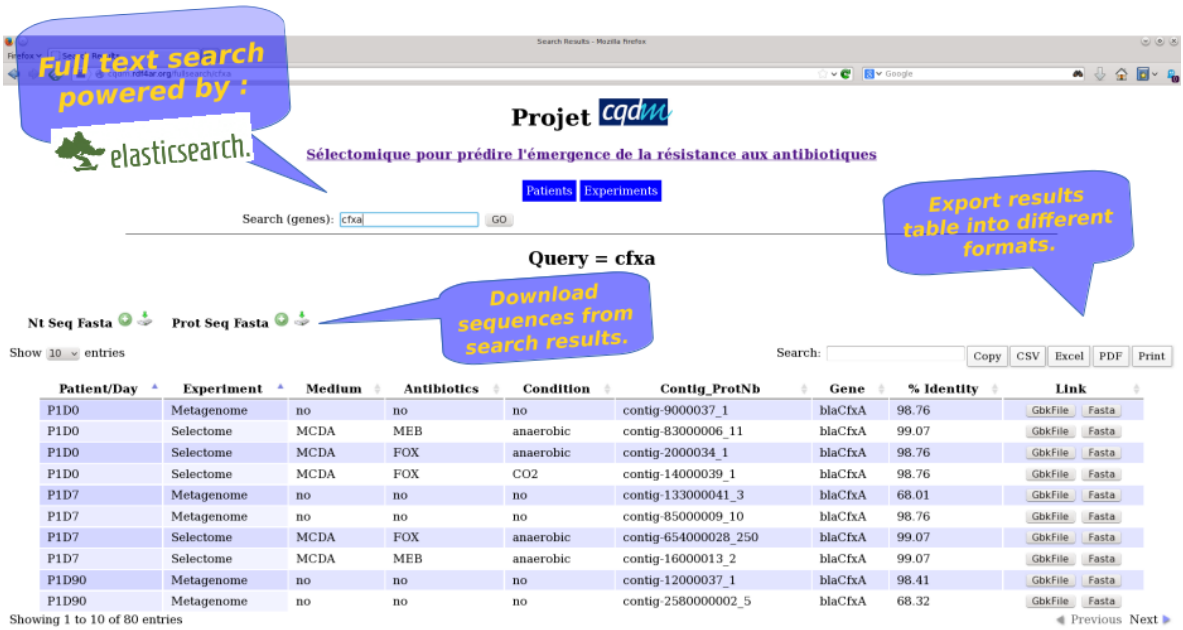
  - **Version du logiciel** : 2.3.1
- **Prodigal**
  - **Utilité** : détecter les cadres de lectures ouverts dans les séquences assemblées par Ray Assembler (*contigs*), générer les fichiers de séquences d'ADN et de protéines pour les gènes identifiés et construire le gabarit des fichiers d'annotations au format GFF (« *Generic Feature Format* ») et GenBank.
  - **Article référence** : (Hyatt *et al.*, 2010)
  - **URL pour téléchargement** : http://prodigal.ornl.gov/downloads.php
  - **Version du logiciel** : 2.60
- **Fasta36**
  - **Utilité** : réaliser les alignements en masse sur un superordinateur dans le but d'identifier les séquences codantes sorties de Prodigal et de réaliser leur annotation fonctionnelle.
  - **Article référence** : (Pearson, 2000)
  - **URL pour téléchargement** : http://faculty.virginia.edu/wrpearson/fasta/fasta36/
  - **Version du logiciel** : 36.3.5c
- **Logiciels utilitaires** (développés spécifiquement pour les analyses du projet)
  - **Prodigal-SumUp-Blast.pl**
  - **Utilité** : créer des fichiers résumés pour les résultats d'alignements de séquences, précédemment réalisés avec Fasta36. Générer des fichiers au format TSV (« Tabular-Separated Values »), visualisable avec un chiffrier comme Microsoft Excel. Identifiers les plus proches homologues des résultats d'alignements pour chaque séquence codante avec le pourcentage d'identités et de couvertures des alignement dans les deux sens (pour la séquence requête et celle du gène type qui sert à l'annotation).
  - **Dépendances** : BioPerl (http://bioperl.org)
  - **Parse-Annotations-to-Gbk.rb**
  - **Utilité** : rapporter les gènes identifiés avec les alignements, depuis les fichiers résumés, à l'intérieur des fichiers d'annotations générés par Prodigal.
  - **Dépendances** : BioRuby (http://bioruby.org)

4.1.1 Application Web

L' application Web du projet CQDM fut dévelopée afin de fournir un outil qui serait facilement utilisable par des biologistes pour l'exploration des données produites par les analyses bioinformatiques. Les technologies utilisées pour bâtir l'application sont en partie les mêmes que pour le site Web de MERGEM, soit le langage de programmation Ruby et la librairie Sinatra.

Le système de gestion de base de données n'est cependant pas Virtuoso, mais plutôt CouchDB (http://couchdb.apache.org/). CouchDB est un engin de base de données de documents qui peut être interrogé comme un magasin clé-valeur et qui utilise le format JSON pour le stockage, le JavaScript pour l'indexage et fournit une interface de programmation d'applications (API) via le protocole HTTP. De plus, l'outil Elasticsearch (http://www.elasticsearch.org/) est utilisé pour indexer les données chargées dans la BD CouchDB et offrir une recherche plein texte à l'utilisateur. L'application Web fournit un accès aux données de la BD avec de simples requêtes bâties au sein de l'URL qui sont gérées par les routes avec la librairie Sinatra. Par exemple, cette adresse URL :
http ://cqdm.genome.ulaval.ca/download/gbk/Metagenome_7_90_no_no_contig-xxxxx,
les 3 paramètres suivants [download][gbk][Metagenome_7_90_no_no_contig-xxxxx] donneraient comme résultat le fichier d'annotations genbank sélectionné au troisième paramètre de l'URL, soit le contig-xxxxx du métagénome du patient 7 au jour 90 (voir l'exemple de la figure annexe A.7).
Un autre bel exemple d'URL est celui de la recherche plein texte pour un gène en particulier dans tous les échantillons analysés : http ://cqdm.genome.ulaval.ca/fullsearch/cfxa. Avec les deux paramètres [fullsearch] [cfxa], l'action entreprise par l'application serait de faire une recherche plein texte pour le gène *cfxa* dans tous les gènes annotés de toutes les expériences. À noter que le gène *blaCfxA* code pour une β-lactamase qui est connue pour être retrouvée dans les espèces bactériennes du microbiote intestinal humain comme en fait foit les 80 résultats à la figure 4.2.

Figure 4.2 – Recherche plein texte pour un gène d'intérêt.



On peut aussi remarquer sur la figure que les résultats sont rapportés dans un tableau de style Excel offrant des fonctionnalités de base intéressante pour l'exploration des résultats. On peut entre autres trier les colonnes, il est même possible de faire un tri multiple sur plusieurs colonnes à la fois en tenant la touche « SHIFT » du clavier. On peut aussi filtrer dynamiquement les résultats avec un mot clé dans la boîte texte juste au dessus du tableau ; cela peut s'avérer utile pour regrouper les résultats d'une expérimentation, d'un médium ou d'une condition quelconque. Il est aussi possible d'exporter le tableau des résultats dans les différents formats que proposent les boutons situés à côté de la boîte texte pour le filtre - CSV, Excel, PDF et PRINT ou COPY. L'utilisateur pourrait donc travailler au besoin avec le tableur de son choix (comme Microsoft Excel) à partir du tableau des résultats généré par l'application Web. Il est aussi possible de télécharger toutes les séquences nucléotidiques ou protéiques des gènes identifiés à l'aide de la recherche plein texte.

Étant donné que deux des éléments fondamentaux du projet CQDM étaient les patients et l'expérimentation du séquençage, il était logique de placer ces deux rubriques dans l'en tête des pages Web. Les figures annexes A.3 et A.4 illustrent respectivement la navigation pour les patients et les expérimentations. Le listage des données est donc dépendant de la rubrique choisie question d'offrir l'information pertinente à partir de la sélection. Un autre listage intéressant, voir figure annexe A.4 est celui réalisé pour un échantillon quelconque (par exemple : le métagénome du patient 3 à jour 0) où tous les gènes de résistances sont rapportés dans le tableur avec la possibilité d'accéder aux séquences des gènes listés (figure annexe A.6) où bien encore au fichier d'annotations GenBank du *contig* (figure annexe A.7). Dans les deux cas, la séquence (du gène ou du contig) peut être soumise à un BLAST directement au NCBI qui s'ouvrira automatiquement dans une nouvelle fenêtre du navigateur Web de l'utilisateur.

4.2 Projets Pseudomonas

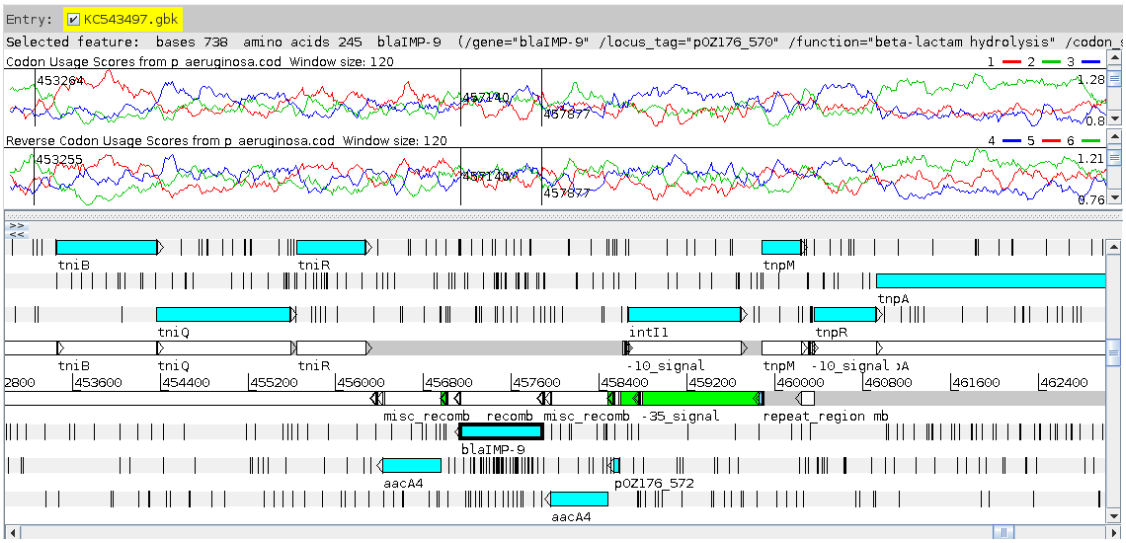
4.2.1 Plasmid pOZ176

Le plasmide pOZ176 est très intéressant puisque, déjà avec une taille imposante de plus de 500 kb (« kilobases » ou millier de paires de bases (pdb)), il est le premier plasmide du groupe d'incompatibilité IncP-2 a être publié. De plus, il contient deux intégrons de classe I lesquels sont positionnés sur des transposons qui furent nommés Tn*6016* (avec blaIMP-9) et Tn*6217*. À noter que mon travail se voulait d'abord et avant tout un projet d'annotations manuelles des gènes du génome du plasmide qui allaient être déposées dans la base de données publique GenBank du NCBI, au moment de la soumission de l'article. Un choix parfois judicieux devait être fait pour s'assurer d'attribuer la bonne nomenclature à chacun des gènes, soit la meilleure sémantique possible pour les produits protéiques.

Le fichier d'annotations GenBank est présenté en partie dans l'annexe A.4.1 (l'en-tête, la source, l'annotation du premier gène et du gène *blaIMP-9*) ; le fichier complet est toutefois

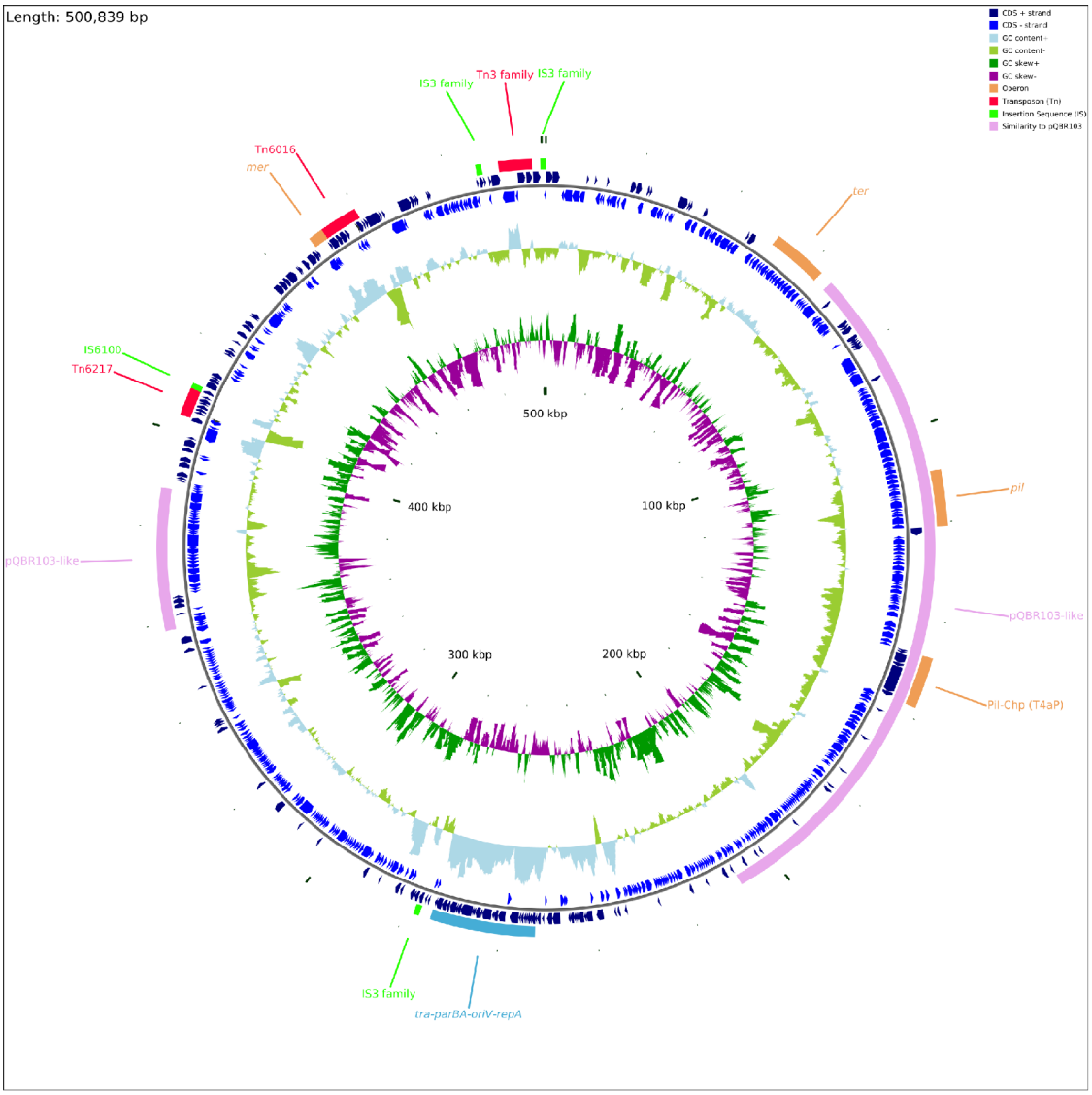
disponible à l'adresse suivante http://www.ncbi.nlm.nih.gov/nuccore/496684371. Les statistiques ressorties avec le logiciel Artemis du Wellcome Trust Sanger Institute (Rutherford *et al.*, 2000) sur le fichier GenBank sont disponibles à l'annexe A.4.2. Le nombre de protéines hypothétiques (« hypothetical proteins ») a cependant été extrait avec des outils développés maison pour manipuler les propriétés d'un fichier d'annotations GenBank. Il est à noter que les protéines hypothétiques (396) représentent 66% des protéines totales du plasmide (596). La densité génique est aussi impressionnante avec un total de ∼1.2 kilo bases, soit moins de 837 pdb par gène. La moyenne de pdb par séquence codante est de 719 (∼240 acides aminés), donnant un pourcentage codant de 85.9% au plasmide. Le résumé des entrées d'annotation distingue entre autres 4 éléments mobiles, 6 régions répétées, 23 sites de recombinaisons et 1 origine de réplication comme propriétés.

Figure 4.3 – Intégron de type I sur le transposon Tn*6016* de pOZ176.



La capture d'écran du logiciel Artemis, à la figure 4.3, illustre une partie du fichier d'annotations du plasmide axée sur la région du premier des deux intégrons, localisé sur Tn*6016* et qui contient la β-lactamase IMP-9 . Il est intéressant de regarder la fenêtre du haut sur l'utilisation de codon des gènes qui se base sur un fichier compilé spécialement pour les *Pseudomonas* (fichier d'utilisation de codon utilisé : http://www.kazusa.or.jp/codon/[..]/species=208964). On voit entre autres que le gène *blaIMP-9* ne possède pas une bonne utilisation de codon, comparativement à intI1 par exemple, suggérant une récente acquisition du gène chez les *aeruginosa*. L'importance clinique du gène en question (blaIMP-9) est due à la résistance qu'il confère aux antibiotiques carbapénems, comme méropénem et imipénem, qui sont souvent considérés comme des antibiotiques de dernières lignes de défense dans les cas de multirésistances, comme c'est le cas chez PA96 principalement à cause de la présence du plasmide pOZ176 . Étant donnée la structure mosaïque des génomes de plasmides, faire une annotation manuelle de ceux-ci s'avère souvent nécessaire et assure une meilleure caractérisation comparativement à une annotation automatique.

Figure 4.4 – Carte du plasmide pOZ176 avec ses différents éléments géniques importants. Tiré de Xiong *et al.* (2013) avec permission.



Un article sur le projet du plasmide pOZ176 fut publié dans le journal Antimicrobial Agents and Chemotherapy (AAC) (Xiong *et al.*, 2013). La figure 4.4, réalisée à l'aide du logiciel CGView (Stothard et Wishart, 2005), illustre la carte génique circulaire du plasmide avec les éléments d'intérêts indiqués en couleur sur les arcs externes. On aperçoit en rouge trois éléments mobiles dont les deux intégrons, en vert les séquences d'insertions [1] dont 3 sont de la famille IS3 et l'autre de la famille IS6100. En orange sont indiqué quatre opérons identifiés comme suit :

— l'opéron *mer* provoquant la résistance au mercure ;

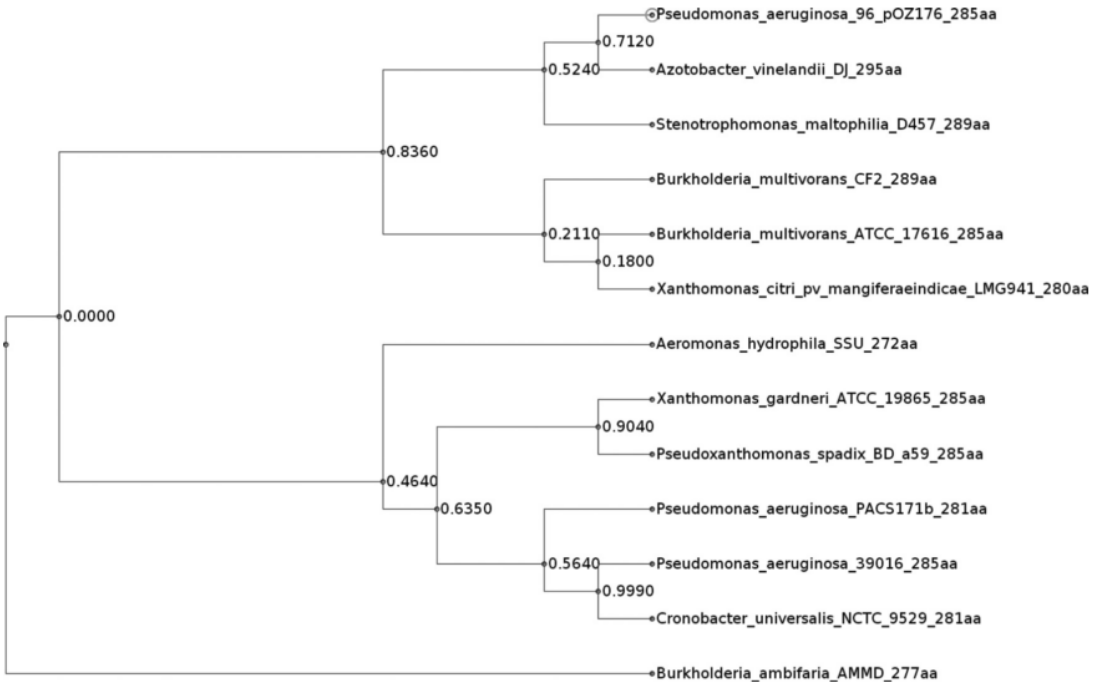
1. À noter que les séquences d'insertions, étant transposables, sont fréquemment répétées dans les génomes et plasmides ce qui rend la vie difficile aux assembleurs basés sur le graphe de Bruijn.

— l'opéron *ter* provoquant la résistance au tellure ;

— 2 opérons *pil* associé à l'assemblage de pili.

Les pili ont au moins deux phénotypes connus : les pili communs servent à la motilité de la bactérie alors que les pili sexuels peuvent servir au transfert de matériel génétique d'une cellule bactérienne à une autre par conjugaison. Il serait peut-être important de mentionner que le deuxième opéron Pil-Chp est vraisemblablement de type T4aP, tel que décrit dans Burrows (2012), et aurait une influence sur la pathogénicité de la bactérie en étant impliqué dans la formation de biofilm et dans la colonisation de l'hôte. À savoir si le système de pili est totalement fonctionnel, étant situé sur un plasmide au lieu du chromosome, et si tel est le cas est-ce qu'il serait aussi fonctionnel chez d'autres espèces bactériennes sont des questions qui demeurent à répondre. Les plasmides sont très souvent un amalgame de gènes acquis de différents hôtes et peuvent parfois être transmis grâce à leur système de conjugaison.

Figure 4.5 – Arbre phylogénétique du gène d'origine de réplication *repA*. Arbre ML (Maximum Likelihood) réalisé avec le logiciel MEGA. Tiré de Xiong *et al.* (2013) avec permission.



L'origine de pOZ176 peut être discutée à l'aide de génomique comparative, la figure 4.4 arbore un arc de similarité (en mauve) avec un autre plasmide, nommé pQBR103, aussi trouvé chez un *Pseudomonas aeruginosa* isolé à la surface d'une feuille de betterave (Tett *et al.*, 2007). À l'oeil, la carte génique suggère qu'environ le tiers du plasmide est similaire à pQBR103. À noter que la similarité fut évaluée au niveau protéique puisque les séquences nucléotidiques étaient trop divergentes pour ressortir une bonne synténie. Cela indique que malgré leur ressemblance, leur origine est plutôt lointaine considérant le temps évolutif nécessaire pour accumuler suffi-

samment de mutations pour briser les alignements de leur séquence nucléotidique. En revanche, les protéines doivent garder une certaine similarité pour ne pas faire perdre l'activité et cela nous permit donc de trouver les ressemblances entre les deux plasmides. Une autre stratégie utilisée dans le but d'investiguer l'origine du plasmide fut la construction d'un arbre phylogénétique du gène *repA* avec ses plus proches homologues dans GenBank, ressorties à l'aide d'un BLAST. La figure 4.5 expose l'arbre en question avec le gène *repA* de pOZ176 positionné en première position. Il est intéressant de voir l'étendue des genres bactériens retrouvés avec un *repA* semblable à celui de pOZ176 : *Azotobacter*, *Stenotrophomonas*, *Burkholderia*, *Aeromonas*, *Xanthomonas*, *Pseudoxanthomonas*, *Cronobacter* et bien sur *Pseudomonas*. Tous ces genres sont du phylum des protéobactéries et sont tous des gammaprotéobactéries à l'exception de *Burkholderia* qui est une bêtaprotéobactérie. En fin de compte, ce sont toutes des bactéries à Gram négatif et leur phylum englobe un nombre important de bactéries pathogènes. Cela démontre bien l'étendue du danger que peut provoquer la dissémination d'éléments mobiles autant pour causer la résistance aux antibiotiques que la virulence chez des agents pathogènes.

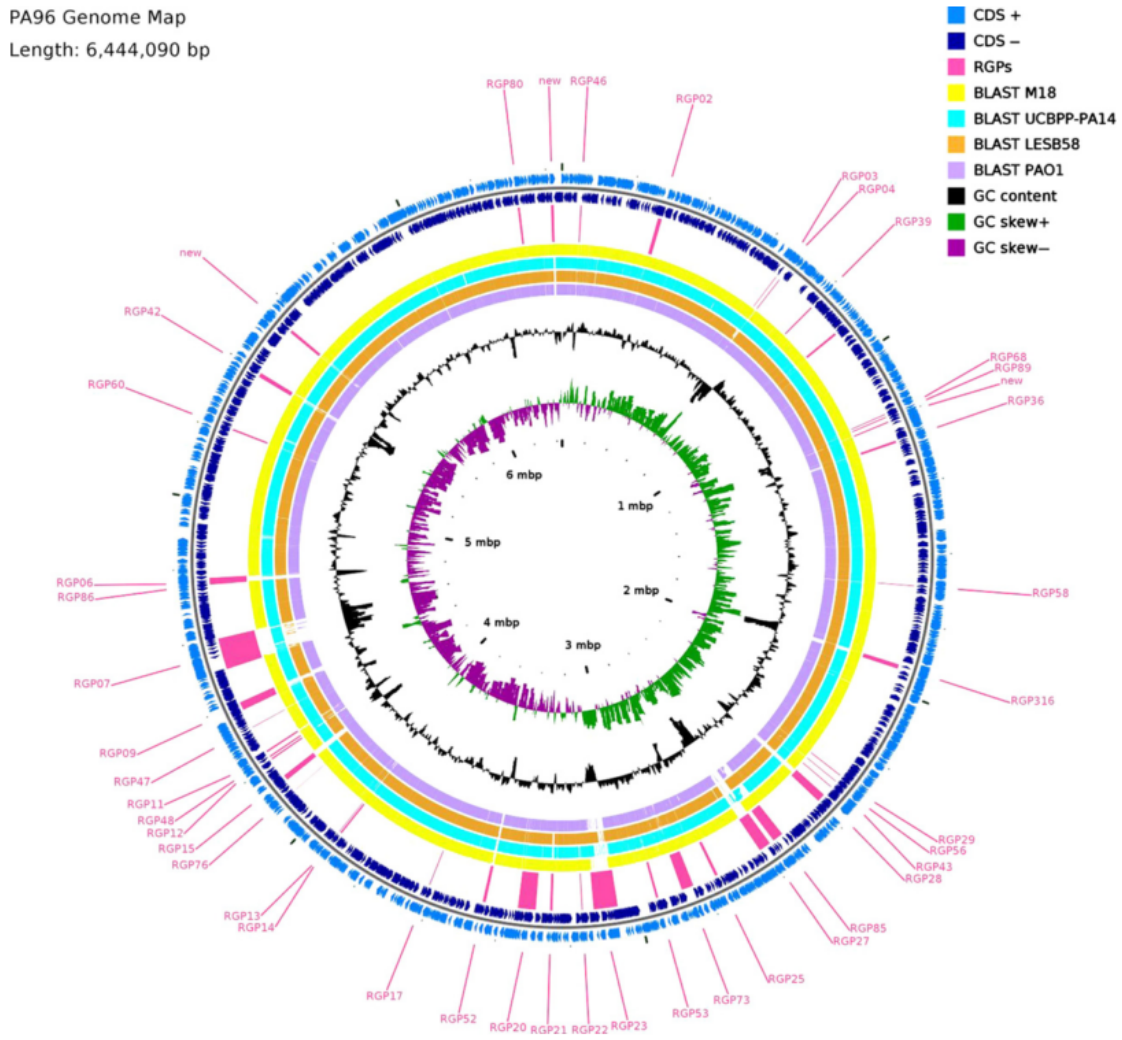
4.2.2 Chromosome de la souche PA96

Suite à la conférence Pseudomonas 2013 (prochaine section), un numéro spécial de la revue FEMS [2] allait être dédié à la biologie des *Pseudomonas* (voir http://issuu.com/fems/[..]). L'opportunité était belle pour soumettre une lettre à cette édition spéciale à propos du génome PA96 hôte du plasmide pOZ176 (Déraspe *et al.*, 2014). Encore une fois, la génomique comparative nous permit de discuter de l'origine de cette souche multirésistante aux antibiotiques, isolée de Guangzhou en Chine.

Une grosse partie de l'analyse génomique de PA96 se concentrait sur son génome accessoire. Pour se faire, la nomenclature des îlots génomique définie dans Mathee *et al.* (2008) allait être utilisée dans les analyses. Cette nomenclature définit les îlots comme des régions de plasticité génomique (RGP [3]) et leur associe une numérotation spécifique et uniforme. La synténie et la phylogénie avec plusieurs autres génomes de références de *Pseudomonas aeruginosa* allaient être réalisées dans le but de fournir une vue comparative et caractéristique de PA96. La figure 4.6 montre la carte du génome de PA96 ainsi que sa synténie avec les 4 génomes de références suivants : PAO1, LESB58, UCBPP-PA14 et M18. De plus, la figure 4.7 de l'arbre phylogénétique [4] supporte aussi l'idée de comparaison génomique et est en accord avec celui de Stewart *et al.* (2014), précédemment publié et qui distinguait trois groupes différents d'*aeruginosa* : les PAO1-*likes* à gauche, les PA14-*likes* à droite et les PA7-*likes* en bas. PA96 se situe dans le groupe des PAO1-*likes* et partage une branche commune avec la souche M18. On peut évidemment suggérer qu'un tel rapprochement dans la phylogénie entre M18 et PA96

2. *Federation of European Microbiological Societies*
3. *Regions of Genomic Plasticity*
4. Les génomes inclus dans l'arbre sont tous les génomes complets de *Pseudomonas aeruginosa* trouvés dans GenBank en date de janvier 2014.

Figure 4.6 – Carte génomique de *Pseudomonas aeruginosa* PA96, tiré de Déraspe *et al.* (2014) avec permission.

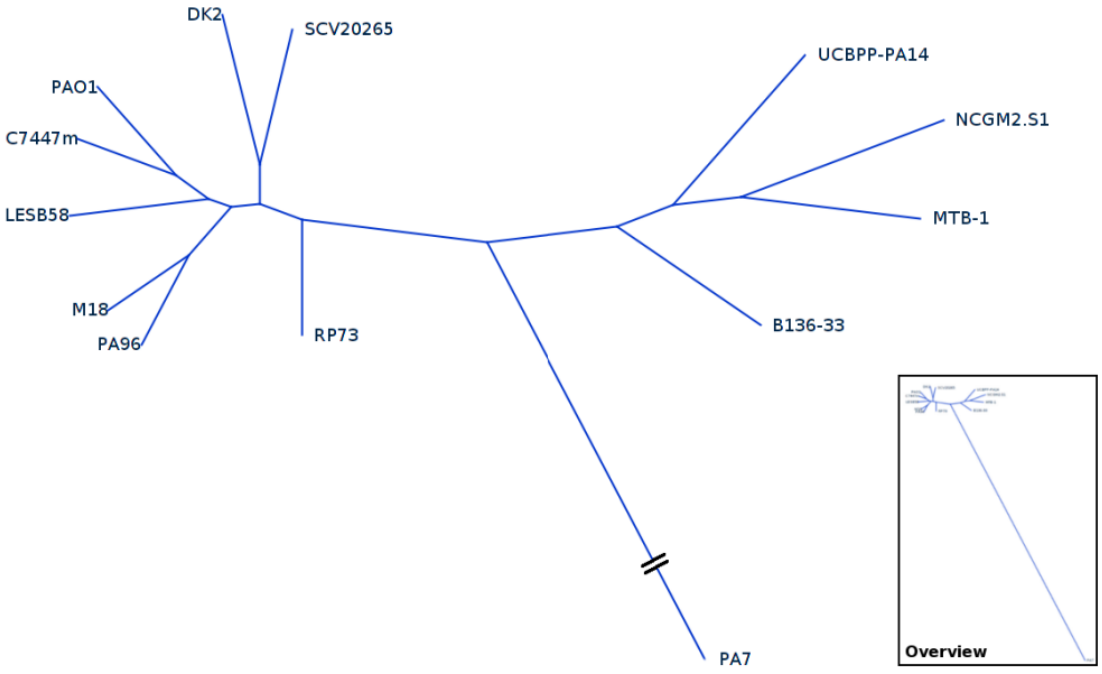


prouve l'existence d'un ancêtre commun aux deux souches. M18 (Wu *et al.*, 2011) est une souche environnementale isolée de la rhizosphère et démontre bien le potentiel adaptatif que possèdent les *Pseudomonas aeruginosa*. Ils sont en effet capable de coloniser différentes niches dont l'humain fait partie et chez lequel ils peuvent causer diverses infections - pulmonaires, urinaires, intraabdominales, cutanées, etc. L'article met donc en évidence le lien entre souches cliniques et environnementales chez les *Pseudomonas aeruginosa*.

4.2.3 Analyse de plusieurs souches de *Pseudomonas* de type PA7-*likes*

En 2013, avait lieu la conférence annuelle de *Pseudomonas* à Lausanne en Suisse à laquelle allait participer le Dr. Paul H. Roy avec la présentation d'une affiche (Déraspe *et al.*, 2013). L'affiche concernait l'analyse comparative de différentes souches de PA7-*likes* (voir Roy *et al.* (2010)) issues du séquençage de plusieurs de leur génome, sur un Illumina MiSeq. Le travail se déroulait

Figure 4.7 – Arbre phylogénétique de *Pseudomonas aeruginosa* PA96, Déraspe *et al.* (2014) avec permission.



en collaboration avec un groupe de recherche de l'Université de Buenos Aires. Je m'occupais principalement des analyses bio-informatiques qui sont nécessaires après les expériences de séquençage. Les génomes étaient donc tous des PA7-*likes* et il devenait intéressant d'analyser les îlots génomiques qui se distinguent du génome essentiel (« core genome ») par leur absence dans le pangénome. Pour pouvoir distinguer les îlots génomiques, la solution apportée était de prendre le génome ou plutôt le protéome - toutes les séquences codante (CDS) - du génome original PA7 et de trouver les homologues de ces protéines dans les autres souches de PA7-*likes*. Un aperçu du fichier résultat est en annexe A.5 et donne une idée du type d'analyse de génomique comparative avec lesquels nous pouvons travailler. De plus, le même type de comparaison fut produit pour les gènes de virulence rapportés dans VFDB. Les conclusions de l'affiche apportaient la constatation que les PA7-*likes* étaient tous dépourvus du système de sécrétion T3SS et de ses facteurs de virulence (protéines sécrétées par le système) ExoS, ExoU, ExoT, ExoY ainsi que la toxine ToxA (voir A.5). L'affiche fut d'ailleurs citée dans l'article de Elsen *et al.* (2014) qui proposa la découverte des facteurs de virulence nécessaire à la pathogénicité des PA7-*likes*, soit la toxine ExlA qui serait exportée via ExlB.

4.2.4 Pseudomonas aeruginosa - AstraZeneca

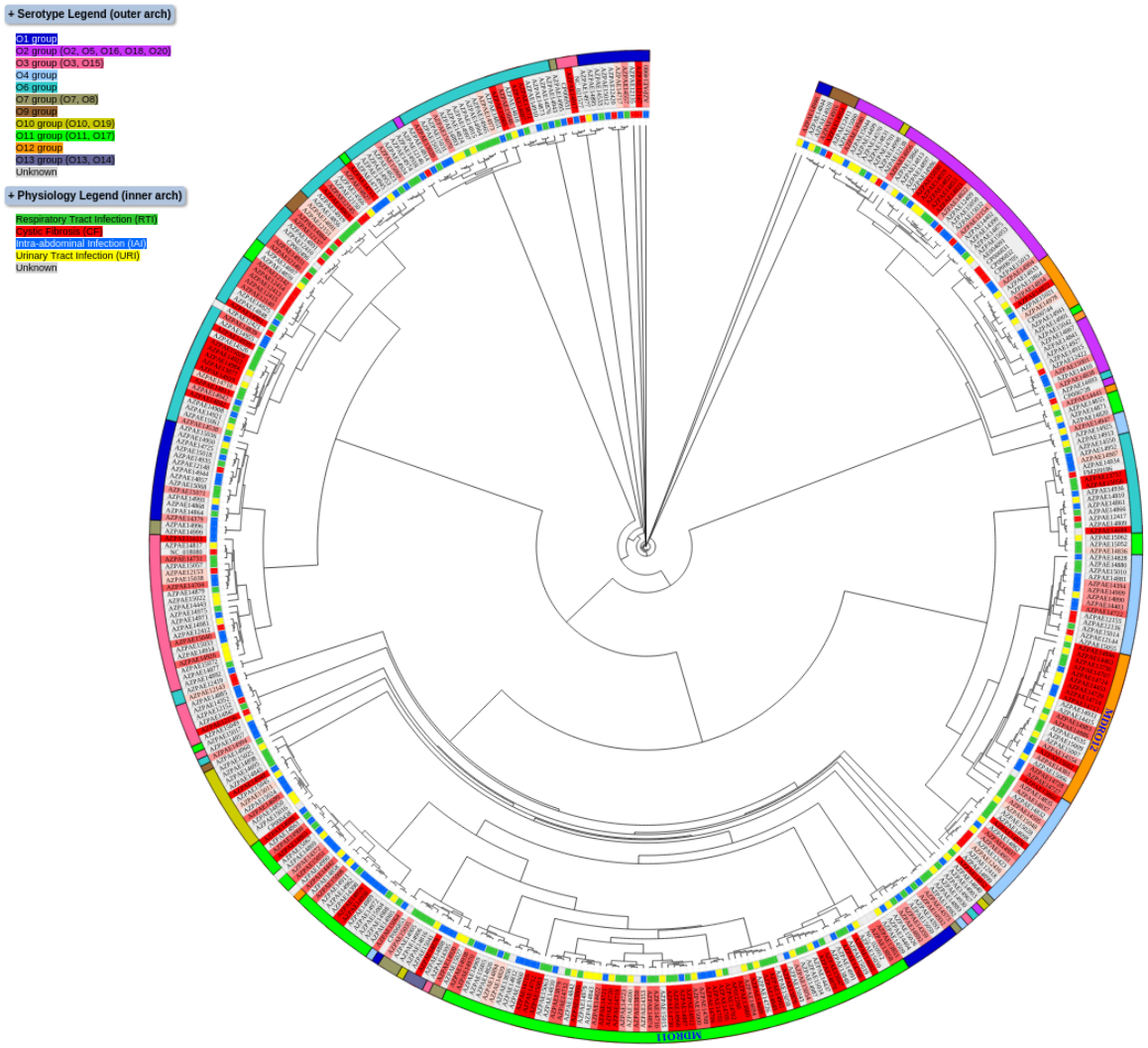
Le projet CQDM avait trois compagnies pharmaceutiques mentors : Merck, Pfizer et AstraZeneca. Suite au projet, la compagnie AstraZeneca fut intéressée à la BD MERGEM et au pipeline d'annotations pour un de leurs projets. Ces nouveaux collaborateurs avaient en main

les données de séquençages de 390 génomes de *Pseudomonas aeruginosa* qui avaient été choisis pour leur profile de résistance et souhaitaient, entre autres, caractériser le résistome de ceux-ci. Ils avaient aussi les concentrations minimales inhibitrices (CMI) pour quatre antibiotiques (méropénem, doripénem, lévofloxacine et amikacine) et désiraient les corréler aux données génomiques. On entreprit donc cette collaboration concernant l'analyse de leur collection de génomes de *Pseudomonas*.

Les gènes de résistances contenus dans les 390 génomes furent donc identifiés avec le même pipeline d'annotations que celui développé pour le projet CQDM. On découvrit ou confirma l'ubiquité de 4 gènes de résistances chez les *Pseudomonas aeruginosa*, soit les β-lactamases AmpC et OXA-50, la chloramphénicol acétyltransférase CatB7 et l'aminoglycoside 3' phosphotransférase APH(3')-IIb. Au moins 125 génomes sur les 390 (32%) possédaient un intégron de classe I et leur contenu en cassettes fut manuellement inspecté. De plus, 41 génomes possédaient au moins une carbapénemase (β-lactamases de type VIM, IMP, KPC, GES ou SPM), des gènes horizontalement acquis et cliniquement préoccupants à cause de l'importance des antibiotiques carbapénems - souvent utilisés en dernier recours lors d'une infection par une souche multirésistante (Papp-Wallace *et al.*, 2011). Avec les corrélations, on s'aperçut que les modifications délétères dans la porine OprD étaient l'une des raisons majeures de la résistance aux carbapénems puisqu'elles étaient plus fréquentes que la présence d'une carbapénemase. Pour la résistance à amikacine, divers éléments génétiques furent identifiés pour causer la résistances, comme les enzymes d'inactivations des aminoglycosides (*ant(4')-IIb*, *aph15*, *aph(3')-VIb*, *aph(6)-Id*, *aac(6')-Ib*, *aac(6')-Iai*, *aac(6')-31*,*aacA29*, *aacA7*) et les pompes d'efflux (MexXY-OprM), mais un nombre important de souches résistantes ne possédaient aucun des ces éléments suggérant l'existence d'autres mécanismes inconnus ou encore reliés à l'expression des gènes. La résistance à lévofloxacine, quant à elle, s'expliquait très bien avec les mutations dans les régions QRDRs des gyrase A et B (gènes *gyrA* et *gyrB*) et de la topoisomérase IV (gènes *parC* et *parE*), telles que décrites au chapitre 1.

Un autre aspect important du projet concernait la phylogénie élargie de l'espèce avec d'autres génomes de références. Pour ce faire, la première étape consistait à trouver le génome essentiel ou tous les gènes étant présents dans tous les génomes de *Pseudomonas aeruginosa* analysés. À partir des 5571 gènes du génome de référence PAO1, chaque homologue était identifié dans les autres génomes lorsque présent. À titre d'exemple, avec un seuil de 99% des génomes (401/405) se devant de posséder le gène, on obtenait un total de 4038 gènes représentatifs du génome essentiel. Cependant, pour la phylogénie finale les critères d'inclusions d'un gène furent plus stricts avec un seuil de présence de 100% et seulement pour les gènes qui étaient assemblés sur leur pleine longueur. Un total de 1278 gènes respectaient donc ces règles, en raison des assemblages parfois incomplets et du manque de profondeur du séquençage Illumina. À noter que 4038 gènes est un chiffre beaucoup plus réaliste pour caractériser le génome essentiel. Par la suite, des alignements multiples indépendants pour chacun de ces gènes étaient

Figure 4.8 – Arbre phylogénétique des *Pseudomonas aeruginosa* de la compagnie AstraZeneca avec plusieurs autres génomes de référence. À noter que la figure ne fait pas partie de l'article, mais fut réalisée dans le cadre du projet.



réalisés avant de les concaténer. Cela donna un imposant alignement multiple caractéristique de la comparaison génomique de tous les génomes. La phylogénie fut ensuite réalisée avec la méthode maximum de vraisemblance (« maximum likelihood ») depuis la version parallèle (MPI) du logiciel RAxML (Stamatakis, 2014). Une centaine d'itérations, avec la méthode d'autoamorçage (« bootstrap »), furent réalisées pour ajouter une confiance statistique à la topologie des noeuds. La figure 4.8 montre la capture d'écran de l'arbre phylogénétique interactif développé à l'aide de la librairie jsPhyloSVG (Smits et Ouverney, 2010) et un peu de programmation Web. L'arc interne de l'arbre illustre le type d'infection causée par la souche alors que l'arc externe dénote son sérotype. On s'aperçoit que les sérotypes regroupent très bien les génomes entre eux, par rapport à la topologie de l'arbre, alors que ce n'est pas le cas pour le type d'infection. Deux groupes importants associés à leur phénotype et reconnus pour

leur multirésistance sont les MDR[5] O11 et O12 (Pirnay *et al.*, 2009) et sont explicitement identifiés sur l'arc externe.

Finalement, les résultats de l'étude furent publiés dans le journal AAC (Kos *et al.*, 2014). L'article démontre toute la puissance qu'offrent les nouvelles générations de séquençages pour l'étude de la résistance aux antibiotiques dans une large cohorte de génomes bactériens et plus spécifiquement pour les *Pseudomonas aeruginosa*. Il met aussi l'accent sur la caractérisation du résistome de l'espèce en lien avec les données expérimentales (CMI) des antibiotiques testés. L'approfondissement des connaissances sur la résistance des *Pseudomonas* qui ressort de l'ouvrage aura un impact important sur le développement futur d'outils diagnostiques et thérapeutiques pour les infections reliées à ce pathogène.

5. *Multidrug Resistant*

Conclusion

La résistance aux antibiotiques est un sujet d'actualité et la conscientisation au problème est maintenant beaucoup mieux ancrée chez les différents intervenants en santé publique comme en font foi toutes les mesures mises en place durant les dernières années. De toutes ces mesures, la génomique a assurément une place prépondérante dans les techniques utilisées puisqu'elle permet la caractérisation des facteurs qui rendent les bactéries résistantes aux traitements par antibiotiques utilisés pour contrer les infections humaines. Cependant, le boom de la génomique a aussi entraîné le besoin de mieux gérer les données produites en masse par les nouvelles générations de séquençages qui affluent dans les bases de données publiques du INSDC. La création de bases de données spécialisées est un processus souvent fastidieux, mais qui rapporte énormément en matière de qualité des analyses génomiques qui en découlent par la suite.

Le projet de recherche, ici présenté, a mis de l'avant de bonnes techniques pour créer une base de données compréhensible et organisée de manière à refléter au mieux l'état actuel des connaissances sur un sujet d'intérêt, la résistance aux antibiotiques. Une bonne maîtrise des mécanismes des antibiotiques et de leurs résistances a été apportée avec comme support une revue substantielle de la littérature sur le sujet. Le travail se veut un reflet de la nomenclature actuelle et généralement acceptée par la communauté scientifique pour les gènes de résistances et certains éléments mobiles qui mènent à leurs disséminations. La création de MERGEM a aussi exploité des concepts innovateurs comme le Web sémantique pour élargir ses horizons et faire partie du mouvement qui est en vogue chez les grandes institutions fournisseuses de données biologiques, tels le EBI et le NCBI. De plus, les différents projets d'annotations des *Pseudomonas aeruginosa* ont contribué a l'avancement des connaissances génomiques sur cet organisme.

L'utilisation de la base de données MERGEM s'est aussi prouvée efficace dans divers projets génomiques. Plusieurs études sur les *Pseudomonas aeruginosa* auront bénéficié de celle-ci (Xiong *et al.* (2013), Déraspe *et al.* (2014), Kos *et al.* (2014)) ainsi qu'un important projet de métagénomique sur la résistance aux antibiotiques du microbiote intestinal humain (projet CQDM). Le développement d'outils comme le pipeline d'annotations est venu complémenter le travail de recherche et permis d'assurer une utilisation efficace de la BD MERGEM dans

les analyses bio-informatiques. En plus du site Web pour MERGEM, une application Web fut développée spécifiquement pour le projet CQDM dans le but de faciliter l'exploration des résultats de manière conviviale par des biologistes ou autres personnes participantes au projet.

Perspectives

La base de données MERGEM, bien qu'intégrale dans son ensemble, pourrait être enrichie avec l'intégration de plusieurs autres ressources de données pertinentes. Le RDF apporte cette flexibilité de modélisation et d'intégration des données qui laisse la porte ouverte à cette perspective pour le futur. Un autre aspect qui pourrait s'avérer intéressant d'approfondir davantage est la mise à jour automatique de MERGEM à l'aide de module de programmation. En effet, il serait possible de synchroniser les différentes sources de données utilisées dans MERGEM, principalement les sites Web des responsables de nomenclature. Des prototypes fonctionnels furent développés durant le cours du projet et ont été utilisés avec succès pour réaliser quelques mises à jour de façon semi-automatique. Cependant, un système générique pour gérer l'intégration automatique des données au sein de MERGEM serait nécessaire pour porter les modules en production. De plus, le système pourrait être synchronisé avec une base de données publique, comme GenBank, question d'avoir toujours une vue en temps réel de la diversité et de l'étendue des gènes de résistances à travers tous les génomes bactériens rendus publique.

D'autre part, la génomique comparative est un domaine qui gagne en importance étant donnée la quantité toujours grandissante de génomes séquencés. Un outil prometteur auquel j'ai contribué au développement est Ray Surveyor, qui fait partie de la suite logicielle Ray, et qui sert à comparer une multitude de génomes entre eux. La comparaison se fait à base de K-mers, d'une longueur prédéfinie, et est très efficace puisqu'il utilise les mêmes technologies que Ray pour paralléliser ses calculs. En plus de pouvoir comparer des centaines voire des milliers de génomes entre eux, Ray Surveyor pourra servir à la création de bases de données spécialisées à base de K-mers. Il sera aussi possible de ressortir des interprétations statistiques encore plus poussées, à l'aide d'algorithme d'apprentissage automatique, basées sur les résultats de Ray Surveyor.

Annexe A

Titre de l'annexe

A.1 Antibiotiques et résistance

A.1.1 Classification des β-lactamines

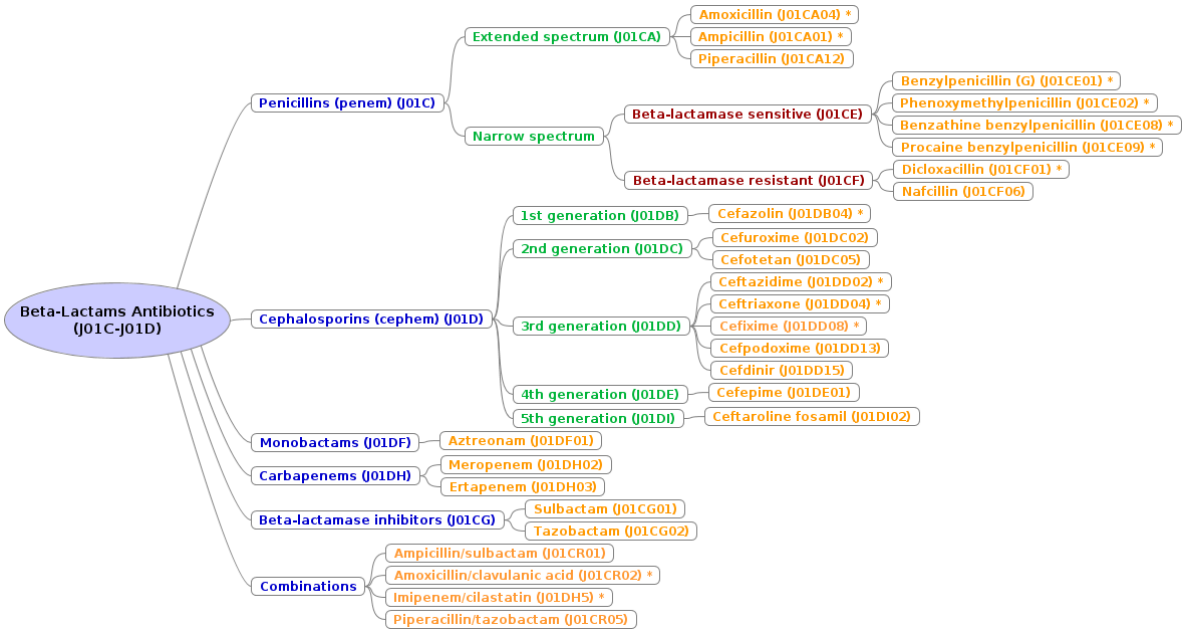


Figure A.1 – Classification des principales antibiotiques β-lactamines utilisées en cliniques dans le guide de The Johns Hopkins Hospital (2014). Les codes de classification ATC (*Anatomical Therapeutic Chemical*) de l'OMS apparaissent entre parenthèses. Les antibiotiques qui sont considérés comme médecine essentiel par l'OMS possèdent une $*$ (World Health Organization (WHO), 2013).

A.1.2 Métabolisme du folate

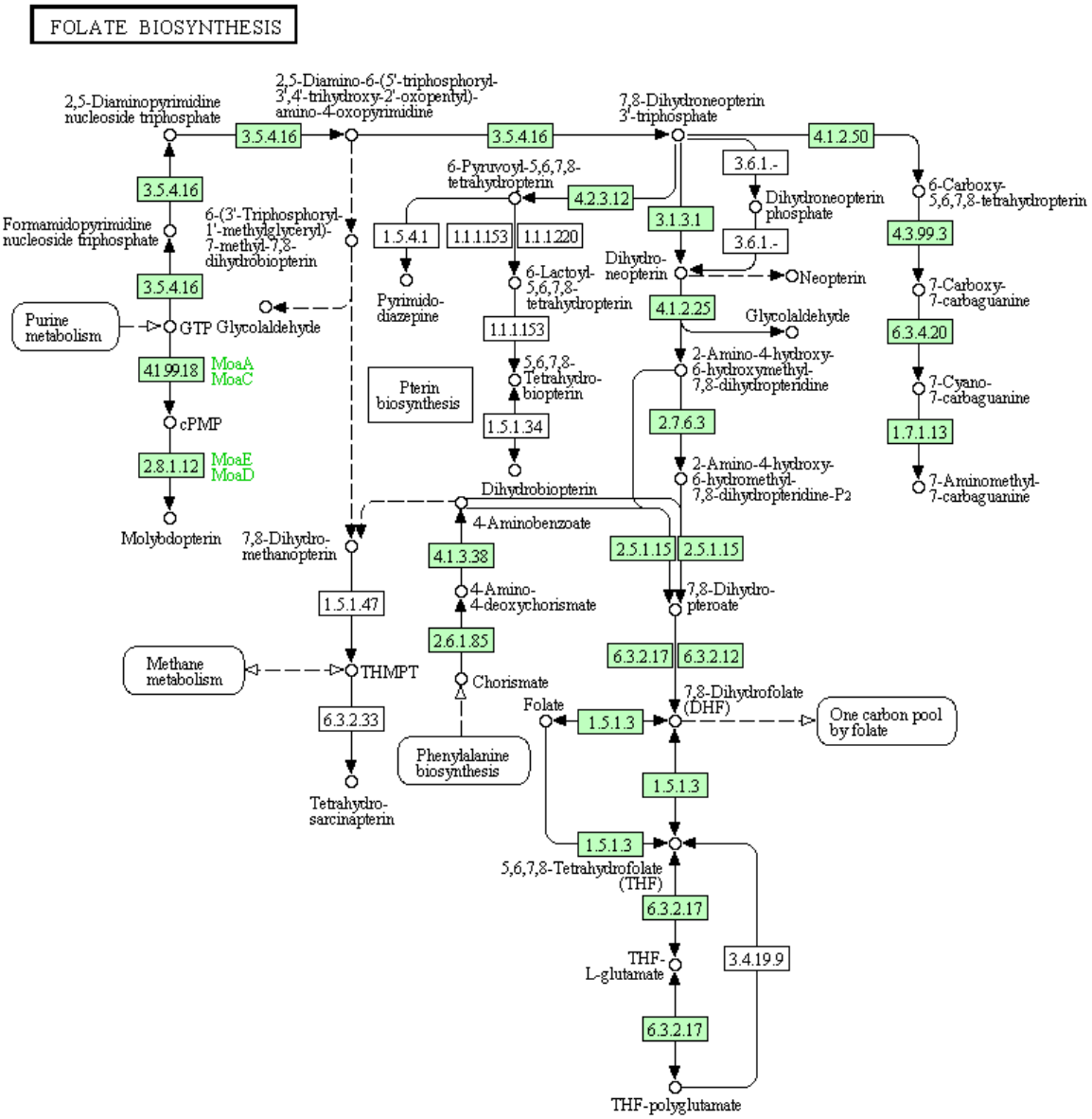


Figure A.2 – Sentier métabolique du folate chez *Escherichia coli* (KEGG : map00790). Utilisation avec la permission de KEGG (Kanehisa et Goto, 2000), (Kanehisa *et al.*, 2014).

A.2 Formats de fichiers RDF

RDF/XML

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:ns0="http://mergem.org/property/" xmlns:ns1
    ="http://www.w3.org/2000/01/rdf-schema#">
  <rdf:Description rdf:type="http://mergem.org/dataSource/" rdf:about="http://mergem.org/dataSource/pubmed">
    <ns0:hasURI rdf:resource="http://www.ncbi.nlm.nih.gov/pubmed"/>
    <ns1:label>Pubmed</ns1:label>
  </rdf:Description>
</rdf:RDF>
```

N-triples

```
<http://mergem.org/dataSource/pubmed> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://mergem.org/
    dataSource/> .
<http://mergem.org/dataSource/pubmed> <http://www.w3.org/2000/01/rdf-schema#label> "Pubmed" .
<http://mergem.org/dataSource/pubmed> <http://mergem.org/property/hasURI> <http://www.ncbi.nlm.nih.gov/pubmed> .
```

N-quads

```
<http://mergem.org/dataSource/pubmed> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://mergem.org/
    dataSource/> <http://graph.mergem.org> .
<http://mergem.org/dataSource/pubmed> <http://www.w3.org/2000/01/rdf-schema#label> "Pubmed" <http://graph.mergem.
    org> .
<http://mergem.org/dataSource/pubmed> <http://mergem.org/property/hasURI> <http://www.ncbi.nlm.nih.gov/pubmed> <
    http://graph.mergem.org> .
```

Turtle

```
<http://mergem.org/dataSource/pubmed> a <http://mergem.org/dataSource/>;
   <http://www.w3.org/2000/01/rdf-schema#label> "Pubmed";
   <http://mergem.org/property/hasURI> <http://www.ncbi.nlm.nih.gov/pubmed>;
```

JSON-LD

```
[
  {
    "@id": "http://mergem.org/dataSource/pubmed",
    "@type": [
      "http://mergem.org/dataSource/"
    ],
    "http://mergem.org/property/hasURI": [
      {
        "@id": "http://www.ncbi.nlm.nih.gov/pubmed"
      }
    ],
    "http://www.w3.org/2000/01/rdf-schema#label": [
      {
        "@value": "Pubmed"
      }
    ]
  }
]
```

A.3 Projet CQDM

A.3.1 Application Web

Figure A.3 – Navigation et listage par patient.

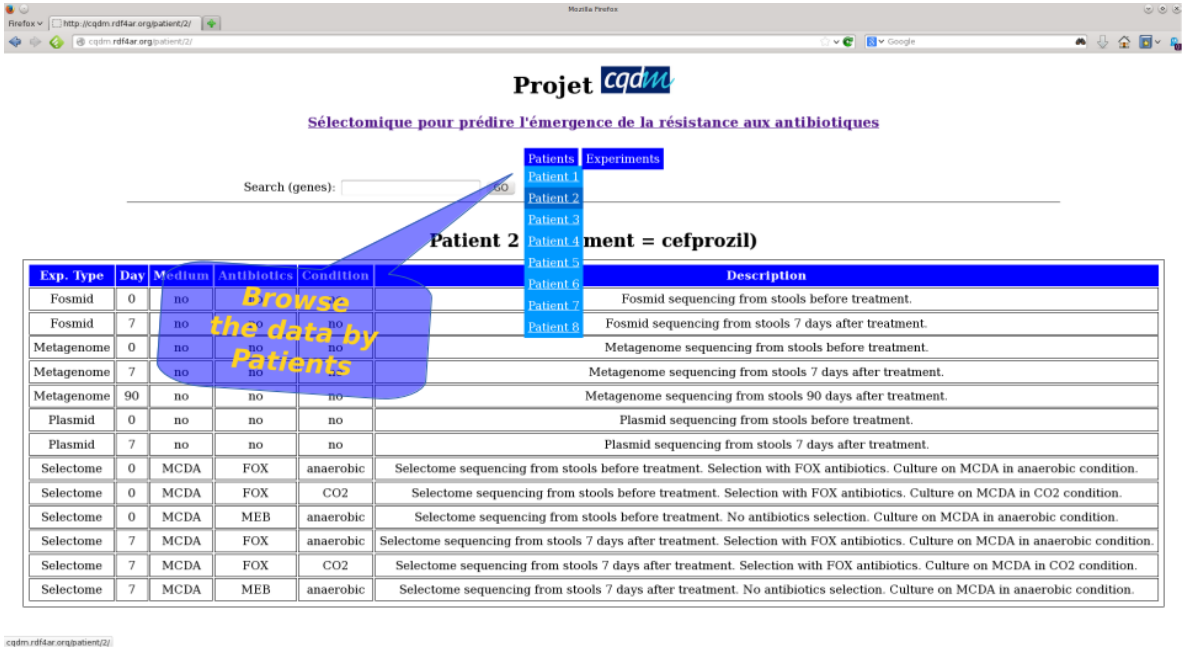


Figure A.4 – Navigation et listage par expérimentation.

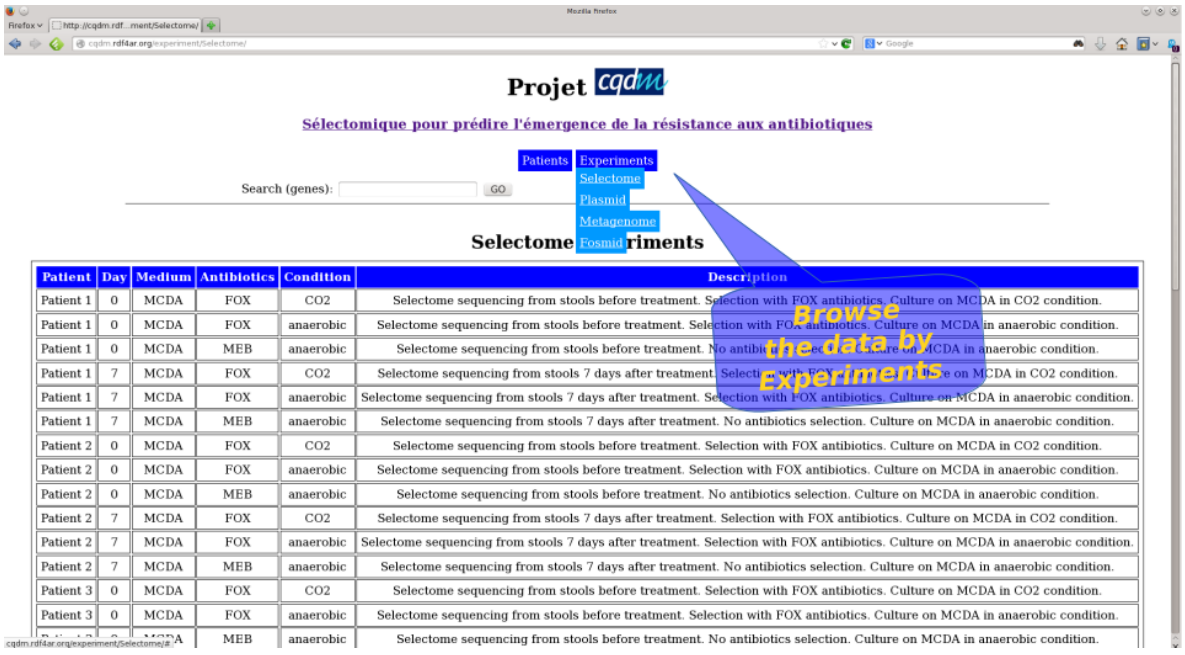


Figure A.5 – Listage des gènes de résistances pour un échantillon donné.

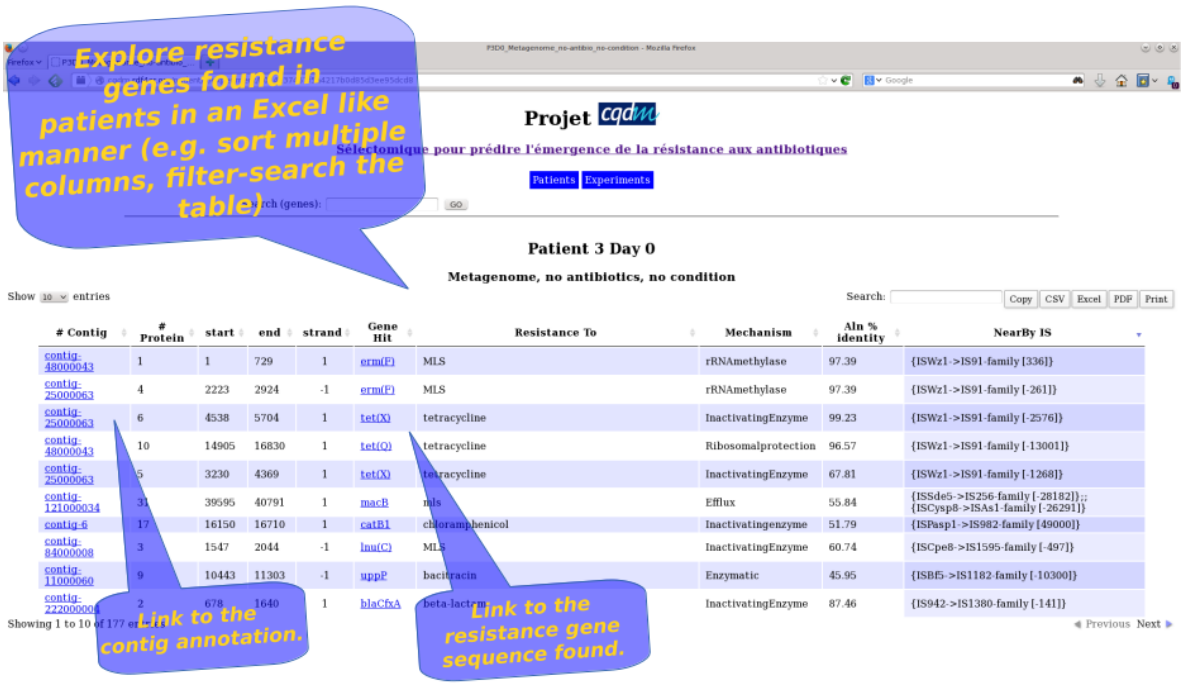
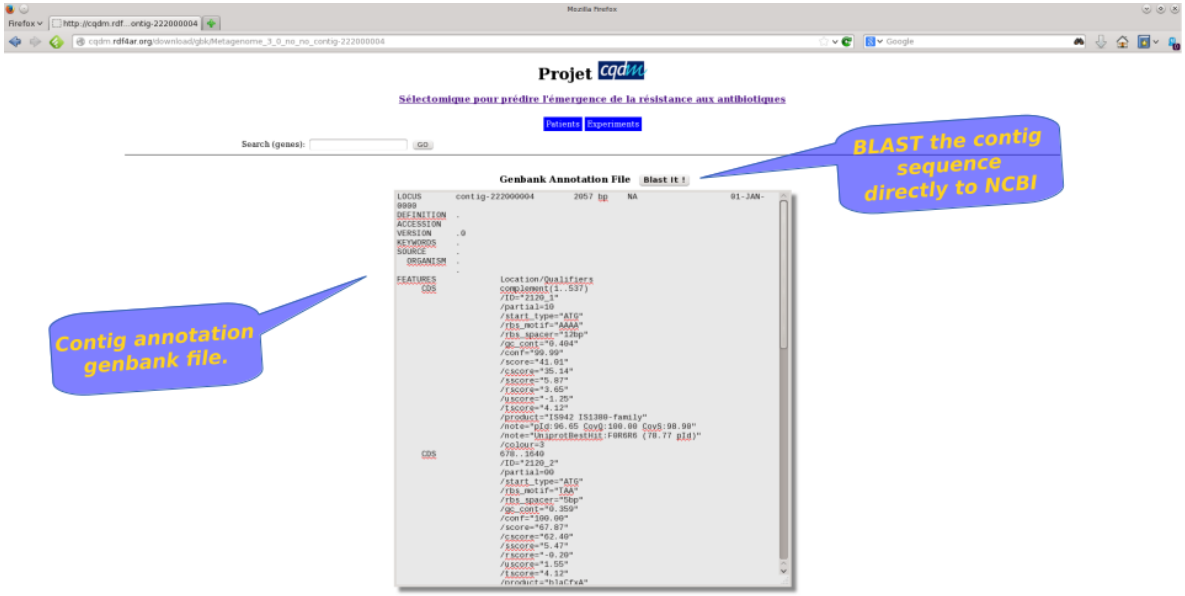


Figure A.6 – Séquences d'un gène de résistance identifié dans un échantillon.

Figure A.7 – Fichier d'annotation GenBank contenant le gène de résistance.



A.4 pOZ176

A.4.1 Fichier d'annotation GenBank de pOZ176 (KC543497)

```
LOCUS       KC543497              500839 bp    DNA     circular BCT 30-SEP-2013
DEFINITION  Pseudomonas aeruginosa plasmid pOZ176, complete sequence.
ACCESSION   KC543497 AY033653 EU886981
VERSION     KC543497.1  GI:496684371
KEYWORDS    .
SOURCE      Pseudomonas aeruginosa
  ORGANISM  Pseudomonas aeruginosa
            Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales;
            Pseudomonadaceae; Pseudomonas.
REFERENCE   1  (bases 1 to 500839)
  AUTHORS   Xiong,J., Alexander,D.C., Ma,J.H., Deraspe,M., Low,D.E.,
            Jamieson,F.B. and Roy,P.H.
  TITLE     Complete Sequence of pOZ176, a 500-Kilobase IncP-2 Plasmid Encoding
            IMP-9-Mediated Carbapenem Resistance, from Outbreak Isolate
            Pseudomonas aeruginosa 96
  JOURNAL   Antimicrob. Agents Chemother. 57 (8), 3775-3782 (2013)
   PUBMED   23716048
REFERENCE   2  (bases 1 to 500839)
  AUTHORS   Xiong,J., Hawkey,P.M. and Roy,P.H.
  TITLE     Novel class 1 integrons on large plasmids in multiresistant
            Pseudomonas aeruginosa isolated from a multicenter survey in
            Guangzhou, PRC
  JOURNAL   Unpublished
REFERENCE   3  (bases 1 to 500839)
  AUTHORS   Xiong,J.H., Hynes,M.F., Ye,H.F., Chen,H.L., M'Zali,F. and
            Hawkey,P.M.
  TITLE     Characterization of blaIMP-9 and Its Association with Large
            Plasmids Carried by Pseudomonas aeruginosa Isolates from the
            People's Republic of China
  JOURNAL   Unpublished
REFERENCE   4  (bases 1 to 500839)
  AUTHORS   Xiong,J., M'Zali,F.H., Chen,H., Ye,H., Lai,F., Wei,Y., Su,D. and
            Hawkey,P.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (30-APR-2001) Microbiology, University of Leeds, Thoresby
            Place, Leeds LS2 9JT, United Kingdom
```

```
REFERENCE   5  (bases 1 to 500839)
  AUTHORS   Xiong,J.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (28-SEP-2005) Division of Immunology and Infection,
            Medical School, University of Birmingham, Edgbaston, Birmingham B15
            2TT, United Kingdom
  REMARK    Sequence update by submitter
REFERENCE   6  (bases 1 to 500839)
  AUTHORS   Xiong,J.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (14-JUL-2008) Infectious Diseases Research Center,
            Universite Laval, Quebec City, Quebec G1V 4G2, Canada
  REMARK    Sequence update by submitter
REFERENCE   7  (bases 1 to 500839)
  AUTHORS   Xiong,J., Hawkey,P.M. and Roy,P.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (15-JUL-2008) Infectious Diseases Research Center,
            Universite Laval, 2705 boul. Laurier, suite RC-709, Quebec, Quebec
            G1V 4G2, Canada
REFERENCE   8  (bases 1 to 500839)
  AUTHORS   Roy,P.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (07-MAY-2013) Infectious Diseases Res. Ctr., Universite
            Laval, 2705 Boul. Laurier, suite RC-709, Quebec, QC G1V 4G2, Canada
COMMENT     On or before Sep 30, 2013 this sequence version replaced
            gi:208436690, gi:194359396.
FEATURES             Location/Qualifiers
     source          1..500839
                     /organism="Pseudomonas aeruginosa"
                     /mol_type="genomic DNA"
                     /strain="96"
                     /db_xref="taxon:287"
                     /plasmid="pOZ176"
     gene            39..1442
                     /locus_tag="pOZ176_001"
                     /note="nuclease-related domain protein, N-terminal
                     truncated"
                     /pseudo
     gene            1435..2991
```

```
                     /locus_tag="pOZ176_002"
     CDS             1435..2991
                     /locus_tag="pOZ176_002"
                     /codon_start=1
                     /transl_table=11
                     /product="hypothetical protein"
                     /protein_id="AGL45968.1"
                     /db_xref="GI:496684372"
                     /translation="MNDEYQAADASGFRICNTISLLVPAYQYQINCAWTKEVSLPAVE
                     EFTCRLLLALQEVLPGEIRDYFGLSKRECDVLIETLIRNKLAVYTNDGHLTPSSMLMD
                     RTKGSSSASPSLTKYEERIERPIFELLTKTIMPPSQHNRTRWGLPQIPVPPESKGWSV
                     LAVADAFGDQYRAFLDFSKLSESETRKTRLYKVGTCDQMAPVNIQVDLEIGLLPTQAG
                     NVEIIKRVAEKVGGTRQRPLSMDLEAKISDYLNSLRMPKDGMSPQEYCQEFKDEVLAR
                     YLDDRGLDINSWLIDHKDRKTGYGNQETRAMIGPLYDNNNRITLGRMLEDLSKDWPEG
                     TIHSALWLSSSVPLWAANGTLLSDFCRKTAEKLSEAPHVKGKITAILPFDDKKEFGQL
                     RSTYHNRIPNGIAFEGSDLQDRFEIFLIPGQLAVVQYHFQPSDDSAATVPIGYITRDP
                     VRVAHIDNFLNSRLSGRGEGFVVWSEDSEKDITNHMEKDRLELIQSSSLGFPMTSQVK
                     LTIRKPPRKW"
[...]

     CDS             complement(457140..457877)
                     /gene="blaIMP-9"
                     /locus_tag="pOZ176_570"
                     /function="beta-lactam hydrolysis"
                     /note="molecular class B beta-lactamase"
                     /codon_start=1
                     /transl_table=11
                     /product="IMP-9 metallo-beta-lactamase"
                     /protein_id="AGL46523.1"
                     /db_xref="GI:496684927"
                     /translation="MSKLFVFFMFLFCSITAAGESLPDLKIEKLDEGVYVHTSFEEVN
                     GWGVIPKHGLVVLVNTDAYLIDTPFTAKDTENLVNWFVERGYRIKGSISSHFHSDSTG
                     GIEWLNSQSIPTYASELTNELLKKDGKVQAKYSFSGVSYWLVKKKIEVFYPGPGHAPD
                     NVVVWLPENRVLFGGCFVKPYGLGNLGDANLEAWPKSAKLLMSKYSKAKLVVPSHSDI
                     GDSSLLKLTWEQTVKGFNESKKSTTAH"
[...]
```

A.4.2 Statistique du fichier d'annotation GenBank de pOZ176 (KC543497)

```
# Général
Nombre de paire de bases: 500839
Nombre de caractéristiques génomique: 1274
Pourcentage GC: 57.89

# Séquences Codante (CDS)
Nombre de CDS: 598
Nombre total de pdb associées aux CDS: 430386
Densité génique: 1.193 gènes par kilo pdb (837 pdb par gène).
Moyenne de pdb par CDS: 719
Pourcentage codant: 85.9

# Résumé des entrées d'annotation
source: 1
misc_feature: 9
-35_signal: 8
CDS: 598
operon: 1
mobile_element: 4
-10_signal: 8
repeat_region: 6
misc_recomb: 23
gene: 615
rep_origin: 1

# hypotetical protein: 396 (66% du total de 599)
```

A.5 PA7-likes

```
# Comparaison des CDS du génome de référence PA7 avec les souches PA7-likes.
Reference PA7 DSM1128 PA5196  PAE413  PAE802  PAE815 PAE832  VRFPAO1  ZW26  ZW26-V2
PSPA7_0001      +       +       +       +       +       +       +       +       +
PSPA7_0002      +       +       +       +       +       +       +       +       +
PSPA7_0003      +       +       +       +       +       +       +       +       +
PSPA7_0004      +       +       +       +       +       +       +       +       +
PSPA7_0005      +       +       +       +       +       +       +       +       +
PSPA7_0006      +       +       +       +       +       +       +       +       +
PSPA7_0007      -       -       -       -       -       -       -       -       -
PSPA7_0008      +       +       +       +       +       +       -       +       +
PSPA7_0009      +       +       +       +       +       +       -       +       +
[...]

# Comparaison des CDS des PA7-likes avec les facteurs de virulence de VFDB
GeneName DSM1128 PA5196 PAE413  PAE802  PAE815  PAE832  ZW26
[...]
exoS    -NA-    -NA-    -NA-    -NA-    -NA-    -NA-    -NA-
exoT    -NA-    -NA-    -NA-    -NA-    -NA-    -NA-    -NA-
exoU    -NA-    -NA-    -NA-    -NA-    -NA-    -NA-    -NA-
exoY    -NA-    -NA-    -NA-    -NA-    -NA-    -NA-    -NA-
toxA    -NA-    -NA-    -NA-    -NA-    -NA-    -NA-    -NA-
[...]
```

Bibliographie

Jeffrey R Aeschlimann : The role of multidrug efflux pumps in the antibiotic resistance of Pseudomonas aeruginosa and other gram-negative bacteria. Insights from the Society of Infectious Diseases Pharmacists. *Pharmacotherapy*, 23(7):916–24, juillet 2003. ISSN 0277-0008. URL http://www.ncbi.nlm.nih.gov/pubmed/12885104.

Hiroyuki Akama, Misa Kanemaki, Masato Yoshimura, Tomitake Tsukihara, Tomoe Kashiwagi, Hiroshi Yoneyama, Shin-ichiro Narita, Atsushi Nakagawa et Taiji Nakae : Crystal structure of the drug discharge outer membrane protein, OprM, of Pseudomonas aeruginosa : dual modes of membrane anchoring and occluded cavity end. *The Journal of biological chemistry*, 279(51):52816–9, décembre 2004. ISSN 0021-9258. URL http://www.ncbi.nlm.nih.gov/pubmed/15507433.

Katie J Aldred, Robert J Kerns et Neil Osheroff : Mechanism of quinolone action and resistance. *Biochemistry*, 53(10):1565–74, mars 2014. ISSN 1520-4995. URL http://www.ncbi.nlm.nih.gov/pubmed/24576155.

R Allmansberger, B Bräu et W Piepersberg : Genes for gentamicin-(3)-N-acetyltransferases III and IV. II. Nucleotide sequences of three AAC(3)-III genes and evolutionary aspects. *Molecular & general genetics : MGG*, 198(3):514–20, janvier 1985. ISSN 0026-8925. URL http://www.ncbi.nlm.nih.gov/pubmed/3892230.

S F Altschul, W Gish, W Miller, E W Myers et D J Lipman : Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–10, octobre 1990. ISSN 0022-2836. URL http://www.ncbi.nlm.nih.gov/pubmed/2231712.

R P Ambler : The structure of beta-lactamases. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 289(1036):321–31, mai 1980. ISSN 0962-8436. URL http://www.ncbi.nlm.nih.gov/pubmed/6109327.

Eric Antezana, Mikel Egaña, Bernard De Baets, Martin Kuiper et Vladimir Mironov : ONTO-PERL : an API for supporting the development and analysis of bio-ontologies. *Bioinformatics (Oxford, England)*, 24(6):885–7, mars 2008. ISSN 1367-4811. URL http://www.ncbi.nlm.nih.gov/pubmed/18245124.

M Ashburner, C J Mungall et S E Lewis : Ontologies for biologists : a community model for the annotation of genomic data. *Cold Spring Harbor symposia on quantitative biology*, 68:227–35, janvier 2003. ISSN 0091-7451. URL http://www.ncbi.nlm.nih.gov/pubmed/15338622.

A Bairoch et R Apweiler : The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic acids research*, 27(1):49–54, janvier 1999. ISSN 0305-1048. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=148094&tool=pmcentrez&rendertype=abstract.

A Bairoch et R Apweiler : The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research*, 28(1):45–8, janvier 2000. ISSN 0305-1048. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102476&tool=pmcentrez&rendertype=abstract.

Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski et Others : SPAdes : a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5):455–477, 2012.

François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault et Jean Morissette : Bio2RDF : towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, 41(5):706–16, octobre 2008. ISSN 1532-0480. URL http://www.ncbi.nlm.nih.gov/pubmed/18472304.

Ronald Bentley et J W Bennett : What is an antibiotic ? Revisited. *Advances in applied microbiology*, 52:303–31, janvier 2003. ISSN 0065-2164. URL http://www.ncbi.nlm.nih.gov/pubmed/12964249.

Tim Berners-Lee, James Hendler, Ora Lassila et Others : The semantic web. *Scientific american*, 284(5):28–37, 2001.

John Besemer et Mark Borodovsky : GeneMark : web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic acids research*, 33(Web Server issue):W451–4, juillet 2005. ISSN 1362-4962. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1160247&tool=pmcentrez&rendertype=abstract.

Christian Bizer, Tom Heath et Tim Berners-Lee : Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, janvier 2009. ISSN 1552-6283. URL http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/jswis.2009081901.

Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J Martin, Karine Michoud, Claire O'Donovan,

Isabelle Phan, Sandrine Pilbout et Michel Schneider : The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research*, 31(1):365–70, janvier 2003. ISSN 1362-4962. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=165542&tool=pmcentrez&rendertype=abstract.

Sébastien Boisvert, François Laviolette et Jacques Corbeil : Ray : simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of computational biology : a journal of computational molecular cell biology*, 17(11):1519–33, novembre 2010. ISSN 1557-8666. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3119603&tool=pmcentrez&rendertype=abstract.

Sébastien Boisvert, Frédéric Raymond, Elénie Godzaridis, François Laviolette et Jacques Corbeil : Ray Meta : scalable de novo metagenome assembly and profiling. *Genome biology*, 13(12):R122, décembre 2012. ISSN 1465-6914. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3119603&tool=pmcentrez&rendertype=abstracthttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4056372&tool=pmcentrez&rendertype=abstract.

Evan Bolton, Gang Fu, Paul Thiessen et Asta Gindulyte : PubChemRDF : Towards a semantic description of PubChem. *In ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY*, volume 246. AMER CHEMICAL SOC 1155 16TH ST, NW, WASHINGTON, DC 20036 USA, 2013.

M H Brown, I T Paulsen et R A Skurray : The multidrug efflux protein NorM is a prototype of a new family of transporters. *Molecular microbiology*, 31(1):394–5, janvier 1999. ISSN 0950-382X. URL http://www.ncbi.nlm.nih.gov/pubmed/9987140.

L E Bryan et J Bedard : Impermeability to quinolones in gram-positive and gram-negative bacteria. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology*, 10(4):232–9, avril 1991. ISSN 0934-9723. URL http://www.ncbi.nlm.nih.gov/pubmed/1864283.

J L Burns, L A Hedin et D M Lien : Chloramphenicol resistance in Pseudomonas cepacia because of decreased permeability. *Antimicrobial agents and chemotherapy*, 33(2):136–41, février 1989. ISSN 0066-4804. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=171444&tool=pmcentrez&rendertype=abstract.

J L Burns, P M Mendelman, J Levy, T L Stull et A L Smith : A permeability barrier as a mechanism of chloramphenicol resistance in Haemophilus influenzae. *Antimicrobial agents and chemotherapy*, 27(1):46–54, janvier 1985. ISSN 0066-4804. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=176203&tool=pmcentrez&rendertype=abstract.

Lori L Burrows : Pseudomonas aeruginosa twitching motility : type IV pili in action. *Annual review of microbiology*, 66(June):493–520, janvier 2012. ISSN 1545-3251. URL http://www.ncbi.nlm.nih.gov/pubmed/22746331.

Karen Bush et George a Jacoby : Updated functional classification of beta-lactamases. *Antimicrobial agents and chemotherapy*, 54(3):969–76, mars 2010. ISSN 1098-6596. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2825993&tool=pmcentrez&rendertype=abstract.

Karen Bush, George A Jacoby et Antone A Medeiros : A functional classification scheme for beta-lactamases and its correlation with molecular structure. *Antimicrobial Agents and Chemotherapy*, 39(6):1211–1233, juin 1995. ISSN 0066-4804. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=162717&tool=pmcentrez&rendertype=abstracthttp://aac.asm.org/cgi/doi/10.1128/AAC.39.6.1211.

CDC : Antibiotic Resistance Threats in the United States, 2013. Rapport technique, U.S. Department of Health and Human Services, 2013.

Geoffrey Chang : Multidrug resistance ABC transporters. *FEBS letters*, 555(1):102–5, novembre 2003. ISSN 0014-5793. URL http://www.ncbi.nlm.nih.gov/pubmed/14630327.

Lei Chen, Haifei Zhang, Ying Chen et Wenping Guo : Blank Nodes in RDF. *Journal of Software*, 7(9):1993–1999, 2012a.

Lihong Chen, Zhaohui Xiong, Lilian Sun, Jian Yang et Qi Jin : VFDB 2012 update : toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic acids research*, 40(Database issue):D641–5, janvier 2012b. ISSN 1362-4962. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245122&tool=pmcentrez&rendertype=abstract.

Lihong Chen, Jian Yang, Jun Yu, Zhijian Yao, Lilian Sun, Yan Shen et Qi Jin : VFDB : a reference database for bacterial virulence factors. *Nucleic acids research*, 33(Database issue):D325–8, janvier 2005. ISSN 1362-4962. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=539962&tool=pmcentrez&rendertype=abstract.

Yen-Ju Chen, Owen Pornillos, Samantha Lieu, Che Ma, Andy P Chen et Geoffrey Chang : X-ray structure of EmrE supports dual topology model. *Proceedings of the National Academy of Sciences of the United States of America*, 104(48):18999–9004, novembre 2007. ISSN 1091-6490. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2141897&tool=pmcentrez&rendertype=abstract.

I Chopra et M Roberts : Tetracycline antibiotics : mode of action, applications, molecular biology, and epidemiology of bacterial resistance. *Microbiology and molecular biology*

reviews : MMBR, 65(2):232–60 ; second page, table of contents, juin 2001. ISSN 1092-2172. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=99026&tool=pmcentrez&rendertype=abstract.

Anne E Clatworthy, Emily Pierson et Deborah T Hung : Targeting virulence : a new paradigm for antimicrobial therapy. *Nature chemical biology*, 3(9):541–8, septembre 2007. ISSN 1552-4450. URL http://www.ncbi.nlm.nih.gov/pubmed/17710100.

Frédéric Collin, Shantanu Karkare et Anthony Maxwell : Exploiting bacterial DNA gyrase as a drug target : current state and perspectives. *Applied microbiology and biotechnology*, 92(3):479–97, novembre 2011. ISSN 1432-0614. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3189412&tool=pmcentrez&rendertype=abstract.

Nicholas R Cozzarelli : UPSIDE : Uniform principle for sharing integral data and materials expeditiously. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3721–2, mars 2004. ISSN 0027-8424. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=374308&tool=pmcentrez&rendertype=abstract.

CQDM : Le réservoir de résistance aux antibiotiques, 2011. URL http://www.cqdm.org/fr/portefeuille-de-projets/projet/7.

Sandra Da Re et Marie-Cécile Ploy : Antibiotiques et réponse SOS bactérienne : Une voie efficace d'acquisition des résistances aux antibiotiques. *MS. Médecine sciences*, 28(2):179–184, 2012.

John Day-Richter : The OBO Flat File Format Specification, version 1.2, 2006. URL http://oboformat.googlecode.com/svn/trunk/doc/GO.format.obo-1_2.html.

Vanessa M D'Costa, Christine E King, Lindsay Kalan, Mariya Morar, Wilson W L Sung, Carsten Schwarz, Duane Froese, Grant Zazula, Fabrice Calmels, Regis Debruyne, G Brian Golding, Hendrik N Poinar et Gerard D Wright : Antibiotic resistance is ancient. *Nature*, 477(7365):457–61, septembre 2011. ISSN 1476-4687. URL http://www.ncbi.nlm.nih.gov/pubmed/21881561.

A L Delcher, D Harmon, S Kasif, O White et S L Salzberg : Improved microbial gene identification with GLIMMER. *Nucleic acids research*, 27(23):4636–41, décembre 1999. ISSN 0305-1048. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=148753&tool=pmcentrez&rendertype=abstract.

Maxime Déraspe, David C Alexander, Jianhui Xiong, Jennifer H Ma, Donald E Low, Frances B Jamieson et Paul H Roy : Genomic analysis of Pseudomonas aeruginosa PA96, the host of carbapenem resistance plasmid pOZ176. *FEMS microbiology letters*, mars 2014. ISSN 1574-6968. URL http://www.ncbi.nlm.nih.gov/pubmed/24673340.

Maxime Déraspe, Elisabet Vilacoba, Daniela Centron et Paul H. Roy : Comparative genomics of PA7-like Pseudomonas aeruginosa, 2013. URL http://www3.unil.ch/wpmu/pseudomonas2013/.

Yohei Doi et Yoshichika Arakawa : 16S ribosomal RNA methylation : emerging resistance mechanism against aminoglycosides. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 45(1):88–94, juillet 2007. ISSN 1537-6591. URL http://www.ncbi.nlm.nih.gov/pubmed/17554708.

Jack A Dunkle, Liqun Xiong, Alexander S Mankin et Jamie H D Cate : Structures of the Escherichia coli ribosome with antibiotics bound near the peptidyl transferase center explain spectra of drug action. *Proceedings of the National Academy of Sciences of the United States of America*, 107(40):17152–7, octobre 2010. ISSN 1091-6490. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2951456&tool=pmcentrez&rendertype=abstract.

M. Dürst et M. Suignard : Internationalized Resource Identifiers (IRIs) (RFC3987), 2005. URL http://www.ietf.org/rfc/rfc3987.txt.

Sylvie Elsen, Philippe Huber, Stéphanie Bouillot, Yohann Couté, Pierre Fournier, Yohann Dubois, Jean-François Timsit, Max Maurin et Ina Attrée : A type III secretion negative clinical strain of Pseudomonas aeruginosa employs a two-partner secreted exolysin to induce hemorrhagic pneumonia. *Cell host & microbe*, 15(2):164–76, février 2014. ISSN 1934-6069. URL http://www.ncbi.nlm.nih.gov/pubmed/24528863.

Elif Eren, Jagamya Vijayaraghavan, Jiaming Liu, Belete R Cheneke, Debra S Touw, Bryan W Lepore, Mridhu Indic, Liviu Movileanu et Bert van den Berg : Substrate specificity within a family of outer membrane carboxylate channels. *PLoS biology*, 10 (1):e1001242, janvier 2012. ISSN 1545-7885. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3260308&tool=pmcentrez&rendertype=abstract.

M Ettayebi, S M Prasad et E A Morgan : Chloramphenicol-erythromycin resistance mutations in a 23S rRNA gene of Escherichia coli. *Journal of bacteriology*, 162(2):551–7, mai 1985. ISSN 0021-9193. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=218883&tool=pmcentrez&rendertype=abstract.

Matthew E Falagas et Sofia K Kasiakou : Colistin : the revival of polymyxins for the management of multidrug-resistant gram-negative bacterial infections. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 40(9):1333–41, mai 2005. ISSN 1537-6591. URL http://www.ncbi.nlm.nih.gov/pubmed/15825037.

a S Fauci : Infectious diseases : considerations for the 21st century. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 32(5):675–85, mars 2001. ISSN 1058-4838. URL http://www.ncbi.nlm.nih.gov/pubmed/11229834.

K R Fiebelkorn, S A Crawford et J H Jorgensen : Mutations in folP associated with elevated sulfonamide MICs for Neisseria meningitidis clinical isolates from five continents. *Antimicrobial agents and chemotherapy*, 49(2):536–40, février 2005. ISSN 0066-4804. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=547345&tool=pmcentrez&rendertype=abstract.

V A Fischetti et American Society for Microbiology : *Gram-positive Pathogens*. American Society Mic Series. ASM Press, 2006. ISBN 9781555813437. URL http://books.google.ca/books?id=DUX1G0ddWlAC.

Laura S Frost, Raphael Leplae, Anne O Summers et Ariane Toussaint : Mobile genetic elements : the agents of open source evolution. *Nature reviews. Microbiology*, 3(9):722–32, septembre 2005. ISSN 1740-1526. URL http://www.ncbi.nlm.nih.gov/pubmed/16138100.

F Garnier, S Taourit, P Glaser, P Courvalin et M Galimand : Characterization of transposon Tn1549, conferring VanB-type resistance in Enterococcus spp. *Microbiology (Reading, England)*, 146 (Pt 6:1481–9, juin 2000. ISSN 1350-0872. URL http://www.ncbi.nlm.nih.gov/pubmed/10846226.

S Goussard, W Sougakoff, C Mabilat, A Bauernfeind et P Courvalin : An IS1-like element is responsible for high-level synthesis of extended-spectrum beta-lactamase TEM-6 in Enterobacteriaceae. *Journal of general microbiology*, 137(12):2681–7, décembre 1991. ISSN 0022-1287. URL http://www.ncbi.nlm.nih.gov/pubmed/1665171.

Ray Greek : The Discovery and Development of Penicillin. *ACS Commemorative Booklet*, 1999. URL http://www.acs.org/content/acs/en/education/whatischemistry/landmarks/flemingpenicillin.html.

Barry G Hall et Miriam Barlow : Revised Ambler classification of {beta}-lactamases. *The Journal of antimicrobial chemotherapy*, 55(6):1050–1, juin 2005. ISSN 0305-7453. URL http://jac.oxfordjournals.org/cgi/content/long/55/6/1050.

Holly Heaslet, Melissa Harris, Kelly Fahnoe, Ronald Sarver, Henry Putz, Jeanne Chang, Chakrapani Subramanyam, Gabriela Barreiro et J Richard Miller : Structural comparison of chromosomal and exogenous dihydrofolate reductase from Staphylococcus aureus in complex with the potent inhibitor trimethoprim. *Proteins*, 76(3):706–17, août 2009. ISSN 1097-0134. URL http://www.ncbi.nlm.nih.gov/pubmed/19280600.

Tom Heath et Christian Bizer : Linked Data : Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web : Theory and Technology*, 1(1):1–136, février 2011. ISSN 2160-4711. URL http://www.morganclaypool.com/doi/abs/10.2200/S00334ED1V01Y201102WBE001.

Weiling Hong, Jie Zeng et Jianping Xie : Antibiotic drugs targeting bacterial RNAs. *Acta Pharmaceutica Sinica B*, 4(4):258–265, août 2014. ISSN 22113835. URL http://www.sciencedirect.com/science/article/pii/S2211383514000641.

H Huang et R E Hancock : The role of specific surface loop regions in determining the function of the imipenem-specific pore protein OprD of Pseudomonas aeruginosa. *Journal of bacteriology*, 178(11):3085–90, juin 1996. ISSN 0021-9193. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=178056&tool=pmcentrez&rendertype=abstract.

Jennifer R Huddleston : Horizontal gene transfer in the human gastrointestinal tract : potential spread of antibiotic resistance genes. *Infection and drug resistance*, 7:167–76, janvier 2014. ISSN 1178-6973. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4073975&tool=pmcentrez&rendertype=abstract.

P Huovinen : Resistance to trimethoprim-sulfamethoxazole. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 32(11):1608–14, juin 2001. ISSN 1058-4838. URL http://www.ncbi.nlm.nih.gov/pubmed/11340533.

P Huovinen, L Sundström, G Swedberg et O Sköld : Trimethoprim and sulfonamide resistance. *Antimicrobial agents and chemotherapy*, 39(2):279–89, février 1995. ISSN 0066-4804. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=162528&tool=pmcentrez&rendertype=abstract.

Doug Hyatt, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer et Loren J Hauser : Prodigal : prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11:119, janvier 2010. ISSN 1471-2105. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2848648&tool=pmcentrez&rendertype=abstract.

Illumina : De Novo Assembly Using Illumina Reads. Rapport technique, 2010. URL http://res.illumina.com/documents/products/technotes/technote_denovo_assembly_ecoli.pdf.

IWG-SCC : Classification of staphylococcal cassette chromosome mec (SCCmec) : guidelines for reporting novel SCCmec elements. *Antimicrobial agents and chemotherapy*, 53 (12):4961–7, décembre 2009. ISSN 1098-6596. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2786320&tool=pmcentrez&rendertype=abstract.

George Jacoby, Vincent Cattoir, David Hooper, Luis Martínez-Martínez, Patrice Nordmann, Alvaro Pascual, Laurent Poirel et Minggui Wang : qnr Gene nomenclature. *Antimicrobial agents and chemotherapy*, 52(7):2297–9, juillet 2008. ISSN 1098-6596. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2443900&tool=pmcentrez&rendertype=abstract.

Mario Juhas : Horizontal gene transfer in human pathogens. *Critical reviews in microbiology*, juillet 2013. ISSN 1549-7828. URL http://www.ncbi.nlm.nih.gov/pubmed/23862575.

Simon Jupp, James Malone, Jerven Bolleman, Marco Brandizi, Mark Davies, Leyla Garcia, Anna Gaulton, Sebastien Gehant, Camille Laibe, Nicole Redaschi, Sarala M Wimalaratne, Maria Martin, Nicolas Le Novère, Helen Parkinson, Ewan Birney et Andrew M Jenkinson : The EBI RDF platform : linked open data for the life sciences. *Bioinformatics (Oxford, England)*, 30(9):1338–9, mai 2014. ISSN 1367-4811. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3998127&tool=pmcentrez&rendertype=abstract.

M Kanehisa et S Goto : KEGG : kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, janvier 2000. ISSN 0305-1048. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409&tool=pmcentrez&rendertype=abstract.

Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi et Mao Tanabe : Data, information, knowledge and principle : back to metabolism in KEGG. *Nucleic acids research*, 42(Database issue):D199–205, janvier 2014. ISSN 1362-4962. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965122&tool=pmcentrez&rendertype=abstract.

Kristie M Keeney, Sophie Yurist-Doutsch, Marie-Claire Arrieta et B Brett Finlay : Effects of antibiotics on human microbiota and subsequent disease. *Annual review of microbiology*, 68:217–35, janvier 2014. ISSN 1545-3251. URL http://www.ncbi.nlm.nih.gov/pubmed/24995874.

Veronica N Kos, Maxime Déraspe, Robert E McLaughlin, James D Whiteaker, Paul H Roy, Richard A Alm, Jacques Corbeil et Humphrey Gardner : The resistome of Pseudomonas aeruginosa in relationship to phenotypic susceptibility. *Antimicrobial agents and chemotherapy*, novembre 2014. ISSN 1098-6596. URL http://www.ncbi.nlm.nih.gov/pubmed/25367914.

Ivan Laponogov, Maninder K Sohi, Dennis A Veselkov, Xiao-Su Pan, Ritica Sawhney, Andrew W Thompson, Katherine E McAuley, L Mark Fisher et Mark R Sanderson : Structural insight into the quinolone-DNA cleavage complex of type IIA topoisomerases. *Nature structural & molecular biology*, 16(6):667–9, juin 2009. ISSN 1545-9985. URL http://www.ncbi.nlm.nih.gov/pubmed/19448616.

Tim-Berner Lee : Linked Data - Design Issues, 2006. URL http://www.w3.org/DesignIssues/LinkedData.html.

Hanna-Kirsti S Leiros, Pardha S Borra, Bjørn Olav Brandsdal, Kine Susann Waade Edvardsen, James Spencer, Timothy R Walsh et Orjan Samuelsen : Crystal Structure

of the Mobile Metallo-β-Lactamase AIM-1 from Pseudomonas aeruginosa : Insights into Antibiotic Binding and the Role of Gln157. *Antimicrobial agents and chemotherapy*, 56 (8):4341–53, août 2012. ISSN 1098-6596. URL http://www.ncbi.nlm.nih.gov/pubmed/22664968.

Stuart B Levy : *Le paradoxe des antibiotiques. Comment le miracle tue le miracle.* BELIN, belin édition, 1999. ISBN 2-7011-2407-7.

Stuart B Levy, Laura M Mcmurry, Teresa M Barbosa, Vickers Burdett, Patrice Courvalin, Wolfgang Hillen, C Roberts, Julian I Rood, Diane E Taylor, Laura M M C Murry et Marilyn C Roberts : Nomenclature for New Tetracycline Resistance Determinants. pages 1523–1525, 1999.

Kim Lewis : Platforms for antibiotic discovery. *Nature reviews. Drug discovery*, 12(5):371–87, mai 2013. ISSN 1474-1784. URL http://www.ncbi.nlm.nih.gov/pubmed/23629505.

Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen et Jun Wang : SOAP2 : an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, 2009.

Bo Liu et Mihai Pop : ARDB–Antibiotic Resistance Genes Database. *Nucleic acids research*, 37(Database issue):D443–7, janvier 2009. ISSN 1362-4962. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686595&tool=pmcentrez&rendertype=abstract.

Pauline Macheboeuf, Carlos Contreras-Martel, Viviana Job, Otto Dideberg et Andréa Dessen : Penicillin binding proteins : key players in bacterial cell cycle and drug resistance processes. *FEMS microbiology reviews*, 30(5):673–91, septembre 2006. ISSN 0168-6445. URL http://www.ncbi.nlm.nih.gov/pubmed/16911039.

Y Masaoka, Y Ueno, Y Morita, T Kuroda, T Mizushima et T Tsuchiya : A two-component multidrug efflux pump, EbrAB, in Bacillus subtilis. *Journal of bacteriology*, 182(8):2307–10, avril 2000. ISSN 0021-9193. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=111282&tool=pmcentrez&rendertype=abstract.

Kalai Mathee, Giri Narasimhan, Camilo Valdes, Xiaoyun Qiu, Jody M Matewish, Michael Koehrsen, Antonis Rokas, Chandri N Yandava, Reinhard Engels, Erliang Zeng, Raquel Olavarietta, Melissa Doud, Roger S Smith, Philip Montgomery, Jared R White, Paul a Godfrey, Chinnappa Kodira, Bruce Birren, James E Galagan et Stephen Lory : Dynamics of Pseudomonas aeruginosa genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 105(8):3100–5, février 2008. ISSN 1091-6490. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2268591&tool=pmcentrez&rendertype=abstract.

Andrew G McArthur, Nicholas Waglechner, Fazmin Nizam, Austin Yan, Marisa a Azad, Alison J Baylay, Kirandeep Bhullar, Marc J Canova, Gianfranco De Pascale, Linda Ejim, Lindsay Kalan, Andrew M King, Kalinka Koteva, Mariya Morar, Michael R Mulvey, Jonathan S O'Brien, Andrew C Pawlowski, Laura J V Piddock, Peter Spanogiannopoulos, Arlene D Sutherland, Irene Tang, Patricia L Taylor, Maulik Thaker, Wenliang Wang, Marie Yan, Tennison Yu et Gerard D Wright : The Comprehensive Antibiotic Resistance Database. *Antimicrobial agents and chemotherapy*, 57(7):1–11, mai 2013. ISSN 1098-6596. URL http://www.ncbi.nlm.nih.gov/pubmed/23650175.

John D McPherson : A defining decade in DNA sequencing. *Nature methods*, 11(10):1003–5, octobre 2014. ISSN 1548-7105. URL http://www.ncbi.nlm.nih.gov/pubmed/25264775.

Michael L Metzker : Emerging technologies in DNA sequencing. *Genome research*, 15(12): 1767–76, décembre 2005. ISSN 1088-9051. URL http://www.ncbi.nlm.nih.gov/pubmed/16339375.

R H Mosher, D J Camp, K Yang, M P Brown, W V Shaw et L C Vining : Inactivation of chloramphenicol by O-phosphorylation. A novel resistance mechanism in Streptomyces venezuelae ISP5230, a chloramphenicol producer. *The Journal of biological chemistry*, 270(45):27000–6, novembre 1995. ISSN 0021-9258. URL http://www.ncbi.nlm.nih.gov/pubmed/7592948.

Alexandra Moura, Mário Soares, Carolina Pereira, Nuno Leitão, Isabel Henriques et António Correia : INTEGRALL : a database and search engine for integrons, integrases and gene cassettes. *Bioinformatics (Oxford, England)*, 25(8):1096–8, avril 2009. ISSN 1367-4811. URL http://www.ncbi.nlm.nih.gov/pubmed/19228805.

Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne Storey, Christopher G Chute et Mark A Musen : BioPortal : ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, 37(Web Server issue):W170–3, juillet 2009. ISSN 1362-4962. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2703982&tool=pmcentrez&rendertype=abstract.

Barack Obama : Executive Order – Combating Antibiotic-Resistant Bacteria. Rapport technique, Washington, 2014. URL http://www.whitehouse.gov/the-press-office/2014/09/18/executive-order-combating-antibiotic-resistant-bacteria.

G a Pankey et L D Sabath : Clinical relevance of bacteriostatic versus bactericidal mechanisms of action in the treatment of Gram-positive bacterial infections. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 38(6):864–70, mars 2004. ISSN 1537-6591. URL http://www.ncbi.nlm.nih.gov/pubmed/14999632.

Krisztina M Papp-Wallace, Andrea Endimiani, Magdalena A Taracila et Robert A Bonomo : Carbapenems : past, present, and future. *Antimicrobial agents and chemotherapy*, 55(11):4943–60, novembre 2011. ISSN 1098-6596. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3195018&tool=pmcentrez&rendertype=abstract.

David L Paterson et Robert A Bonomo : Extended-spectrum beta-lactamases : a clinical update. *Clinical microbiology reviews*, 18(4):657–86, octobre 2005. ISSN 0893-8512. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1265908&tool=pmcentrez&rendertype=abstract.

W R Pearson : Flexible sequence similarity searching with the FASTA3 program package. *Methods in molecular biology (Clifton, N.J.)*, 132:185–219, janvier 2000. ISSN 1064-3745. URL http://www.ncbi.nlm.nih.gov/pubmed/10547837.

Johann Peltier, Pascal Courtin, Imane El Meouche, Manuella Catel-Ferreira, Marie-Pierre Chapot-Chartier, Ludovic Lemée et Jean-Louis Pons : Genomic and expression analysis of the vanG-like gene cluster of Clostridium difficile. *Microbiology (Reading, England)*, 159(Pt 7):1510–20, juillet 2013. ISSN 1465-2080. URL http://www.ncbi.nlm.nih.gov/pubmed/23676437.

Jean-Paul Pirnay, Florence Bilocq, Bruno Pot, Pierre Cornelis, Martin Zizi, Johan Van Eldere, Pieter Deschaght, Mario Vaneechoutte, Serge Jennes, Tyrone Pitt et Daniel De Vos : Pseudomonas aeruginosa population structure revisited. *PloS one*, 4(11):e7740, janvier 2009. ISSN 1932-6203. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2777410&tool=pmcentrez&rendertype=abstract.

Hester Plumridge : Drug Makers Tiptoe Back Into Antibiotic R&D. *The Wall Street Journal*, janvier 2014. URL http://online.wsj.com/news/articles/SB10001424052702303465004579322601579895822.

Gerrit J Poelarends, Piotr Mazurkiewicz et Wil N Konings : Multidrug transporters and antibiotic resistance in Lactococcus lactis. *Biochimica et biophysica acta*, 1555(1-3):1–7, septembre 2002. ISSN 0006-3002. URL http://www.ncbi.nlm.nih.gov/pubmed/12206883.

Keith Poole : Efflux-mediated antimicrobial resistance. *The Journal of antimicrobial chemotherapy*, 56(1):20–51, juillet 2005. ISSN 0305-7453. URL http://www.ncbi.nlm.nih.gov/pubmed/15914491.

Steven J Projan : Why is big Pharma getting out of antibacterial drug discovery ? *Current Opinion in Microbiology*, 6(5):427–430, octobre 2003. ISSN 13695274. URL http://linkinghub.elsevier.com/retrieve/pii/S1369527403001097.

M Putman, H W van Veen et W N Konings : Molecular properties of bacterial multidrug transporters. *Microbiology and molecular biology reviews : MMBR*, 64(4):672–93, décembre

2000. ISSN 1092-2172. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=99009&tool=pmcentrez&rendertype=abstract.

Sara Reardon : Phage therapy gets revitalized. *Nature*, 510(7503):15–6, juin 2014. ISSN 1476-4687. URL http://www.ncbi.nlm.nih.gov/pubmed/24899282.

M I Recht et J D Puglisi : Aminoglycoside resistance with homogeneous and heterogeneous populations of antibiotic-resistant ribosomes. *Antimicrobial agents and chemotherapy*, 45 (9):2414–9, septembre 2001. ISSN 0066-4804. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=90670&tool=pmcentrez&rendertype=abstract.

Nicole Redaschi, UniProt Consortium et Others : Uniprot in RDF : Tackling data integration and distributed annotation with the semantic web. 2009.

Adam P Roberts, Michael Chandler, Patrice Courvalin, Gérard Guédon, Peter Mullany, Tony Pembroke, Julian I Rood, C Jeffery Smith, Anne O Summers, Masataka Tsuda et Douglas E Berg : Revised nomenclature for transposable genetic elements. *Plasmid*, 60(3):167–73, novembre 2008. ISSN 1095-9890. URL http://www.ncbi.nlm.nih.gov/pubmed/18778731.

Marilyn C Roberts : Update on macrolide-lincosamide-streptogramin, ketolide, and oxazolidinone resistance genes. *FEMS microbiology letters*, 282(2):147–59, mai 2008. ISSN 0378-1097. URL http://www.ncbi.nlm.nih.gov/pubmed/18399991.

Marilyn C Roberts, Joyce Sutcliffe, Patrice Courvalin, Lars Bogo Jensen, Julian Rood et Helena Seppala : MINIREVIEW Nomenclature for Macrolide and Macrolide-Lincosamide- Streptogramin B Resistance Determinants. 43(12):2823–2830, 1999.

Ari Robicsek, Jacob Strahilevitz, George A Jacoby, Mark Macielag, Darren Abbanat, Chi Hye Park, Karen Bush et David C Hooper : Fluoroquinolone-modifying enzyme : a new adaptation of a common aminoglycoside acetyltransferase. *Nature medicine*, 12(1):83–8, janvier 2006. ISSN 1078-8956. URL http://www.ncbi.nlm.nih.gov/pubmed/16369542.

J I Ross, E A Eady, J H Cove et W J Cunliffe : 16S rRNA mutation associated with tetracycline resistance in a gram-positive bacterium. *Antimicrobial agents and chemotherapy*, 42(7):1702–5, juillet 1998. ISSN 0066-4804. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=105669&tool=pmcentrez&rendertype=abstract.

Paul H Roy, Sasha G Tetu, André Larouche, Liam Elbourne, Simon Tremblay, Qinghu Ren, Robert Dodson, Derek Harkins, Ryan Shay, Kisha Watkins, Yasmin Mahamoud et Ian T Paulsen : PA7 - Complete genome sequence of the multiresistant taxonomic outlier Pseudomonas aeruginosa PA7. *PloS one*, 5(1):e8842, janvier 2010. ISSN 1932-6203. URL http://www.ncbi.nlm.nih.gov/pubmed/20107499.

K Rutherford, J Parkhill, J Crook, T Horsnell, P Rice, M A Rajandream et B Barrell : Artemis : sequence visualization and annotation. *Bioinformatics (Oxford, England)*, 16(10):944–5, octobre 2000. ISSN 1367-4803. URL http://www.ncbi.nlm.nih.gov/pubmed/11120685.

Abigail A Salyers, Anamika Gupta et Yanping Wang : Human intestinal bacteria as reservoirs for antibiotic resistance genes. *Trends in microbiology*, 12(9):412–6, septembre 2004. ISSN 0966-842X. URL http://www.ncbi.nlm.nih.gov/pubmed/15337162.

Eric Sauvage, Frédéric Kerff, Mohammed Terrak, Juan A Ayala et Paulette Charlier : The penicillin-binding proteins : structure and role in peptidoglycan biosynthesis. *FEMS microbiology reviews*, 32(2):234–58, mars 2008. ISSN 0168-6445. URL http://www.ncbi.nlm.nih.gov/pubmed/18266856.

Alexandra M Schnoes, Shoshana D Brown, Igor Dodevski et Patricia C Babbitt : Annotation error in public databases : misannotation of molecular function in enzyme superfamilies. *PLoS computational biology*, 5(12):e1000605, décembre 2009. ISSN 1553-7358. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2781113&tool=pmcentrez&rendertype=abstract.

Stefan Schwarz, Corinna Kehrenberg, Benoît Doublet et Axel Cloeckaert : Molecular basis of bacterial resistance to chloramphenicol and florfenicol. *FEMS microbiology reviews*, 28(5):519–42, novembre 2004. ISSN 0168-6445. URL http://www.ncbi.nlm.nih.gov/pubmed/15539072.

K J Shaw, P N Rather, R S Hare et G H Miller : Molecular genetics of aminoglycoside resistance genes and familial relationships of the aminoglycoside-modifying enzymes. *Microbiological reviews*, 57(1):138–63, mars 1993. ISSN 0146-0749. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=372903&tool=pmcentrez&rendertype=abstract.

P Siguier, J Perochon, L Lestrade, J Mahillon et M Chandler : ISfinder : the reference centre for bacterial insertion sequences. *Nucleic acids research*, 34(Database issue):D32–6, janvier 2006a. ISSN 1362-4962. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1347377&tool=pmcentrez&rendertype=abstract.

Patricia Siguier, Jonathan Filée et Michael Chandler : Insertion sequences in prokaryotic genomes. *Current opinion in microbiology*, 9(5):526–31, octobre 2006b. ISSN 1369-5274. URL http://www.ncbi.nlm.nih.gov/pubmed/16935554.

Lynn L Silver : Challenges of antibacterial discovery. *Clinical microbiology reviews*, 24 (1):71–109, janvier 2011. ISSN 1098-6618. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3021209&tool=pmcentrez&rendertype=abstract.

Jared T Simpson, Kim Wong, Shaun D Jackman, Jacqueline E Schein, Steven J M Jones et Inanç Birol : ABySS : a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–1123, 2009.

Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel et Suzanna Lewis : The OBO Foundry : coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–5, novembre 2007. ISSN 1087-0156. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2814061&tool=pmcentrez&rendertype=abstract.

Samuel A Smits et Cleber C Ouverney : jsPhyloSVG : a javascript library for visualizing interactive and vector-based phylogenetic trees on the web. *PloS one*, 5(8):e12267, janvier 2010. ISSN 1932-6203. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2923619&tool=pmcentrez&rendertype=abstract.

Morten O a Sommer, Gautam Dantas et George M Church : Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science (New York, N.Y.)*, 325(5944):1128–31, août 2009. ISSN 1095-9203. URL http://www.ncbi.nlm.nih.gov/pubmed/19713526.

Brad Spellberg : Dr. William H. Stewart : mistaken or maligned ? *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 47(2):294, juillet 2008. ISSN 1537-6591. URL http://www.ncbi.nlm.nih.gov/pubmed/18564938.

Alexandros Stamatakis : RAxML Version 8 : A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics (Oxford, England)*, pages 2010–2011, janvier 2014. ISSN 1367-4811. URL http://www.ncbi.nlm.nih.gov/pubmed/24451623.

Judith N Steenbergen, Jeff Alder, Grace M Thorne et Francis P Tally : Daptomycin : a lipopeptide antibiotic for the treatment of serious Gram-positive infections. *The Journal of antimicrobial chemotherapy*, 55(3):283–8, mars 2005. ISSN 0305-7453. URL http://www.ncbi.nlm.nih.gov/pubmed/15705644.

Lewis Stewart, Amy Ford, Vartul Sangal, Julie Jeukens, Brian Boyle, Iréna Kukavica-Ibrulj, Shabhonam Caim, Lisa Crossman, Paul A Hoskisson, Roger Levesque et Nicholas P Tucker : Draft genomes of 12 host-adapted and environmental isolates of Pseudomonas aeruginosa and their positions in the core genome phylogeny. *Pathogens and disease*, 71(1):20–5, juin 2014. ISSN 2049-632X. URL http://www.ncbi.nlm.nih.gov/pubmed/24167005.

Paul Stothard et David S Wishart : Circular genome visualization and exploration using CGView. *Bioinformatics (Oxford, England)*, 21(4):537–9, février 2005. ISSN 1367-4803. URL http://www.ncbi.nlm.nih.gov/pubmed/15479716.

S M Swaney, H Aoki, M C Ganoza et D L Shinabarger : The oxazolidinone linezolid inhibits initiation of protein synthesis in bacteria. *Antimicrobial agents and chemotherapy*, 42(12):3251–5, décembre 1998. ISSN 0066-4804. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=106030&tool=pmcentrez&rendertype=abstract.

Adrian Tett, Andrew J Spiers, Lisa C Crossman, Duane Ager, Lena Ciric, J Maxwell Dow, John C Fry, David Harris, Andrew Lilley, Anna Oliver, Julian Parkhill, Michael A Quail, Paul B Rainey, Nigel J Saunders, Kathy Seeger, Lori A S Snyder, Rob Squares, Christopher M Thomas, Sarah L Turner, Xue-Xian Zhang, Dawn Field et Mark J Bailey : Sequence-based analysis of pQBR103 ; a representative of a unique, transfer-proficient mega plasmid resident in the microbial community of sugar beet. *The ISME journal*, 1(4):331–40, août 2007. ISSN 1751-7362. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2656933&tool=pmcentrez&rendertype=abstract.

The Johns Hopkins Hospital : *Antibiotic Guidelines 2014-2015 - Treatment Recommendations For Adult Inpatients*. Johns Hopkins Medecine, 2014-2015 édition, 2014. URL http://www.hopkinsmedicine.org/amp/guidelines/Antibiotic_guidelines.pdf.

C S Toro, S R Lobos, I Calderón, M Rodríguez et G C Mora : Clinical isolate of a porinless Salmonella typhi resistant to high levels of chloramphenicol. *Antimicrobial agents and chemotherapy*, 34(9):1715–9, septembre 1990. ISSN 0066-4804. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=171911&tool=pmcentrez&rendertype=abstract.

John H Tran, George A Jacoby et David C Hooper : Interaction of the plasmid-encoded quinolone resistance protein Qnr with Escherichia coli DNA gyrase. *Antimicrobial agents and chemotherapy*, 49(1):118–25, janvier 2005. ISSN 0066-4804. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=538914&tool=pmcentrez&rendertype=abstract.

Angela H a M van Hoek, Dik Mevius, Beatriz Guerra, Peter Mullany, Adam Paul Roberts et Henk J M Aarts : Acquired antibiotic resistance genes : an overview. *Frontiers in microbiology*, 2(September):203, janvier 2011. ISSN 1664-302X. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3202223&tool=pmcentrez&rendertype=abstract.

Waldemar Vollmer, Didier Blanot et Miguel A de Pedro : Peptidoglycan structure and architecture. *FEMS microbiology reviews*, 32(2):149–67, mars 2008. ISSN 0168-6445. URL http://www.ncbi.nlm.nih.gov/pubmed/18194336.

W3C : RDF 1.1 Concepts and Abstract Syntax, 2014a. URL http://www.w3.org/TR/rdf11-concepts/.

W3C : RDF Schema 1.1. 2014b. URL http://www.w3.org/TR/rdf-schema/.

W3C Recommendation : OWL Web Ontology Language Overview. 2004. URL http://www.w3.org/TR/2004/REC-owl-features-20040210/.

W3C Recommendation : SPARQL Query Language for RDF. 2008. URL http://www.w3.org/TR/rdf-sparql-query/.

W3C Recommendation : OWL 2 Web Ontology Language Document Overview (Second Edition), 2012. URL http://www.w3.org/TR/owl2-overview/.

W3C Recommendation : SPARQL 1.1 Overview, 2013. URL http://www.w3.org/TR/sparql11-overview/.

W3C Recommendation : RDF 1.1 XML Syntax, 2014. URL http://www.w3.org/TR/rdf-syntax-grammar/.

S A Waksman : What is an antibiotic or an antibiotic substance ? *Mycologia*, 39(5):565–9, 1947. ISSN 0027-5514. URL http://www.ncbi.nlm.nih.gov/pubmed/20264541.

Fergus Walsh : Antibiotic resistance : Cameron warns of medical 'dark ages', 2014. URL http://www.bbc.com/news/health-28098838.

Linda M Weigel, Don B Clewell, Steven R Gill, Nancye C Clark, Linda K McDougal, Susan E Flannagan, James F Kolonay, Jyoti Shetty, George E Killgore et Fred C Tenover : Genetic analysis of a high-level vancomycin-resistant isolate of Staphylococcus aureus. *Science (New York, N.Y.)*, 302(5650):1569–71, novembre 2003. ISSN 1095-9203. URL http://www.ncbi.nlm.nih.gov/pubmed/14645850.

Kris Wetterstrand : DNA Sequencing Costs : Data from the NHGRI Genome Sequencing Program (GSP). 2014. URL http://www.genome.gov/sequencingcosts.

Michael Widmann, Jürgen Pleiss et Peter Oelschlaeger : Systematic Analysis of Metallo-β-Lactamases Using an Automated Database. *Antimicrobial agents and chemotherapy*, 56 (7):3481–91, juillet 2012. ISSN 1098-6596. URL http://www.ncbi.nlm.nih.gov/pubmed/22547615.

Daniel N Wilson : The A-Z of bacterial translation inhibitors. *Critical reviews in biochemistry and molecular biology*, 44(6):393–433, 2009. ISSN 1549-7798. URL http://www.ncbi.nlm.nih.gov/pubmed/19929179.

World Health Organization (WHO) : WHO model list of essential medicines : 18th list, April 2013. 2013. URL http://apps.who.int/iris/bitstream/10665/93142/1/EML_18_eng.pdf.

Gerard D Wright : Bacterial resistance to antibiotics : enzymatic degradation and modification. *Advanced drug delivery reviews*, 57(10):1451–70, juillet 2005. ISSN 0169-409X. URL http://www.ncbi.nlm.nih.gov/pubmed/15950313.

Da-Qiang Wu, Jing Ye, Hong-Yu Ou, Xue Wei, Xianqing Huang, Ya-Wen He et Yuquan Xu : Genomic analysis and temperature-dependent transcriptome profiles of the rhizosphere originating strain Pseudomonas aeruginosa M18. *BMC genomics*, 12:438, janvier 2011. ISSN 1471-2164. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3189399&tool=pmcentrez&rendertype=abstract.

Jianhui Xiong, David C Alexander, Jennifer H Ma, Maxime Déraspe, Donald E Low, Frances B Jamieson et Paul H Roy : Complete sequence of pOZ176, a 500-kilobase IncP-2 plasmid encoding IMP-9-mediated carbapenem resistance, from outbreak isolate Pseudomonas aeruginosa 96. *Antimicrobial agents and chemotherapy*, 57(8):3775–82, août 2013. ISSN 1098-6596. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3719692&tool=pmcentrez&rendertype=abstract.

Jian Yang, Lihong Chen, Lilian Sun, Jun Yu et Qi Jin : VFDB 2008 release : an enhanced web-based resource for comparative pathogenomics. *Nucleic acids research*, 36(Database issue):D539–42, janvier 2008. ISSN 1362-4962. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2238871&tool=pmcentrez&rendertype=abstract.

Mi-Kyung Yun, Yinan Wu, Zhenmei Li, Ying Zhao, M Brett Waddell, Antonio M Ferreira, Richard E Lee, Donald Bashford et Stephen W White : Catalysis and sulfa drug resistance in dihydropteroate synthase. *Science (New York, N.Y.)*, 335(6072):1110–4, mars 2012. ISSN 1095-9203. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531234&tool=pmcentrez&rendertype=abstract.

Alexandre Prehn Zavascki, Luciano Zubaran Goldani, Jian Li et Roger L Nation : Polymyxin B for the treatment of multidrug-resistant pathogens : a critical review. *The Journal of antimicrobial chemotherapy*, 60(6):1206–15, décembre 2007. ISSN 0305-7453. URL http://www.ncbi.nlm.nih.gov/pubmed/17878146.

Daniel R Zerbino et Ewan Birney : Velvet : algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5):821–829, 2008.

HongMin Zhang et Quan Hao : Crystal structure of NDM-1 reveals a common β-lactam hydrolysis mechanism. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 25(8):2574–82, août 2011. ISSN 1530-6860. URL http://www.ncbi.nlm.nih.gov/pubmed/21507902.