Université du Québec
Institut National de la Recherche Scientifique
Centre Armand-Frappier

DÉCOUVERTE ET ÉTUDE D'ARN NONCODANTS RÉGULANT DES MÉTHYLASES D'ARN

Par
Katia Smail

Mémoire pour l'obtention du grade de
Maître ès sciences (M.Sc.) en microbiologie appliquée

Jury d'évaluation

Examinateur externe	Dr. Marc Drolet Université de Montréal
Examinatrice interne et présidente du jury	Dre. Maritza Jaramillo INRS -Centre Armand-Frappier Santé Biotechnologie
Directeur de recherche	Dr. Jonathan Perreault INRS – Centre Armand-Frappier Santé Biotechnologie

REMERCIEMENTS

Les travaux de recherche présentés dans ce mémoire ont été réalisés au sein du laboratoire du Dr. Jonathan Perreault à l'INRS-Institut Armand-Frappier, et ont été financés par la fondation Armand-Frappier et le conseil de recherche en sciences naturelles et en génie du Canada.

Je tiens très sincèrement à remercier toutes les personnes qui ont contribué à la réalisation de ces travaux de recherche et au bon déroulement de cette maîtrise, particulièrement mon directeur de recherche Dr. Jonathan Perreault qui m'a accueilli dans son laboratoire et grâce à qui j'ai accompli plusieurs défis que je ne me serais pas cru capable de faire. Je le remercie d'avoir cru en mes capacités, pour son soutien, ses bons conseils et surtout son humanité. Je remercie également notre ancienne assistante de recherche, madame Julie Motard, pour son aide et ses bons conseils. Je remercie tous les membres de mon laboratoire avec qui j'ai vécu deux années très enrichissantes durant lesquelles j'ai beaucoup appris sur le plan scientifique mais également sur le plan humain. Finalement, je remercie mes parents pour leur soutien inconditionnel ainsi que mes deux frères à qui je souhaite beaucoup de succès.

RÉSUMÉ

Les ARN noncodants (ARNnc) sont des molécules d'ARN non traduites en protéines ayant un rôle dans la régulation des gènes. Chez les bactéries, ils constituent un élément crucial de la régulation génique car ils permettent une homéostasie, entre autres par des ajustements de leur physiologie en réponse à différentes conditions environnementales. Les régions UTR (untranslated region) sont localisées en 5′ ou en 3′ d'une région codante dans un ARNm. On retrouve dans celles-ci la majorité des ARNnc régulateurs connus tels les *riboswitchs*, les petits ARN, les ARN anti-sens ainsi que des éléments d'autorégulation. Dans ce projet, nous émettons l'hypothèse que des gènes qui codent pour des enzymes qui modifient les ARN pourraient s'autoréguler à travers leur UTR. Nous nous intéressons particulièrement à la méthylation des ARN car de potentiels exemples d'un tel mécanisme ont été rapportés dans la littérature (tel le motif mraW). L'objectif du projet consiste à trouver de bons candidats à l'aide d'outils bio-informatiques et d'étudier ensuite expérimentalement les candidats potentiels au laboratoire pour confirmer leur rôle et leur mécanisme d'action. Les expériences seront réalisées principalement de deux façons : avec des gènes rapporteurs *in vivo* et, lorsque confirmé, par des essais de méthylation *in vitro*. Ainsi, en utilisant la base de données RiboGap, qui permet un accès très simplifié aux régions inter-géniques chez les procaryotes, et le pipeline GraphClust, qui permet de trouver des regroupements conservés de séquences, nous avons sélectionné une liste de motifs à tester au laboratoire. Nous avons pour commencer entrepris de confirmer la structure de certains de nos motifs via des expériences d'*in-line probing*. Par la suite, le motif-28 retrouvé en avant de la méthyltransférase *mnmC* a été testé *in vivo*. Les résultats obtenus combinés à l'outil BPROM ont permis de mettre en évidence une possible régulation à double sens dans la région 5′UTR du gène *mnmC,* ce qui nous a amené à prédire le promoteur potentiel du gène *mnmC* qui était jusqu'alors inconnu. Avec BPROM nous avons raffiné la liste des candidats potentiels que nous avons obtenus.

TABLES DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES TABLEAUX SUPPLEMENTAIRES ACCOMPAGNANT LA VERSION ELECTRONIQUE (*i.e. Excel spreadsheets*)

1. Table S2 : Total sequences found upstream of *mnmC* gene
2. Fasta all : fasta file of all sequences found upstream of *mnmC* gene
3. Table S3 : fasta file of sequences found upstream of *mnmC* gene with Redendancy <98
4. FabR binding sites : table of FabR binding sites description

LISTE DES ABREVIATIONS

ADN : acide désoxyribonucléique

ARN : acide ribonucléique

ARNnc : ARN noncodant

ARNm : ARN messager

ARNt : ARN de transfert

ARNr : ARN ribosomal

pARN : petit ARN

UTR : *UnTranstated Region* (régions non transcrites)

SD : séquence Shine-Dalgarno

TPP : thiamine pyrophosphate

FMN : flavine mononucléotide

RBS : *ribosomal binding site* (Site de liaison au ribosome)

SAM : S-adénosyl-méthionine

NAD : nicotinamide adénine di-nucléotides

FAD : flavine adénine di-nucléotides

T : thymine

A : adénine

C : cytosine

G : guanine

U : uracile

ARN-MTases : ARN méthyl-transférases

RIG : région inter-génique

INTRODUCTION

Les bactéries ayant une taille très limitée, elles ne disposent pas de moyens de changer d'environnement quand celui-ci est défavorable. Cependant, elles sont pourvues d'une capacité exceptionnelle d'adaptation aux changements pouvant se produire dans leur environnement. En effet, le mode de vie d'une bactérie sera déterminé par son environnement et les espèces avec lesquelles elle interagit. Ainsi, les bactéries sont dotées de systèmes leur permettant d'apprécier leur environnement extérieur et d'intégrer l'information via différents signaux et voies métaboliques. Suite à cela, des décisions d'adaptation et de survie sont prises grâce à des « algorithmes » de régulation de gènes essentiels à la survie (Cases & de Lorenzo, 2005), en d'autres mots, en modulant l'expression des gènes qui permettront l'adaptation nécessaire ou le maintient de l'homéostasie. Les génomes bactériens ont la caractéristique d'être optimisés, c'est-à-dire qu'ils ont très peu de gènes non essentiels à leur survie dans l'ensemble de leurs environnements typiques. Par exemple, les bactéries intracellulaires et endosymbiotiques possèdent de petits génomes car débarrassés de certains gènes inutiles dans l'environnement de la cellule hôte. Ainsi, plus une bactérie est présente dans diverses niches écologiques, plus la taille de son génome est importante. En effet, les possibilités de régulation augmentent avec la grosseur du génome, ce qui augmente la capacité d'adaptation à plusieurs environnements (Cases *et al*., 2003). L'évolution a permis une régulation stricte de l'expression des gènes nécessaires à la survie. Les bactéries sont pourvues de plusieurs systèmes de régulation des gènes, pouvant affecter l'initiation de la transcription au niveau des acides désoxyribonucléiques (ADN), la traduction des acides ribonucléiques (ARN) et jusqu'à la régulation post-traductionnelle au niveau protéique. La majeure partie de la régulation se fait au niveau de l'initiation de la transcription via des facteurs de régulation (facteurs sigma), des activateurs et des répresseurs (Cases & de Lorenzo, 2005). Néanmoins, l'information génétique étant transmise via les ARN, ces molécules demeurent très importantes dans le processus de régulation.

1. Les acides ribonucléiques (ARN)

L'ARN est une molécule polymérique issue de la transcription par l'ARN polymérase de l'ADN dont il est une copie. Contrairement à l'ADN, l'ARN est composé d'un ribose au lieu d'un désoxyribose et d'une base azotée uracile au lieu de la thymine. Les bases azotées guanine, cytosine et adénine sont communes aux deux molécules. Outre les différences dans la composition

chimique, l'ARN est simple brin ce qui permet la formation de structures secondaires pouvant être très complexes et éloignées de la structure en double hélice de l'ADN.

On distingue deux types de molécules d'ARN, à savoir les ARN codants et les ARN noncodants (ARNnc). La différence entre les deux réside dans le processus de traduction en protéines. En effet, les ARN peuvent servir de messagers entre l'ADN et la synthèse des protéines (ARNm), ils sont alors dits codants car ils sont traduits par les ribosomes en protéines, ou alors ils peuvent être directement impliqués dans différentes fonctions telle la régulation, ils sont alors dits noncodants car ils ne sont pas traduits en protéines (Kozak, 1983).

1.1. Les ARN codants

Dans les ARN traduits en protéines, dits ARN messagers (ARNm), chaque triplet de nucléotides (codon) est décodé par un ARN de transfert (ARNt) comme codant pour un acide aminé spécifique selon le code génétique (Crick *et al*., 1961). Les acides aminés sont reliés dans un enchainement grâce à des liens peptidiques synthétisés par les ribosomes. Les ribosomes orchestrent le processus de traduction en permettant aux ARNt de reconnaître leurs codons respectifs et en synthétisant les liens peptidiques des protéines.

Chez les bactéries, la majorité du génome est codant. En effet, tel que mentionné plus haut, les génomes bactériens sont optimisés et ne contiennent que très peu ou pas de séquences non utilisées par les bactéries, contrairement aux eucaryotes pour lesquels la majorité du génome ne code pas de protéines (Cases *et al*., 2003). Les procaryotes se caractérisent aussi par la présence d'ARN polycistroniques, c'est-à-dire codants pour plusieurs protéines, qui généralement ont des fonctions reliées et sont souvent soumis aux mêmes éléments régulateurs. Les ARN codants pour une seule protéine sont dits monocistroniques (Kozak, 1983).

1.2. Les ARN non codants

Les ARN noncodants (ARNnc) ne sont pas traduits en protéines, ils agissent dans la cellule directement sous leur forme d'ARN et leur fonction est souvent étroitement liée à la structure. En effet, le repliement des ARNnc (structure secondaire et tertiaire) plus que leur séquence est déterminant dans le rôle qu'ils peuvent jouer dans la cellule. Les ARNnc sont impliqués dans plusieurs processus biologiques telle la traduction avec les ARNt et les ARN ribosomaux (ARNr) et la régulation avec les ARNnc dits régulateurs (Repoila & Darfeuille, 2009).

1.2.1. Les ARNnc régulateurs

Les ARNnc régulateurs forment un groupe hétérogène de molécules qui agissent selon plusieurs mécanismes sur un large spectre de processus biologiques. Ils sont localisés pour la plupart dans les régions intergéniques indépendamment de leur action en *cis* (c'est-à-dire, se trouvant dans la séquence de l'ARNm lui-même) ou en *trans* (c'est-à-dire, une molécule d'ARNnc pouvant agir sur d'autres ARNm). Les ARN comprennent entre autre les petits ARN (pARN), les région 5'UTR des ARNm dont les *riboswitchs*, les ribozymes, les ARN anti-sens et bien d'autres (Waters & Storz, 2009). Dans le cas des pARN, une seule molécule d'ARNnc peut agir sur un ou plusieurs ARNm (Breaker, 2011a). Ils sont cruciaux dans la régulation qui permet à la cellule d'ajuster sa physiologie en fonction des changements de l'environnement. Chez les bactéries, la majorité des ARNnc régulateurs sont des unités de transcription indépendantes et ont leur propre promoteur. Ils peuvent agirent seuls ou être couplés à des protéines, telles les chaperonnes *Hfq* des petits ARN qui facilitent la formation d'un duplex ARNnc-ARNm selon une complémentarité des paires de bases (Repoila & Darfeuille, 2009).

La présence d'ARNnc régulateurs chez les bactéries était connue bien avant les microARN chez les eucaryotes. En 1981 Stougard *et al*. ont démontré que « RNA I » bloque la réplication du plasmide ColE1 et en 1983 Simons et Kleckner ont démonté qu'un petit ARN transcrit à partir du transposon Tn10 inhibe la transposition (Waters & Storz, 2009). Les premiers ARNnc ont été découverts grâce à des techniques d'analyse de gel et des analyses computationnelles. Des techniques de biologie moléculaire ont par la suite été utilisées, ce qui a permis d'augmenter le nombre d'ARNnc régulateurs identifiés (Waters & Storz, 2009). Certaines approches se basent sur la recherche des séquences conservées chez des bactéries phylogénitiquement proches ainsi que l'analyse de séquences proximales pour y chercher des promoteurs et des terminateurs de transcription. Les pARN prédits sont confirmés expérimentalement par des méthodes de *Northern* blot avec l'utilisation de sondes, par des approches de clonage utilisant des librairies d'ADN complémentaire d'ARN d'une taille variant entre 50 et 500 nucléotides ainsi que par immunoprécipitation en utilisant la protéine chaperonne *Hfq* suivie d'une détection directe de l'ARN lié sur puce ADN (Kawano *et al*., 2005).

Les ARNnc régulateurs agissent principalement selon deux mécanismes, soit par la liaison par complémentarité des bases à l'ARN messager (ARNm) soit par liaison directe aux protéines. Dans le premier mécanisme, l'ARNnc peut affecter la transcription, la traduction ou la stabilité de

l'ARNm via la formation de structures stables altérant sa fonction, tandis que dans le second c'est la fonction de la protéine qui est modulée. L'action de l'ARNnc peut se faire au niveau de la région 5′ UTR (*UnTranstated Region*) de l'ARNm, de la séquence codante ou de la région 3′ UTR. Pour cette dernière, la régulation se fait au niveau de la stabilité de l'ARNm, en effet, la liaison d'un ARNnc à une région 3′ UTR d'un ARNm peut soit permettre la stabilité et donc augmenter la traduction soit inversement altérer négativement la stabilité et par conséquent causer la dégradation de l'ARNm. La liaison à la région codante permet l'inhibition de la traduction, à l'exemple du mécanisme d'action des ARN anti-sens. Finalement, la régulation au niveau de la région 5′ UTR se fait soit par l'implication du site de liaison du ribosome (séquence Shine-Dalgarno : SD) ce qui affecte la traduction, ou bien par la formation de structure terminatrice prématurées (Repoila & Darfeuille, 2009).

2. Les régions 5′ UTR (5′ *UnTranslated Regions*)

Les 5′ UTR font partie des régions inter-géniques, régions reliant deux séquences codantes. Les 5′ UTR sont localisées en avant d'un promoteur et co-transcrites dans l'ARNm avec la région codante. Ainsi, ce sont les régions localisées en 5′ de l'ARN messager qui ne font pas partie de la séquence codante mais comprennent des éléments régulateurs souvent déterminant dans le destin de l'ARNm dans la cellule (Repoila & Darfeuille, 2009).

Les 5′ UTR agissent en *cis* dans la régulation de l'expression des gènes. On retrouve au niveau des 5′ UTR plusieurs éléments régulateurs tels les *riboswitchs*, les leaders de protéines ribosomales, les thermorégulateurs et d'autres (Waters & Storz, 2009).

3. Les riboswitchs

Les *riboswitchs* sont des éléments régulateurs localisés presque exclusivement en 5′ UTR de l'ARNm des gènes qu'ils contrôlent (Breaker, 2011a). Ils contiennent un domaine conservé dit aptamère, ayant la capacité de lier un ligand de façon spécifique, suivi d'un domaine variable dit plateforme d'expression (Figure 1). Chez les bactéries les plateformes d'expression renferment des terminateurs de transcription intrinsèques ou des tiges pouvant séquestrer le site de fixation du ribosome. L'ensemble aptamère-plateforme d'expression permet la régulation de l'expression du gène localisé en aval (Serganov & Nudler, 2013). Le terme anglais *switch,* qui peut se traduire par interrupteur, réfère au fait que les *riboswitchs* agissent comme un « interrupteur » génétique

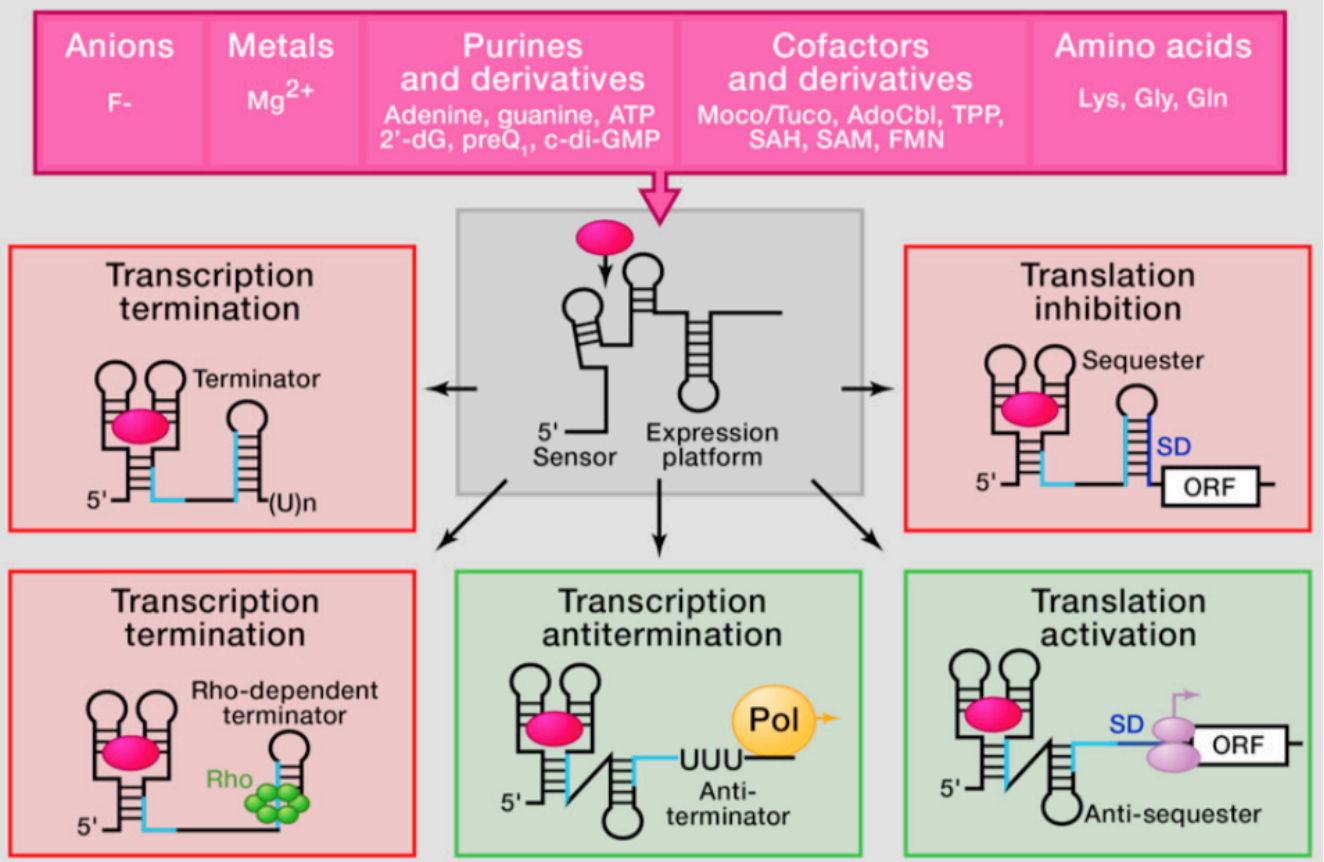

Figure 1. Différents ligands des riboswitchs et leurs mécanismes d'actions

La figure montre les différents mécanismes d'action qui se subdivisent principalement en deux catégories à savoir la régulation de la transcription de l'ADN en ARN et la régulation de la traduction des ARN en protéines. Cette dernière peut se faire en modulant l'accès du site de fixation du ribosome (SD pour séquence Shine-Dalgarno) en présence ou absence du ligand. Pour la transcription, la régulation se fait par la formation de structures anti-terminatrice ou terminatrice, Rho-dépendant ou pas, de la transcription en présence ou absence du ligand. Encadrés rouges : conformations « OFF », encadrés verts : conformations « ON », en bleu : segments d'une tige alternative correspondant à l'autre mode d'action « ON » ou « OFF ». En gris : représentation de la structure d'un *riboswitch* composé de deux parties : l'aptamère liant les ligands et la plateforme d'expression responsable du changement de l'expression. Encadré en rose : les différents types de ligands : les purines et leur dérivés, les cofacteurs d'enzyme et leur dérivés, les acides aminés ainsi que les éléments inorganiques comme les ions et les métaux. Tiré de (Serganov & Nudler, 2013)

qui fait passer l'état d'expression d'un gène de « ON » fonctionnel à « OFF » non fonctionnel et inversement (Breaker, 2011a; Serganov & Nudler, 2013). En effet, lorsque le métabolite pour lequel l'aptamère du *riboswitch* est spécifique est en excès dans la cellule, celui-ci se lie à l'aptamère causant un changement de conformation dans la plateforme d'expression modifiant ainsi l'expression du gène en aval. La caractéristique d'«interrupteur» n'est pas unique aux *riboswitchs*. En effet, d'autres éléments régulateurs ont la capacité de changé l'état d'un ARNm en se liant à des protéines ou des ARNt tels les atténuateurs, les T boxes, et riborégulateurs. Toutefois, les *riboswitchs* ont la caractéristique unique de lier leur ligand sans l'intervention d'aucune autre molécule (Serganov & Nudler, 2013).

3.1. Découverte

Nou et Kadner en 1998 et Miranda-Rios *et al*. en 2001 ont démontré une inhibition de la synthèse des vitamines B1, B2 et B12 par la thiamine, la riboflavin et la cobalamine, le mécanisme restait toutefois incompris, l'hypothèse étant qu'un intermédiaire protéique aurait été le responsable de l'inhibition. Ainsi, en changeant l'hypothèse d'une action indirecte à une possible action directe des molécules sur l'ARNm, l'hypothèse des ARNm senseurs de métabolites a émergé. La démonstration in vivo d'un changement de conformation d'un ARNm en présence et absence d'un métabolite a été observée chez *Salmonella typhimurium* en 2001 par Ravnum et Andersson (Ravnum & Andersson, 2001), et c'est en 2002 que des équipes ont démontré le lien direct entre les dérivés de vitamines thiamine pyrophosphate (TPP), flavine monocléotide (FMN) et AdoCbl et l'inhibition de la synthèse des vitamines B1, B2 et B12, ceci par (Mironov *et al*., 2002), (Winkler *et al*., 2002) et (Nahvi *et al*., 2002) respectivement.

3.2. Mécanismes d'action

Les *riboswitchs* répondent à une variété de ligands incluant les purines et leurs dérivés, les coenzymes de protéines, les acides aminés et certains éléments inorganiques dont les métaux. Les mécanismes de régulation les plus communs à l'action des *riboswitchs* sont l'interruption de la transcription et l'inhibition de la traduction. Chez plusieurs bactéries, la terminaison de la transcription est plus répandue, les bactéries éviteraient ainsi de synthétiser un ARN pleine longueur qui ne sera pas utilisé (Barrick & Breaker, 2007). Dans la plupart des cas, les *riboswitchs* agissent sur les gènes de transport et de biosynthèse des métabolites auxquels il répondent et pour

lesquels les mécanismes sont basés sur la présence de structures mutuellement exclusives en fonction de la présence ou absence du ligand (Serganov & Nudler, 2013).

Dans le cas de la régulation de la transcription, les *riboswitchs* forment des structures en épingle à cheveux anti-terminatrices ou terminatrices de la transcription Rho-indépendantes (Figure 1). Ces structures sont intrinsèques aux plateformes d'expression, sont mutuellement exclusives et dépendent de la liaison du ligand au niveau de l'aptamère. Des mécanisme de terminaison de la transcription Rho-dépendent ont aussi été décrits chez certaines espèces pour lesquelles les structures terminatrices Rho-indépendantes ou les structures séquestrant le site de fixation du ribosome sont absentes (Serganov & Nudler, 2013). Ce type de terminaison contrôlerait le niveau de l'ARNm dans la cellule, en modulant l'accès aux séquences impliquées dans la terminaison de la transcription Rho-dépendante et le recrutement de l'ARNase E responsable de la dégradation de l'ARN (Bastet *et al*., 2018).

Dans la régulation de l'accès aux sites de fixation des ribosomes (RBS pour l'expression en anglais, *ribosomal binding site*), trois mécanismes ont été décrits grâce à l'utilisation de méthodes d'analyse de comparaison de séquences (Breaker, 2018). Dans le premier mécanisme, le RBS peut faire partie de la structure même de l'aptamère. La liaison d'un ligand empêcherait ainsi l'accès au RBS formant un *riboswitch* en « OFF », comme l'exemple du *riboswitch* AdoCbl (Breaker, 2018). Dans le second mécanisme, le RBS est situé à une certaine distance de la structure de l'aptamère, son action impliquerait donc l'intervention d'une structure secondaire capable de se replier suite à la liaison du ligand. Celle-ci est dite plateforme d'expression et elle comprend deux états mutuellement exclusifs. Ainsi, ces *riboswitch* peuvent agir en « OFF » ou bien en « ON », à l'exemple du *riboswitch* TPP (Breaker, 2018). Finalement, le RBS pourrait être localisé avec une structure terminatrice de transcription intrinsèque à la plateforme d'expression ce qui en fait un *riboswitch* à double régulation. Certaines classes du riboswitch TPP fonctionneraient selon ce mécanisme. Des niveaux plus complexes de la régulation de la traduction pourraient exister en combinaison avec d'autres systèmes de régulation (Breaker, 2018).

4. L'autorégulation à travers les régions 5′ UTR

Outre les riboswitchs, les régions 5′ UTR renferment des systèmes d'autorégulation de la traduction. Chez les bactéries, les protéines ribosomales sont un exemple de ce mécanisme par lequel celles-ci lient leur propre ARNm au niveau des régions 5′ UTR pour réguler leur niveau de

traduction. En effet, les régions 5′ UTR se replient dans des structures secondaires spécifiques sur lesquelles se lient une ou plusieurs des protéines ribosomales traduites, formant un duplex empêchant la liaison des ribosomes et ainsi le processus de traduction (Meyer, 2017).

Le mécanisme de régulation le plus commun de ces protéines est la compétition pour le site de fixation du ribosome (RBS) illustré par l'opéron des protéines L1/L11 codées par les gènes *rplK* et *rplA*. Ainsi, la protéine L1 reconnait sur son ARNm une structure en forme de tige boucle similaire à sa cible sur l'ARNr 23S, ce qui cause un blocage de l'accès au RBS, l'arrêt de la traduction et la dégradation de l'ARNm (Figure 2). Un autre exemple de mécanisme d'action est le blocage du ribosome sur l'ARNm. La protéine ribosomale S15 codée par le gène *rpsO* s'autorégule en liant son ARNm dans une conformation tertiaire en pseudo-nœud empêchant le ribosome d'avancer (Figure 2). Des modèles indiquent que ce genre de mécanisme est généralement utilisé lorsque, dans un système de régulation, les répresseurs sont en faible concentration ou présentent une faible affinité (Meyer, 2017).

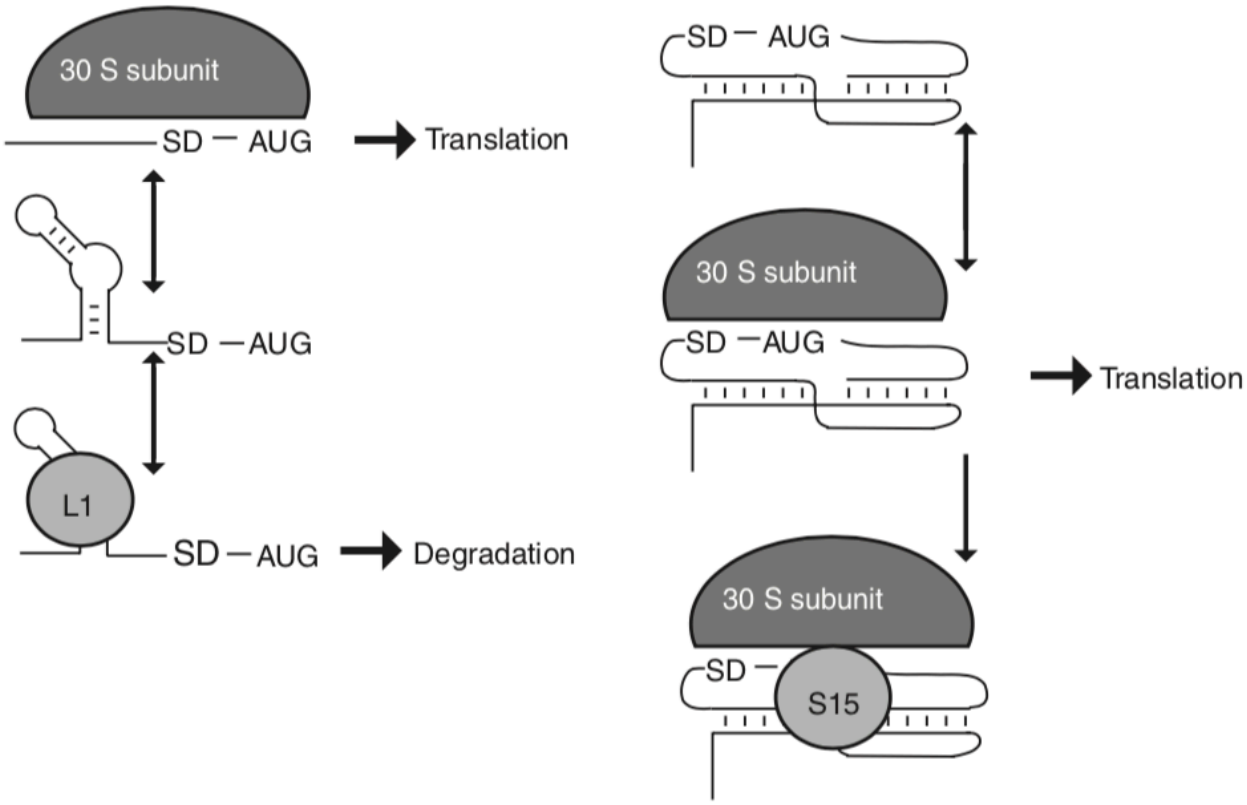

Figure 2. Exemples d'autorégulation des protéines ribosomales

À gauche : mécanisme d'autorégulation par compétition pour le site de fixation du ribosome illustré par la protéine L1, qui lie sur son ARNm une structure en tige boucle obstruant l'accès au ribosome. À droite : mécanisme d'autorégulation par séquestration du ribosome au niveau de son site de fixation à cause de la liaison de la protéine S15 qui forme une conformation en pseudo-nœud. Tiré de (Meyer, 2017)

5. Les modifications post-transcriptionnelles des d'ARN

Les ARN dénombrent de multiples modifications chimiques à différents sites dont la majorité sont introduites post-transcription. Celles-ci peuvent avoir un impact sur le repliement structurel des ARN et leur reconnaissance par les protéines. Ainsi, en vue de l'impact de ces modifications et de leur importance fonctionnelle dans les différents processus biologiques les impliquant, leur expression adéquate est très importante. En effet, les enzymes modifiant l'ARN font partie des protéines reliées au métabolisme des ARN les plus conservées, au même titre que les protéines impliquées dans la transcription et la traduction (Wang & He, 2014). Elles représentent un moyen de régulation post-transcriptionnel dynamique et réversible (Nachtergaele & He, 2017). Elles ont été mises en évidence par des techniques de mutagenèse impactant le métabolisme des ARNt, la résistance aux antibiotiques et la régulation de la traduction pour certains opérons (Marbaniang & Vogel, 2016).

Plus de cent soixante-dix modifications d'ARN ont été identifiées (Boccaletto *et al*., 2018). La plupart de ces modifications surviennent au niveau des ARN de transfert et des ARN ribosomaux. En effet, on dénombre chez ces derniers une soixantaine de modifications. Certaines peuvent être conservées dans plusieurs ARN et d'autres être plus spécifiques, par exemple, les modifications au niveau de la boucle anticodon des ARNt peuvent avoir un rôle déterminant dans le processus de traduction et la reconnaissance des ARNt acylés avec les acides aminés correspondant. Ainsi, on retrouve au niveau des ARN des modifications constitutives et covalentes et d'autres plus dynamiques, réversibles et dépendantes du stress environnemental et de la phase de croissance (Marbaniang & Vogel, 2016)

Les modifications des ARN peuvent être subdivisées en trois grandes catégories, à savoir, les modifications impliquées dans le renforcement de la structure des ARN comme pour les ARN ribosomaux, les modifications impliquées dans la reconnaissance moléculaire comme pour les ARN de transfert et les modifications à des fins de régulation comme celles retrouvées au niveau des ARNm (Wang & He, 2014). Plus spécifiquement, les modifications au niveau des ARNt modulent la reconnaissance des codons en augmentant ou diminuant l'affinité codon/anticodon, elles permettent de déterminer le choix du codon, la conservation du cadre de lecture et seraient même impliquées dans la virulence. Pour les ARNr, les modifications permettent un ajustement de la rigidité et le maintien du repliement de la structure, seraient impliquées dans la liaison aux ARNt et aux ARNm et influenceraient l'action des antibiotiques (Marbaniang & Vogel, 2016).

Le rôle de plusieurs autres modifications, particulièrement au niveau d'autres classes d'ARNnc, reste toutefois à mettre en évidence.

5.1. La méthylation des ARN

La méthylation des ARN est une modification très répandue dans tous les domaines de la vie. Plus de la moitié des modifications d'ARN répertoriées implique l'ajout d'un groupement méthyl. Chez les bactéries, cette modification touche différents types d'ARN tels les ARNt, les ARNr, les ARNtm, les ARNm, etc (Ovcharenko & Rentmeister, 2018). L'impact de la méthylation dépend du type de l'ARN et du nucléotide modifié. En effet, de façon générale la méthylation des ARN est multifonctionnelle et n'est pas associée spécifiquement à une seule fonction (Motorin & Helm, 2011). Les bactéries ont en commun avec les archées et les eucaryotes plusieurs sites de méthylation (Figure 3). Cette distribution et la conservation des modifications indiquent l'importance du rôle que peut jouer la méthylation des ARN dans la biologie de la cellule. Toutefois, au niveau des ARNr par exemple, aucune modification pris à part n'est nécessaire à la survie cellulaire, c'est plutôt l'ensemble des modifications qui peut affecter la survie si altéré (Sergeeva *et al*., 2015).

Les enzymes qui catalysent la réaction de méthylation des ARN sont les ARN méthyl-transférases (ARN-MTases) (Figure 3). Ces dernières peuvent être subdivisées (Motorin & Helm, 2011), en quatre superfamilles. La première superfamille est la famille des RFM, pour *Rossmann-fold* MTases, regroupant la majorité des RNA-MTases, en incluant certaines DNA-MTases. La superfamille SOUT représente le second groupe le plus large. Le nom associé à cette dernière est lié aux premières MTases classées dans le groupe : spoU et TermD MTases. La troisième superfamille est celle dite la famille radical SAM, dû au fait que les enzymes de ce groupe génèrent un radical à partir du S-adénosyl-méthionine (SAM) qui sera utilisé pour la méthylation du substrat. Finalement, la quatrième superfamille est la famille fixatrice des cofacteurs flavine adénine di-nucléotides (FAD) et nicotinamide adénine di-nucléotides (NAD). À quelques exceptions près, toutes ces enzymes utilisent le S-adénosyl-méthionine (SAM) comme donneur de groupement méthyle (Motorin & Helm, 2011).

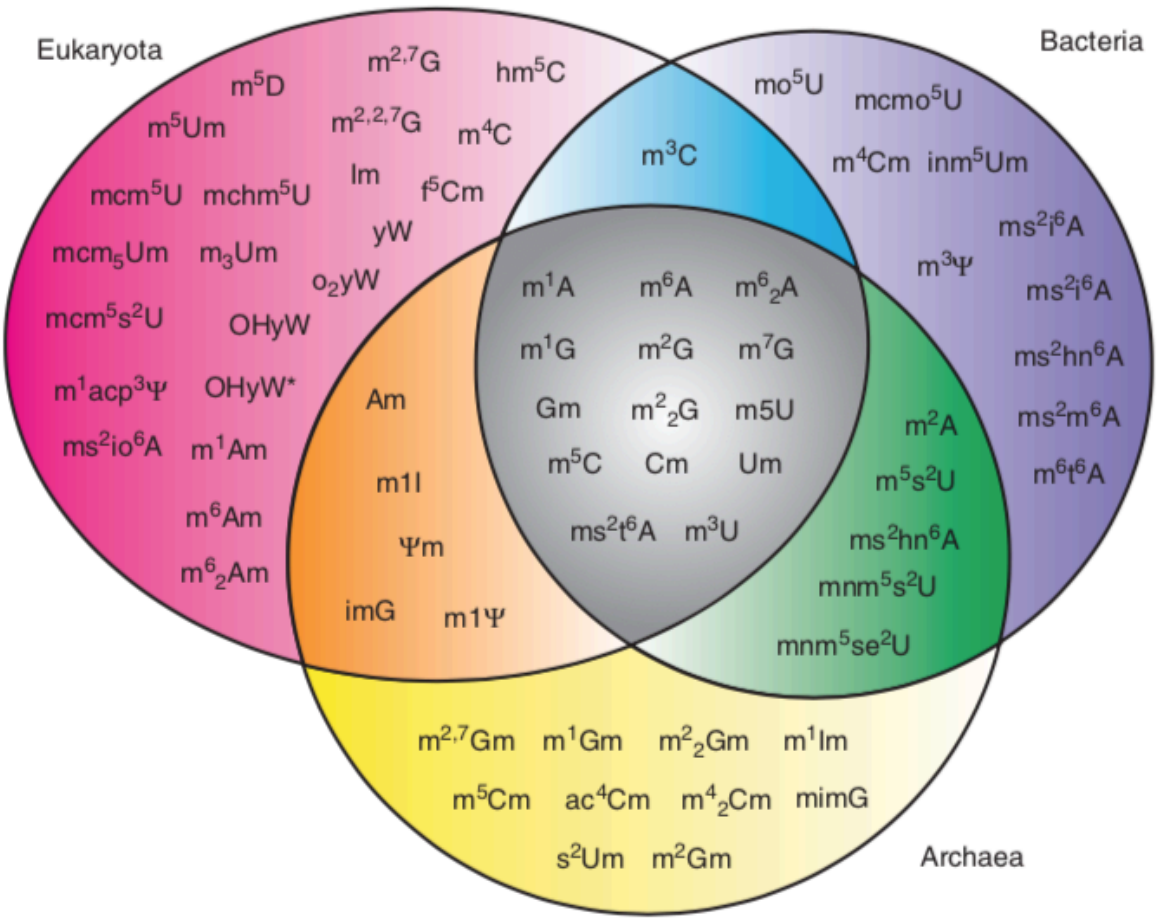

Figure 3. Distribution phylogénique de la méthylation des ARN

La méthylation des ARN présente une distribution large à travers les domaines de la vie, dont certaines (zone grise) sont communes aux trois domaines. On retrouve également des modifications plus spécifiques, toutefois, il se peut que les liens ne soient pas encore tout à fait établis et d'autres modifications communes ou chevauchant plusieurs domaines restent à découvrir. Tiré de (Motorin & Helm, 2011).

5.2. La méthylation des ARN ribosomaux (ARNr)

Le ribosome 70S de *Escherichia coli* est composé des deux sous-unités 30S et 50S, elles même constituées d'un assemblage d'ARNr et de protéines. On retrouve au niveau des ARNr une multitude de sites de méthylation dont la plupart sont situés au centre catalytique du ribosome (Motorin & Helm, 2011). La méthylation des nucléotides de la sous-unité 30S a deux fonctions principales, à savoir le contrôle de l'assemblage et le contrôle de l'initiation du processus de traduction des ARN messagers. Les modifications pré-assemblage telles $m^6{}_2A1518$, $m^6{}_2A1519$ et m^5C967 formeraient un point de contrôle de la qualité de l'ARNr. Tandis que les modifications post assemblage seraient impliquées dans la stabilité et le bon déroulement de l'initiation de la traduction. Exemple, les modifications m^2G966 et m^5C967 améliorent l'initiation de la traduction (Sergeeva *et al*., 2015).

Au niveau de la sous-unité 50S, la fonction principale des modifications seraient liée à la stabilité de la structure de l'ARN et des interactions avec les protéines ribosomales ainsi que la formation d'une sous-unité ribosomale active, exemples : m^1G745, m^6A1618, m^2G1835 et m^6A2030 (Sergeeva *et al*., 2015). Toutefois, on retrouve également au niveau de la sous-unité 50S des modifications impliquées dans l'interaction avec les ARN de transfert telles m^5U747 et m^5U1939, la reconnaissance des substrats de ribosomes telle la méthylation à la position A2503 et la reconnaissance au niveau du site A du ribosome avec la méthylation à la position U2252 (Sergeeva *et al*., 2015).

5.3. La méthylation des ARN de transfert (ARNt)

Les ARNt sont les ARNnc les plus abondants, ils présentent une structure tridimensionnelle en forme de « L » très conservée et sont caractérisés par une teneur et une diversité en modifications très élevée (Motorin & Helm, 2011). Ces modifications sont introduites post-transcription à des positions précises par des enzymes spécifiques (Figures 4 et 5). Elles jouent un rôle important dans le repliement, la stabilité, l'identité et la fonction de l'ARNt (Moukadiri *et al*., 2014). Bien que la méthylation des ARNt jouerait plusieurs rôles, ceux-ci peuvent être subdivisés grossièrement en deux groupes : les fonctions structurales, et les fonctions de la régulation de la traduction.

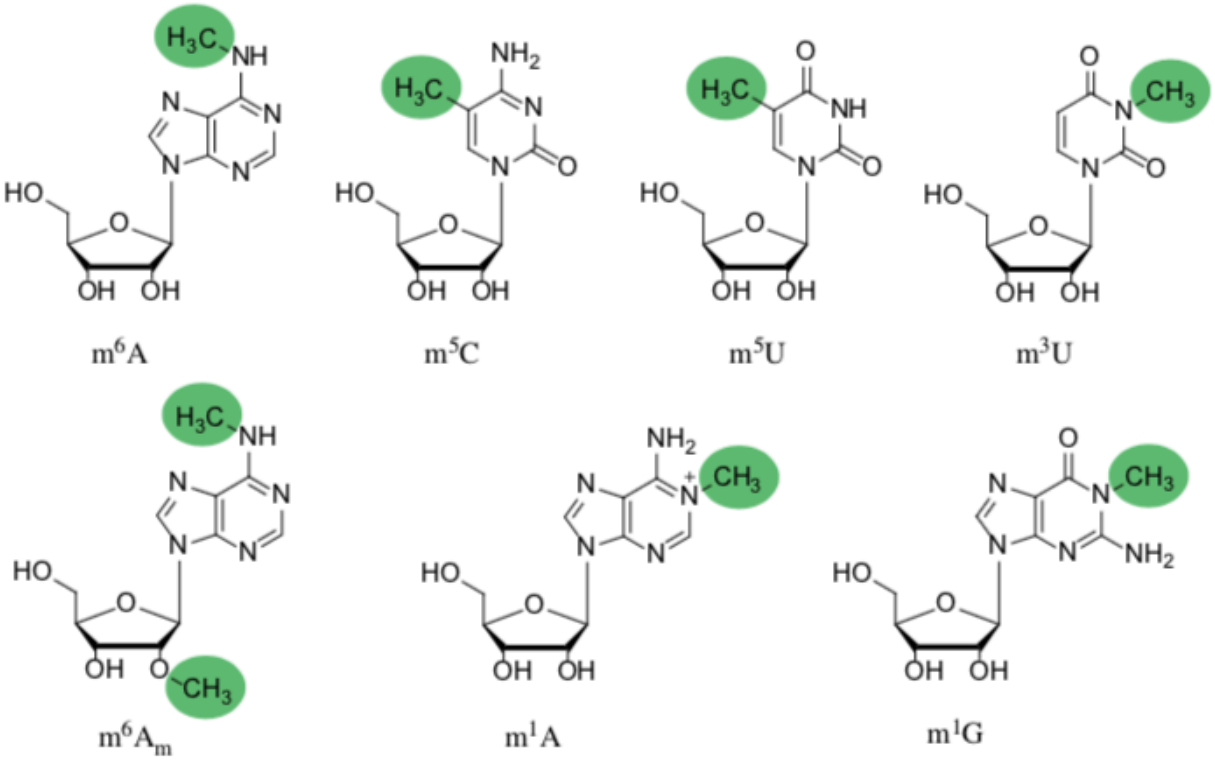

Figure 4. Exemples de sites de méthylations de RNA-MTases

Les MTases présentent plusieurs cibles sur l'ARN. Certains exemples sont illustrés ici, à partir du haut à gauche : N^6-methyladenosine (m^6A), 5-methylcytidine (m^5C), 5-methyluridine (m^5U), 3-methyluridine (m^3U), N^6,2^0-O-dimethyladenosine (m^6A_m), 1-methyladenosine (m^1A) and 1-methylguanosine (m^1G). Tiré de (Ovcharenko & Rentmeister, 2018)

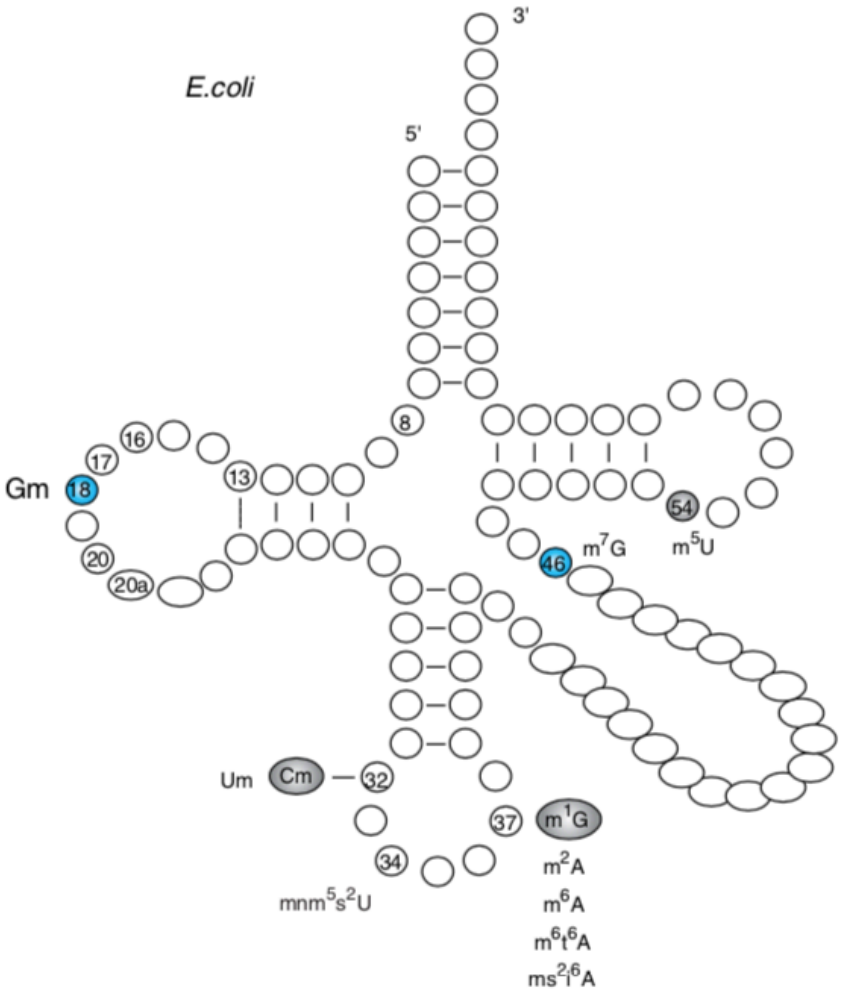

Figure 5. Exemples de sites de méthylation de l'ARN de transfert chez *Escherichia coli*

Les modifications, incluant la méthylation, de la boucle anticodon des ARN de transfert jouent un rôle important dans le bon fonctionnement de la traduction des ARNm en protéines. La position 34 comprend souvent des hyper-modifications complexes impliquées dans la liaison de l'ARNt à l'ARNm dans le ribosome. Par ailleurs, la position 37, dans la boucle, jouerait un rôle important dans le maintien du cadre de lecture. Tiré de (Motorin & Helm, 2011).

La stabilité biophysique des structures d'ARNt implique plusieurs mécanismes dont la liaison des ions magnésium, le blocage des liaisons Watson-Crick pour empêcher les liaisons alternatives et la modulation au niveau du ribose. Il est à considérer que ces modifications dépendent des conditions physiologiques. Par exemple, les modifications Gm18, T54 et Ψ55 chez *Escherichia coli* impliquées dans la stabilité de la structure seraient moins bénéfiques à haute température. Ainsi, plusieurs observations ont été rapportées quant à des mécanismes d'hypo-modifications indiquant une nature dynamique des modifications des ARN sous certaines conditions qui nécessitent une régulation génique complexe (Motorin & Helm, 2011).

Dans la fonction de régulation du processus de traduction des ARNm, les modifications aux positions 34 et 37, retrouvées au niveau de la boucle anticodon des ARNt, jouent un rôle essentiel dans la précision et l'efficacité de la traduction, particulièrement pour le maintien du cadre de lecture. Le nucléotide 37 est universellement une purine (guanine (G) ou adénine (A)). Dans le cas d'une guanine la modification est généralement m^1G37, des hyper-modifications peuvent également s'ajoutées par l'implication de différentes enzymes. Dans le cas où le nucléotide est une adénine, on retrouve différentes positions de méthylation telles m^2A et m^6A sur lesquelles s'ajoutent des hyper-modifications telles ms^2i^6A, ms^2t^6A, ms^2io^6A, et ms^2hn^6A. Le nucléotide à la position 34, dite position *Wobble,* peut être de nature variable ou dégénérative (adénine (A), uracile (U), guanine (G), cytosine (C)). Il est la cible de plusieurs enzymes qui introduisent des hyper-modifications dans des voies de synthèse complexes permettant la régulation de l'appariement des bases avec l'ARNm dans le ribosome (Motorin & Helm, 2011).

6. Motifs en amont de méthylases

À l'instar de ce qui a été discuté plus haut, les méthylases d'ARN (ARN-MTases) jouent un rôle important dans différents processus biologiques, elles sont par conséquent très conservées. Ainsi, le processus de méthylation présente des mécanismes de régulation à plusieurs niveaux. Les ARNnc régulateurs sont un moyen de régulation efficace et économique très utilisé par les bactéries. En effet, il existe plusieurs *riboswitchs* S-Adénosyl-Méthionine (SAM) régulateurs de gènes impliqués dans le métabolisme de SAM. Ce dernier étant le donneur de groupement méthyl le plus commun utilisé par les ARN-MTase. Les gènes codants ces dernières sont ainsi parfois régulés par leur propre substrat.

Des structures conservées trouvées par prédiction bio-informatique en avant de gènes codant des méthylases d'ARN ont été rapportées dans la littérature, notamment les motifs 23S (Weinberg *et al*., 2007) et mraW (Weinberg *et al*., 2010). Ainsi, l'ARN étant le substrat des ARN-MTases, ceci suggère un potentiel mécanisme d'autorégulation impliquant les ARNnc régulateurs à travers les régions 5′ UTR des ARNm, de la même façon que les protéines ribosomales présentées à la section 4.

6.1. Le motif ARN 23S-méthyl

Le motif ARN 23S-méthyl a été retrouvé, suite à des analyses de génomique comparative, en avant de gènes annotés ARNr 23S méthyltransferase chez les Lactobacillales de l'ordre des Firmicutes. Le motif se présente comme deux structures en tige-boucle dont la seconde est suivie d'une queue poly-U (uracile) suggérant la présence d'un terminateur de transcription rho-indépendant (Figure 6). Les deux tiges contiennent une importante co-variations, ce qui représente une forte évidence d'un ARN fonctionnel (voir le point 7.). Ainsi, le motif 23S-méthyl se présente comme un potentiel ARN régulateur en *cis* de la transcription (Weinberg *et al*., 2007).

6.2. Le motif ARN mraW

Le motif ARN mraW est une structure conservée retrouvée chez certaines espèces d'Actinobactérie, particulièrement les Mycobactéries, en avant du gène *mraW* codant une méthyltransférase. Ce dernier est trouvé dans un opéron contenant des gènes impliqués dans la synthèse du peptidoglycane. Le motif se présente comme une structure en tige-boucle avec une boucle terminale très conservée et une tige contenant plusieurs co-variations (Figure 6) suggérant une implication fonctionnelle (Weinberg *et al*., 2010).

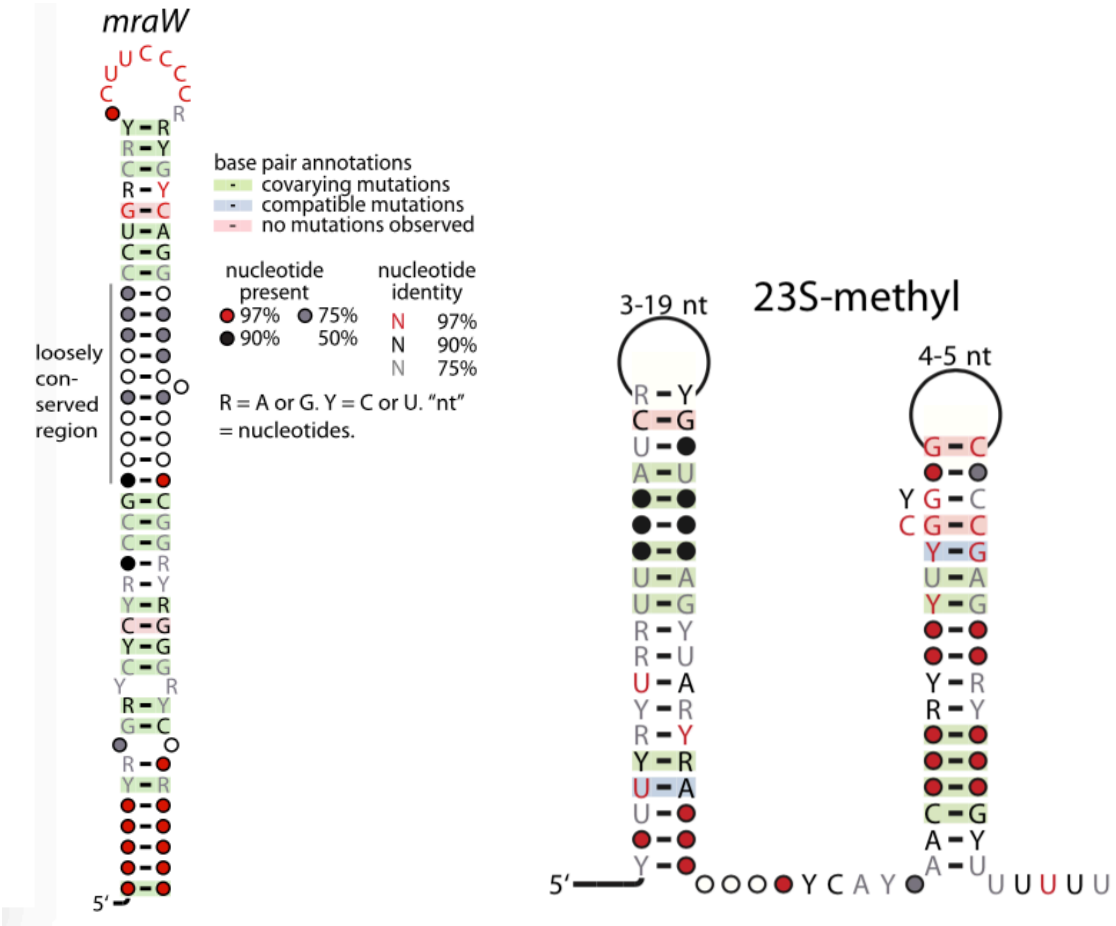

Figure 6. Motifs suspectés de réguler des méthylases d'ARN

Les motifs ARN mraW et 23S-méthyl ont été prédits par bio-informatique. L'ARN mraW est retrouvé en avant du gène *mraW* codant une méthyltransferase. Le gène *mraW* fait partie d'un opéron codant des protéines impliquées dans la synthèse du peptidoglycane. L'ARN mraW est proposé comme régulateurs de ces protéines. Tiré de (Weinberg *et al.*, 2010). L'ARN 23S-méthyl est souvent retrouvé en avant d'une méthylase de l'ARNr 23S suggérant une *cis*-régulation de la transcription rho-indépendante étant donnée la présence de la queue poly-U à la fin de la seconde tige. Tiré de (Weinberg *et al.*, 2007).

7. Recherche et découverte *de novo* de motifs d'ARNnc

Les génomes bactériens sont denses en informations et l'avènement du séquençage à haut débit a permis d'explorer un large ensemble de séquences, ce qui a donné l'accès à de nouvelles découvertes. En effet, l'utilisation d'outils bio-informatiques dans les dernières années a été très fructueux, cela a permis la découverte de centaines d'ARNnc. Toutefois, contrairement aux ARN codants, plusieurs ARNnc sont mal caractérisés et il est à croire que plusieurs restent encore à découvrir.

Tel que discuté plus haut, la majorité des ARNnc connus se retrouvent au niveau des régions inter-géniques (RIG). Ces dernières, contrairement aux régions codantes, ne subissent pas de pressions évolutives pour la conservation des séquences. Cependant, dans les RIG contenant des éléments régulateurs, des conservations de séquences et de structures peuvent être observées. Par ailleurs, les ARNnc ont évolués en conservant leurs structures plutôt que leurs séquences nucléotidiques. En conséquence, la recherche de nouveaux ARNnc devrait être basée sur la prédiction des structures, ce qui implique l'utilisation d'outils capables de capturer des signaux structurels; le plus grand défi étant de réussir à détecter le signal d'une structure fonctionnelle dans un large ensemble de données (Backofen *et al*., 2014a). Pour ce faire, des outils de comparaison génomique et d'analyse de co-variations ont été utilisés. Le point clé de ces outils est la présence de mutations compensatoires (mutation compatibles) ou simultanées (co-variations) qui n'altèrent pas l'appariement des bases, permettant donc la conservation de la structure. Celles-ci sont utilisées comme point de repère indiquant à une possible structure fonctionnelle. Ainsi, en tenant compte des caractéristiques des ARNnc et du large ensemble de séquences disponibles à traiter, le succès d'une recherche *de novo* d'ARNnc pourraient être résumé en quelques points. D'abord, avoir un bon ensemble de données à traiter. Ce dernier doit contenir assez de séquences avec une distance évolutive appropriée pour permettre l'extraction d'un signal structurel le moins faussé possible. En effet, des séquences trop similaires ne présenteraient que très peu de mutations compensatoires et de co-variations, tandis que des séquences très éloignées pourraient présenter plusieurs de celles-ci mais le signal serait difficile à interpréter. Par la suite, il est nécessaire d'effectuer de bons alignements, en tenant compte de la conservation des séquences et de la structure secondaire pour chacune, et d'établir les caractéristiques communes des séquence alignées. La définition de l'alignement local est une étape critique pour déterminer les frontières du motif, cela implique aussi d'écarter le reste des séquences dans l'ensemble de données initial. Finalement, une fois la

structure consensus et le model établi, la dernière étape serait d'effectuer une recherche à l'échelle des génomes pour retrouver davantage d'instances permettant de raffiner le modèle (Ruzzo & Gorodkin, 2014).

7.1. Approches utilisées

L'annotation des ARN est basée sur la similarité des structures et des fonction, déterminée par des méthodes de regroupement en fonction des similarités séquence-structure. L'algorithme Sankoff d'analyses computationnelles phylogénétiques a été la base des méthodes les plus utilisées exploitant de façon simultanée l'alignement et le repliement des séquences dans la découverte de structures d'ARN fonctionnels. De façon générale, ces approches peuvent être classées en deux catégories. La première catégorie utilise des simplifications des représentation des structures et la seconde utilise les informations de séquences comme des connaissances préalables permettant l'accélération des calculs (Heyne *et al*., 2012a).

Dans la catégorie de la simplification des représentations des structures, des outils tels RNAforester (Hochsmann *et al*., 2003) et MARNA (Siebert & Backofen, 2005) ont été utilisés. Ces derniers prennent en compte une seule prédiction de structure par séquence, ils dépendent donc fortement de cette unique prédiction. Cependant les prédictions computationnelles de structures sont sujettes aux erreurs. Dans la même catégorie, on retrouve aussi des approches qui simplifie les modèles pour classer les ARN tel l'outil EvoFold (Pedersen *et al*., 2006).

Dans la catégorie des approches qui utilisent les informations des séquences, celles-ci sont regroupées et alignées et des structures consensus sont ensuite prédites en utilisant des outils tel RNAfold (Lorenz *et al*., 2011). D'autre outils comme CMfinder (Yao *et al*., 2006a) peuvent être directement utilisés sur un ensemble de séquence non alignées pour prédire la structure. Ainsi, contrairement à la première catégorie qui tient compte des structures individuelles, le traitement d'un ensemble de séquences permet l'accélération des calculs (Backofen *et al*., 2014a; Heyne *et al*., 2012a).

7.2. Limites

Les méthodes utilisées jusqu'à présent ont démontré une efficacité considérable dans la découverte de plusieurs structures d'ARN fonctionnels. Cependant, certaines difficultés sont à considérer. Tel que mentionné, il faut avoir un ensemble de données initial avec un nombre suffisant de séquences ayant une distance évolutive appropriée. Toutefois, l'accès aux séquences appropriées en utilisant

les bases de données conventionnelles telles que GenBank et Ensembl peut être laborieux car il nécessite l'utilisation de scripts combiné aux positions dans les génomes (Naghdi *et al*., 2017). Autre problème majeur, notamment des méthodes basées sur l'alignement, est que l'ARN évolue plus rapidement au niveau de sa séquence plutôt que de sa structure ce qui rend la détection de l'homologie des séquences difficile pour des espèces éloignées. D'autre part, le traitement d'un large ensemble de donnée nécessite des ressources computationnelles considérables et cela peut être très limitant (Heyne *et al*., 2012a). Ainsi, un accès plus facile aux séquences et des méthodes de regroupement ne nécessitant pas d'alignement pourraient présenter une solution alternative pour améliorer l'allocation des ressources nécessaires.

7.3. Approche basée sur la fonction

Une des façons de chercher de nouvelles structures d'ARN fonctionnels est d'optimiser la sélection de l'ensemble de données initial, ceci en s'intéressant aux régions intergéniques en avant de gènes fonctionnellement reliés. En effet, les gènes ayant la même fonction sont plus susceptibles d'être régulés par les même éléments régulateurs. Cela est particulièrement intéressant pour la découverte *de novo* d'ARNnc régulateurs *en cis* (Naghdi *et al*., 2017). Ainsi, une recherche effectuée en se basant sur la fonction des gènes pourrait optimiser le signal pour être plus facilement détectable dans le processus d'analyse computationnelle.

7.3.1. RiboGap

RiboGap est une base de données procaryotes, qui permet un accès facile aux séquences intergéniques et aux différents éléments qui y sont annotés. Elle est particulièrement utile pour l'analyse des séquences intergéniques des procaryotes. En effet, elle fournit un moyen de recherche facile en sélectionnant simplement les champs d'intérêt et en y insérant les mots clés appropriés (Figure 7). Les éléments annotés sont regroupés par fonction, taxonomie ou traits phénotypiques. RiboGap facilite également l'extraction de séquences en avant de gènes d'intérêt, ce qui est particulièrement utile dans la recherche de nouveau éléments ARNnc par génomique comparative (Naghdi *et al*., 2017).

7.3.2. GraphClust

Heyne *et al*. ont présenté en 2012 une nouvelle approche qui permet de traiter des ensembles de données de plusieurs milliers de séquences. L'approche est le résultat de l'adaptation d'une

technique de chimio-informatique, développée par la même équipe (Grave & Costa, 2010), pour la détection des similitudes entre les structures secondaires d'ARN. L'approche consiste à utiliser dans un premier temps une technique de « hashage » indépendante de l'alignement dans laquelle les informations sur la structure et la séquence pour chaque ARN sont encodées dans un vecteur. Les positions de l'ensemble des vecteurs obtenus sont alors comparées pour former les regroupements de séquences (clusters), rendant la méthode plus rapide que celles basées sur l'alignement. Des méthodes d'alignement sont ensuite appliquées pour augmenter la qualité des clusters obtenus en éliminant les éléments non consistants (Heyne *et al*., 2012a).

L'approche a été intégrée dans un pipeline d'analyse nommé GraphClust, qui s'exécute en ligne de commande. Celui-ci présente plusieurs dépendances qui sont discuté au chapitre 2. GraphClust permet d'analyser des données issues d'études transcriptomiques ou de source computationnelle. L'exécution du programme se fait en plusieurs étape avec notamment une étape de prétraitement qui permet d'éliminer les séquences dupliquas et d'optimiser les signaux locaux en coupant les longues séquences. Les *clusters* sont produits en tenant compte de la structure et de la séquence pour chaque élément de l'ensemble de données. Suite à la production des clusters, ceux-ci sont raffinés notamment avec l'outil LocARNA (Will *et al*., 2012) qui utilise des méthodes phylogénétiques telle UPGMA pour classer les séquences. Les mieux classées parmi ces dernières sont sélectionnées et réalignées, LocARNA-P permet de calculer un score de fiabilité et identifie les régions de confiance et Infernal (Nawrocki & Eddy, 2013) de produir les models de covariance (Heyne *et al*., 2012a).

7.3.1. Découverte de nouveaux ARN avec RiboGap et GraphClust

L'utilisation de l'approche basée sur la fonction dans la recherche de nouvelles structures d'ARN fonctionnels peut être facilitée par l'utilisation de RiboGap et de GraphClust. En effet, RiboGap permet d'obtenir un ensemble de données initial pertinent en lien avec une fonction d'intérêt. Tandis que GraphClust permet d'effectuer des analyses rapides sur un large ensemble de données. Par ailleurs, GraphClust génère de nombreux résultats nécessitant un tri. En effet, certains critères de sélection doivent être pris en compte, à savoir : la qualité de l'alignement, la conservation des séquences, la présence de covariations, la structure secondaire et dans le cas d'une recherche *de novo*, l'absence d'annotation dans les bases de données telle Rfam (Kalvari *et al*., 2018) (Naghdi *et al*., 2017).

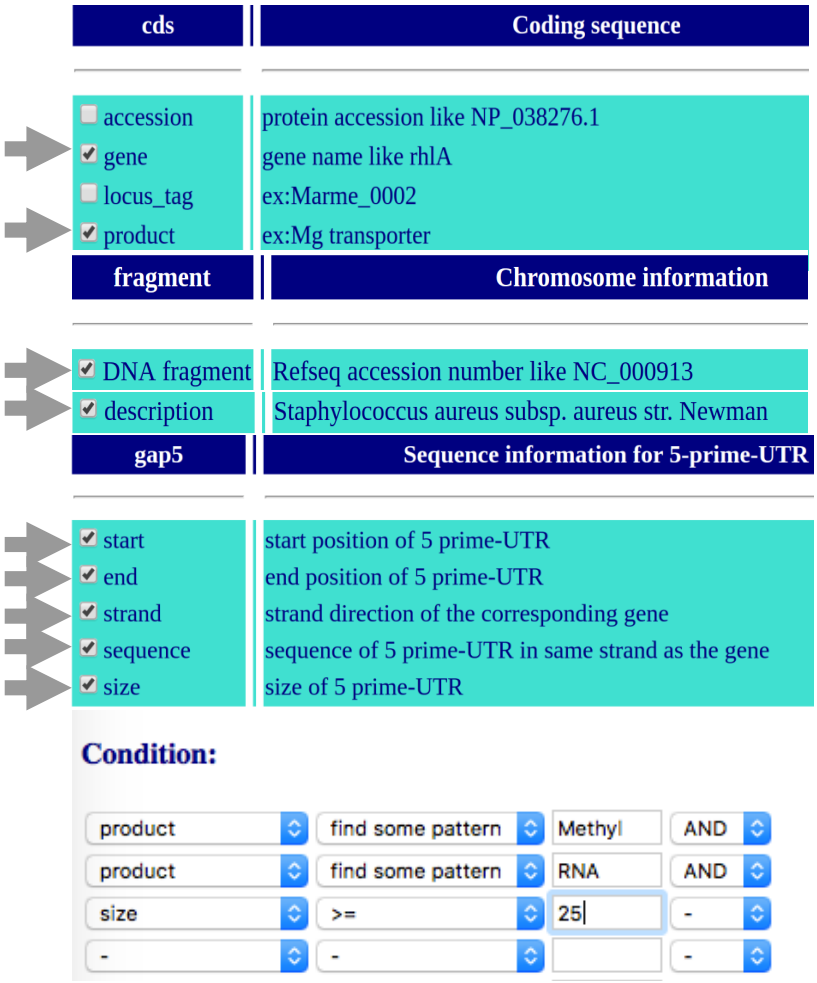

Figure 7. Capture d'écran de la base de données RiboGap

RiboGap est une base de données procaryote qui permet un accès facile aux séquences intergéniques. Dans la figure prise à partir de la page « Advanced_Search » représentée ici, certains champs de recherche ont été coupés pour mettre l'emphase sur l'exemple. La recherche ici permet l'accès aux séquences intergéniques en amont de gènes ayant dans la description de leur produit les mots clés « methyl » et « RNA ». Les mots clés peuvent être inscrits dans la section « condition » pour optimiser la recherche. RiboGap permet également une recherche en utilisant le langage MySQL (se référer au matériel supplémentaire de (Naghdi *et al*., 2017).

Problématique

Ce projet vise à identifier de nouveaux éléments régulateurs notamment des ARNnc chez les bactéries en utilisant des outils bio-informatiques. Les candidats potentiels seront ensuite confirmés et caractérisés expérimentalement. Étant donnée les limites décrites au point 7.2, nous proposons une nouvelle approche pour chercher de nouvelles structures d'ARNnc régulateurs dans les régions 5′ UTR en se basant sur la fonction (Figure 8). La base de données RiboGap sera utilisée pour faciliter l'accès aux séquences intergéniques. Le pipeline GraphClust permettra le traitement de plusieurs milliers de séquences pour trouver des regroupements conservés. Par la suite, des approches biochimiques et microbiologiques seront utilisées pour la confirmation et la caractérisation expérimentale.

Notre intérêt porte sur les fonctions d'ARNnc interagissant avec les produits de gènes en aval. En effet, bien qu'il y ait de nombreux précédents de ces derniers, tels que les *leaders* de protéines ribosomales, d'autres restent à découvrir. Un rare exemple serait une régulation via des enzymes modifiant l'ARN tel que les méthylases d'ARN. De plus, dans le cas où la régulation de ces dernières s'effectuerait via une méthylation directe de l'ARN, cela pourrait fournir un excellent modèle pour l'étude des modifications post-transcriptionnelles, un phénomène très répandu, mais dont l'impact est néanmoins souvent mal compris.

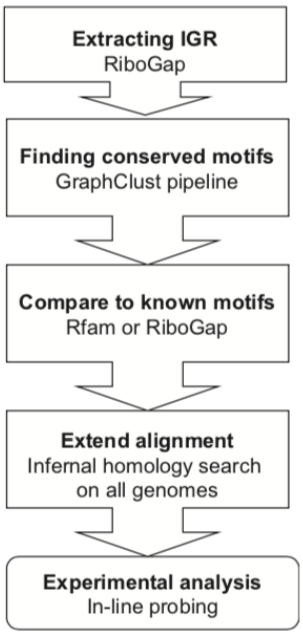

Figure 8. Représentation schématique du pipeline de découverte *de novo* d'ARNnc.

Tiré de (Naghdi *et al*., 2017).

Hypothèse et objectifs

Nous émettons l'hypothèse que les gènes qui codent pour des enzymes responsables de modifier l'ARN, telles que les méthylases d'ARN, pourraient s'autoréguler. En effet, l'autorégulation des méthylases est susceptible de se produire à travers sa région 5′ UTR soit par méthylation directe ou la liaison à celle-ci (Figure 9), comme cela pourrait être le cas pour des motifs tels 23S-méthyl et mraW présenté au point 6. Ainsi, ce projet a pour but de répondre aux objectifs suivants :

1. Trouver de bons candidats en lien avec la méthylation à tester au laboratoire suite aux analyses bio-informatiques.
2. Caractériser les candidats trouvés
    a. Confirmer la structure avec des essais d'*in-line probing*
    b. Tests *in vivo* avec un gène rapporteur
    c. Tests *in vitro*, lorsque confirmé, si la méthylation est directe

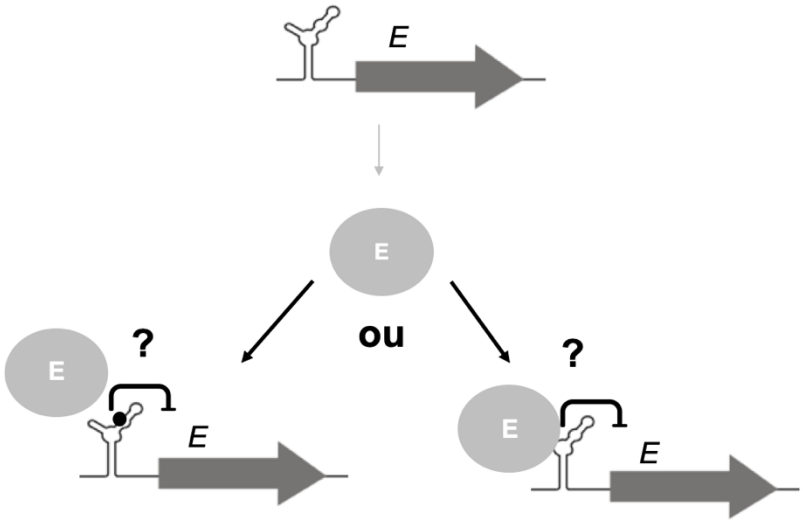

Figure 9. Représentation schématique de mécanismes potentiels d'autorégulation des méthylases d'ARN à travers leurs régions 5'UTR

Motif d'ARN candidat schématiquement illustré en amont du gène de la méthylase. Deux mécanismes potentiels sont illustrés, soit, à gauche, par la méthylation du motif ARN (le point noir illustre un hypothétique site méthylé), ou, à droite, par la liaison directe au motif d'ARN. E : enzyme

Chapitre 1 : Les sites de liaison de FadR et FabR en amont du gène *fabB* regulent aussi l'expression de l'enzyme modifiant l'ARNt U34 MnmC

The FadR and FabR binding sites upstream of *fabB* also regulate the expression of tRNA U(34) modifying enzyme MnmC

Smail, Katia[1] and Jonathan Perreault[1]

[1]Institut National de la Recherche Scientifique (INRS), Institut-Armand Frappier (IAF), 531 boul. des Prairies, Laval, Québec, Canada, H7V-1B7.

Correspondence should be sent to: Jonathan Perreault at jonathan.perreault@iaf.inrs.ca

Key words: *mnmC*, *yfcK*, *trmC*, 5-methylaminomethyl-2-thiouridine, mnm^5s^2U34, FadR, FabR, unsaturated fatty acid, tRNA modification regulation, gene expression

Katia Smail et Jonathan Perreault ont conçu l'étude et les expériences, ils ont aussi co-rédigé l'article. Katia Smail a effectué toutes les expériences.

Résumé

Les modifications de l'ARN sont nombreuses et peuvent être impliquées dans plusieurs fonctions. Les enzymes qui modifient l'ARN sont hautement conservées. Cependant, même si elles sont généralement connues, leurs fonctions exactes et leurs mécanismes de régulation sont souvent inconnus. MnmC est une double enzyme (oxydase / méthylase) impliquée dans la voie de modification complexe d'une uridine en position 34 (wobble) de certains ARNt. Les modifications à la position wobble sont importantes dans l'appariement des bases codon / anticodon et par conséquent le maintien du cadre de lecture. MnmC catalyse les deux dernières étapes de la voie de synthèse augmentant la spécificité de quelques ARNt à leurs codons respectifs. Ainsi, étant donné l'importance de cette modification et la complexité de la voie de synthèse, MnmC et les autres enzymes impliquées ont été bien caractérisées en termes de fonctions moléculaires et structures. Cependant, très peu est connu sur la régulation génétique du gène codant pour MnmC. Notre étude a permis, sur la base d'une caractérisation bio-informatique et des essais avec gène rapporteur, de démontrer la présence d'une régulation potentielle à deux voies dans la région 5'UTR de *mnmC* ainsi que de son promoteur.

Abstract

RNA modifications are numerous and can be involved in several functions. Enzymes that modify RNA are highly conserved. However, even though they are mostly known, exact functions and regulatory mechanisms are often unknown. MnmC is a double enzyme (oxidase / methylase) involved in a complex modification pathway of a uridine at position 34 (wobble) of some tRNAs. Modifications occurring in the wobble position are important in codon / anticodon base pairing, thus in maintaining the reading frame. MnmC catalyzes the last two steps of the biosynthesis pathway which makes tRNAs specific to certain amino acids. Thus, given the importance of this modification and the complexity of the biosynthesis pathway, MnmC and other enzymes involved have been well characterized in terms of molecular function and structure. However, very little is known about the genetic regulation of the gene encoding MnmC. Our study allowed us, based on a bioinformatics and characterization with reporter gene assays, to demonstrate a presence of a potential two-way regulation in the 5'UTR region of *mnmC* as well as its potential promoter.

1. Introduction

Transfer RNA (tRNA) are the most abundant class of noncoding RNA (ncRNA) and count a high number of post transcriptional modifications (Motorin & Helm, 2011). Each modification is introduced at a specific position by a specific enzyme carrying the possibility of an important role in the folding, stability, identity and tRNA function (Moukadiri *et al.*, 2014). Some modifications are complex and require several enzymes for their biosynthesis. Given the impact on tRNA function, many of these enzymes are highly conserved, as are the proteins involved in transcription and translation (Wang & He, 2014).

MnmC, formerly known as YfcK and TrmC (Bujnicki *et al.*, 2004), is a bifunctional enzyme involved in a complex modification of a wobble position (position 34, in the anticodon) of *Escherichia coli* tRNA specific to lysine, glutamate and arginine (Kim & Almo, 2013). It catalyses the last two steps of the biosynthesis of the hypermodified uridine 34; 5-methylaminomethyl-2-thiouridine (mnm^5s^2U34). This modification is part of xm^5s^2U type that limits the oscillating capacity of uridine at the position 34, preventing the misreading of codons ending with C or U (Moukadiri *et al.*, 2014). The biosynthesis pathway (figure 10) of mnm^5s^2U34 modification includes MnmA that catalyses the thiolation (s^2U), which facilitates the base pairing with A and G by stabilizing the C3′-endopuckering of the ribose (Moukadiri *et al.*, 2009), and the complex MnmE/MnmG that catalyzes the 5-carboxymethyl-aminomethyl uridine $cmnm^5U34$ and 5-carboxymethyl-aminomethyl-2-thiouridine, which are conserved from bacteria to eukaryotic organelles (Kim & Almo, 2013; Moukadiri *et al.*, 2014). *mnmG* is negatively regulated by ppGpp (Ogawa & Okazaki, 1991), as well as transcription factors Nac (Poggio *et al.*, 2002) and AnsC (Kolling *et al.*, 1988).

MnmC contains 2 domains: the flavin adenine dinucleotide dependent C-terminal domain (MnmC(o)-FAD dependent), responsible of the oxidation of the Cα-N bond in $cmnm^5s^2U34$ to nm^5s^2U34, and the S-adenosine-methionine N-terminal domain (MnmC(m)-SAM dependent) responsible for the methylation of $nm^5 s^2U34$ producing the final modification m $nm^5 s^2U34$ (Kim & Almo, 2013). The two domains are independent and have different substrates (Moukadiri *et al.*, 2014).

The wobble positions in tRNA often contain highly modified nucleotides which are important for the stabilization and the recognition of codon and anticodon base pairing. MnmC catalyzes the two final steps of an hypermodified uridine specific of $tRNA^{Gly,Glu,Arg\ and\ Lys}$ in *E. coli* and several other gram negative bacteria (Naghdi *et al.*, 2017). Given the importance of MnmC role in tRNA modification, its structure (Kim & Almo, 2013; Kitamura *et al.*, 2011; Roovers *et al.*, 2008) and molecular function including the complex pathway involving MnmC (Bujnicki *et al.*, 2004; Moukadiri *et al.*, 2014; Moukadiri *et al.*, 2009) has been well studied. However, little is known about *mnmC* gene regulation. In a previous study, the regulatory region of *mnmC* was thought to be a regulatory RNA structure within the 5′-UTR of the mRNA (Naghdi *et al.*, 2017). However, the most conserved regions of the motif overlapped conserved binding sites of FabR and FadR(Salgado *et al.*, 2012), which are known to regulate *fabB,* located on the other polarity. Here, we present a gene reporter assay and mutagenesis study of the 5′UTR region of *mnmC* suggesting the dual function of a tandem arrangement of the FabR and FadR binding sites.

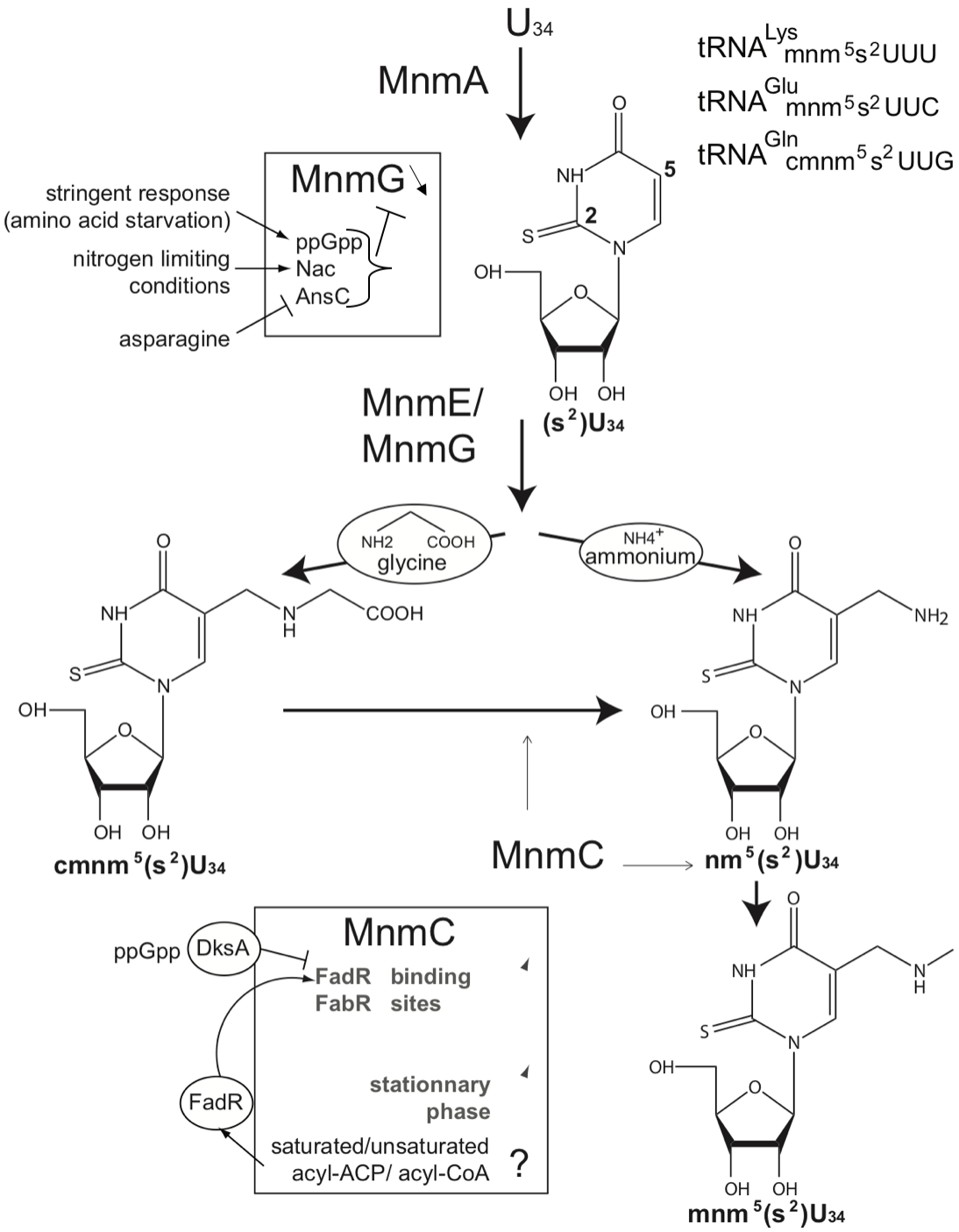

Figure 10. Pathway of tRNA modification by the *mnm* genes.

While complex, the U34 modification by *mnmA mnmE mnmG and mnmC* is relatively well characterized at the biochemical level. However, only *mnmG* had known regulators. Factors first noticed as regulators of this pathway are in bold gray.

2. Materials and methods

2.1. Bioinformatics determination of the regulatory regions

The conserved binding sites of FadR and FabR were found in the E. coli gene regulation database, RegulonDB (Salgado *et al*., 2012), while the mnmC promoter boxes were predicted with BPROM(Solovyev *et al*., 2011). To better study the regulatory sequences upstream of mnmC, we fetched from Ribogap (Naghdi *et al*., 2017) the corresponding intergenic regions (IGR) upstream of all the genes annotated as mnmC or for which the gene product contained the keyword "5-methylaminomethyl-2-thiouridine". One typical name used for mnmC consisted of "bifunctional tRNA (5-methylaminomethyl-2-thiouridine)(34)-methyltransferase MnmD/FAD-dependent 5-carboxymethylaminomethyl-2-thiouridine(34) oxidoreductase", but because the nomenclature varies, and because all the mnmC gene instances we could find had the keyword "5-methylaminomethyl-2-thiouridine", while no other gene had this keyword, we likely got the vast majority of sequences upstream of this gene. Redundancy of the DNA sequences was reduced to produce a fasta file and eliminate sequences with 98% or more identity (with a program on ExPASy). The resulting file was analyzed with MEME (Bailey *et al*., 2015). Conversely, the consensus DNA sequence binding site of FabR "GCGTACA[ACGT][ACGT]TGTACGC" was used in Ribogap to find all genes putatively regulated by this factor.

2.2. Bacterial strains, plasmids, primers and growth conditions

Escherichia coli strains and pRS414 plasmid details are shown in Table 1. Primers and oligonucleotides used are shown in supplementary data (Table S1). The 28-motif, part of *mnmC* 5′UTR region, was amplified by polymerase chain reaction (PCR) from the genomic DNA of *Escherichia coli* str. K12 substr. MG1655 (Hayashi *et al.*, 2006). The *mnmC* 5′UTR was cloned into pRS414 vector using a Gibson assembly cloning kit from New England BioLabs (NEB) under the control of lac promoter (Gibson *et al.*, 2010). Mutant versions of the *mnmC* 5′UTR were produced by PCR assembly and cloned into pRS414 vector the same way (FigureS4). All the inserts were verified by sequencing.

LB (Lysogeny Broth) was used for cultures and LB agar for plating of *E. coli*. Antibiotics were added for strain and plasmid selection at a concentration of 100 µg/ml ampicillin and 100 µg/ml kanamycin. Cell growth, with shaking at 37 °C, was measured using optical density of the cultures at 600 nm (OD_{600}).

Table 1. Strains and plasmids used in this study

Strain/ plasmid	Description	References
Escherichia coli		
MG1655	F-	(Hayashi *et al.*, 2006)
BW25113	F-, *Δ(araD-araB)567*, *ΔlacZ4787*(::rrnB-3), *λ*$^-$, *rph-1*, *Δ(rhaD-rhaB)568*, *hsdR514*	(Baba *et al.*, 2006)
JW5380-2 (***Δ****mnmC*)	F-, *Δ(araD-araB)567*, *ΔlacZ4787*(::rrnB-3), *λ*$^-$, *ΔtrmC732::kan*, *rph-1*, *Δ(rhaD-rhaB)568*, *hsdR514*	(Baba *et al.*, 2006)
Plasmids		
pRS414	ColE1, EcoRI, SmaI, and BamHI, ApR, *Δ*lacZp, *Δ*RBS(*lacZ)*, *Δ*ATG(*lacZ*).	(Simons *et al.*, 1987)
pKS-Ara	ColE1, EcoRI, SmaI, and BamHI, ApR, araBADp, *araC*, rrnBT1, T7Te	This study
pKS-mnmC	ColE1, EcoRI, SmaI, and BamHI, ApR, araBADp, *araC*, *mnmC*, rrnBT1, T7Te, lacZp, RBS(*mnmC)*, ATG(*mnmC*)	This study
pKS-mnmC-5WT	ColE1, EcoRI, SmaI, and BamHI, ApR, araBADp, *araC*, *mnmC*, rrnBT1, T7Te, lacZp, *mnmC-5UTR*, RBS(*mnmC)*, ATG(*mnmC*)	This study
pKS-mnmC-5M1	ColE1, EcoRI, SmaI, and BamHI, ApR, araBADp, *araC*, *mnmC*, rrnBT1, T7Te, lacZp, mutant1, RBS(*mnmC)*, ATG(*mnmC*)	This study
pKS-mnmC-5M2	ColE1, EcoRI, SmaI, and BamHI, ApR, araBADp, *araC*, *mnmC*, rrnBT1, T7Te, lacZp, mutant2, RBS(*mnmC)*, ATG(*mnmC*)	This study
pKS-mnmC-5M3	ColE1, EcoRI, SmaI, and BamHI, ApR, araBADp, *araC*, *mnmC*, rrnBT1, T7Te, lacZp, mutant3, RBS(*mnmC)*, ATG(*mnmC*)	This study

2.3. LacZ expression measurements

The activity of the reporter gene *lacZ* was measured using Miller assays. Overnight cultures were diluted to OD_{600} = 0.05 in fresh LB media containing ampicillin and incubated at 37 °C with shaking. The incubation was stopped at different time points. 1 ml of the culture was centrifuged and resuspended in 1 ml of Z-buffer [6.1 g $Na_2HPO_4.7H_2O$ (0.06 M) 5.5 g $NaH_2PO_4.H_2O$ (0.04 M), 0.75 g KCL (0.01 M), 0.246 g $MgSO_4.7H_2O$ (0.001 M), 2.7 ml β-mercaptoethanol (0.05 M), pH = 7.0)]. 25 µl of sodium dodecyl sulfate (SDS) 0.1% and 50 µl of chloroform were added. The tubes were vortexed for 30 seconds and incubated for 5 min at room temperature (RT). 200 µl of ortho-Nitrophenyl-β-galactoside (ONPG, 4 mg/ml) was added to the tubes and the time recorded. The reaction was stopped by adding 200 µl of sodium carbonate (Na_2CO_3) 1M. The tubes were vortex for 30 seconds and centrifuged for 5 min at 13,000 rpm at RT. The optical density of the supernatant was measured at 420 nm (OD_{420}).

The Miller units were calculated following the formula:

$$Miller\ Units = \frac{(1000 * OD420)}{(t * v * OD600)}$$

t: time in minutes, v: volume of culture used in ml.

3. Results and discussion

3.1. Sequence consensus

We got sequences upstream of *mnmC* genes from the latest version of Ribogap (which includes all complete prokaryotic genomes from NCBI, see Table S2), and, after eliminating redundant sequences, we could select 974 IGR with at least 2% divergence (Table S3). All the sequences were proteobacteria, matching the known phylogeny of *mnmC*. Three of the top five conserved motifs uncovered by MEME were in *Enterobacteriacea* and matched the IGR sequence of *mnmC* in *E. coli* (Figure 11). The #1 ranking motif corresponds to the tandem arrangement of the FabR-FadR binding sites, the #2 motif corresponds to the -10 -35 boxes of the predicted promoter of *mnmC* (Figure 11, in green). The #5 motif corresponds to the 5′-UTR of *fabB*. The #3 and #4, as well as other top 15 motifs, correspond to species that do not have recognizable FabR or FadR binding sites (Fig. S1).

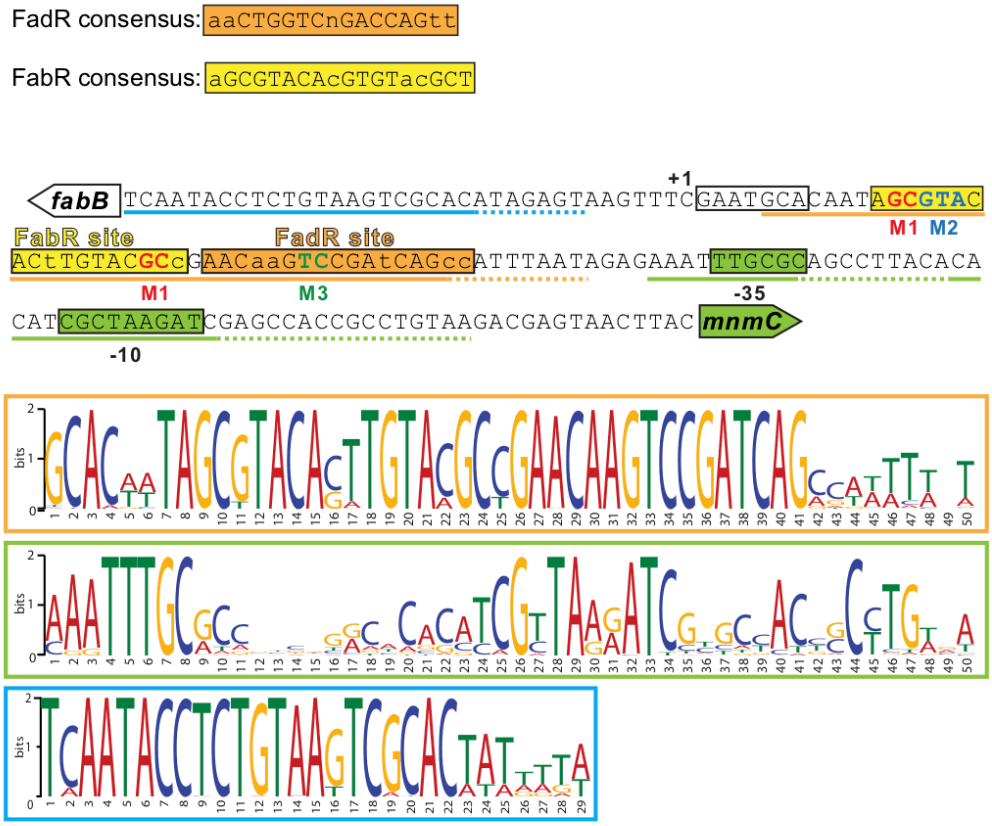

Figure 11. Conserved regulatory elements of the intergenic region (IGR) of *fabB* and *mnmC*.

The known consensus of FabR and FadR are shown on the top and their corresponding site in the IGR are shown in yellow and orange, respectively, with bases not fitting the consensus in lower case. Putative promoter regions are boxed in white (*fabB*) and in green (*mnmC*). The motifs found by MEME are underline with the same color as their boxes, with the more conserved regions underlined with a solid line. Position of mutants assayed are also pictured in bold and color.

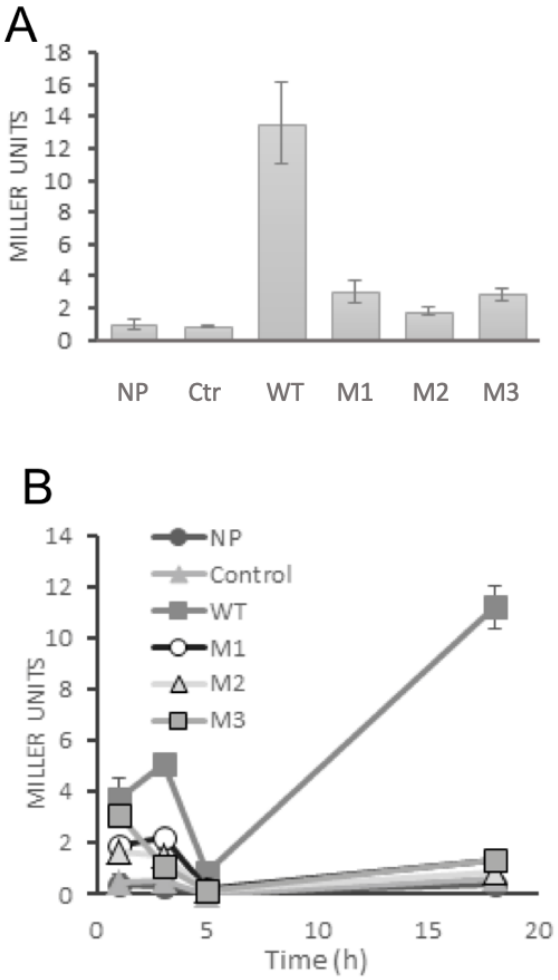

Figure 12. Characterization of the regulating potential of the *mnmC* 5′-UTR using the *lacZ* reporter gene.

[A] LacZ expression under the control of the *mnmC* WT UTR and mutant versions. [B] Miller assays following the growth curve. NP: no plasmid present (no *lacZ*). Control: plasmid pKS-mnmC does not contain the *mnmC* 5′-UTR. WT: plasmid pRS414 contains the *mnmC* WT UTR. M1: Mutant 1. M2: Mutant 2. M3: Mutant 3.

In parallel with the MEME search for consensus motifs, we used the known FabR consensus to find genes associated with it. We found almost the same number of instances upstream of *fabB* (187) as for *mnmC* (180), indicating a conserved synteny, at least for species with the FabR/FadR regulon. We also found many hits for several other genes, especially for *fabA* and *fabR* (Table S4). The latter suggests auto-regulation in γ-Proteobacteria genera such as *Shewanella*, *Idiomarina* and *Kangiella*, in contrast with another γ-Proteobacteria, *E. coli*, which does not exhibit *fabR* autoregulation (Feng & Cronan, 2011)).

3.2. LacZ expression measurements

In a previous study, we described a structured RNA element, (motif 28, see reference (Naghdi *et al.*, 2017) for more details) in *mnmC* 5'UTR region, thought to be a repressing RNA structure. Thus, in order to test our hypothesis, we cloned *mnmC* 5'UTR region in pRS414 vector (Table 1) downstream of a *pLac* promoter and performed Miller assays in presence of isopropyl β-D-1-thiogalactopyranoside (IPTG), activator of *pLac,* in the media. However, instead of a repression, the plasmid containing *mncmC* wild type UTR had more expression of its *lacZ* reporter gene than the mutants and the plasmid control (which does not contain the *mncmC* UTR). In a following assay, we performed the experiment without adding IPTG in the media. The results showed an activity of *lacZ* in presence of the *mncmC* wild type UTR and a decreased expression with the mutants (Figure 12.A) which led us to consider the conserved motif of the *mncmC* UTR as a promoter of the enzyme *mnmC,* rather than as an RNA regulatory element.

The *mnmC* gene shares its 5′-UTR, but on the opposite polarity, with the *fabB* gene, involved in unsaturated fatty acid biosynthesis. The *fabB* gene encodes for β-ketoacyl-ACP synthase I along with *fabA*, which encodes for β-hydroxydecanoyl-ACP dehydratase/isomerase, both activities are essential for monounsaturated fatty acid synthesis (Marrakchi *et al.*, 2002). The 5′-UTR region of *fabB* contains two binding sites for transcription regulators FabR (fatty acid biosynthesis regulon) and FadR (fatty acid degradation regulon). FadR positively regulates *fabB* transcription while FabR acts as a transcription repressor of *fabB* (My *et al.*, 2015; Zhang *et al.*, 2002). The DNA binding of FabR and FadR is dependent on saturated and unsaturated acyl-ACP or acyl-CoA binding (Dirusso *et al.*, 1992) (Zhu *et al.*, 2009). Mutants 1,2,3 in this study affect positions in FadR and FabR binding sites (Fig. 18), causing a significant decrease in *lacZ* gene

expression compared to the wild type version (Fig. 18). Being palindromic sequences, the results suggest that FadR and FabR might have an impact on *mnmC* gene expression. Also, using BPROM, a bioinformatics tool for bacteria promoter prediction (Solovyev *et al.*, 2011), we were able to predict a promoter region for *mnmC* gene downstream of FadR and FabR binding sites (Figure 11).

MnmC is a tRNA modifying enzyme for a few tRNAs, while FabB is a fatty acid synthase, there is no obvious regulatory link between the two. However, it has been shown that FadB regulation increases during stationary phase (Farewell *et al.*, 1996), which corresponds to the results shown in Figure 12.B, representing the activity of the reporter gene over time for which there is a decrease during the growth phase followed by a significant increase during the stationary phase. This is also supported by a previous report indicating that mnm5U(34) modification varies depending on growth phase (Moukadiri *et al.*, 2014). It is also possible that the activator action of FadR would be favored because it is closer to the *mnmC* promoter compared to the binding site of the FabR repressor.

The pathway of U34 modification also involves *mnmA*, *mnmE* and *mnmG*, the latter being the only gene for which we could find published information on its regulation (Figure 10). Interestingly, ppGpp appears to act on both *mnmG* and *mnmC,* even if indirectly in the latter case by repressing FadR expression through DksA. This fits well our mutational data of the FadR binding site, since inhibition of FadR (an activator of *mnmC*) would repress *mnmC*, which means that ppGpp represses both genes, albeit through different pathways.

4. Acknowledgements

This work was funded by Natural Sciences and Engineering Council of Canada (NSERC) [418240 to J.P.]. K.S. was funded by NSERC and the Armand-Frappier foundation.

5. Supplementary material

Table S1. Oligonucleotides used for amplification

Name	Product	sequence
KS_lac28F01	PCR 28-motif region	**attggggatcggaattCCCGGGTTTACACTTTATGCTTCCGGCTCGTATGTTGTCAATACCTCTGTAAGTCGCACATAG**
KS_lac28F02		**TAAAACGACGGGATCCCCGGGGGAGTAGTGTTTCACGTAAGTTACTCG**
KS_F1_mut1	PCR assemply of mutant 1	**CCAGGAATTGGGGATCGGAATTCCCTTTACACTTTATGCTTCCGGCTCGTATGTTGTCAA**
KS_R1_mut1		**TTGTGCATTCGAAACTTACTCTATGTGCGACTTACAGAGGTATTGACAACATACGAGCCG**
KS_F2_mut1		**AGAGTAAGTTTCGAATGCACAATACGTACACTTGTACCGGAACAAGTCCGATCAGCCAT**
KS_R2_mut1		**GCGATGTGTGTAAGGCTGCGCAAATTTCTCTATTAAATGGCTGATCGGACTTGTTC**
KS_F3_mut1		**CAGCCTTACACACATCGCTAAGATCGAGCCACCGCCTGTAAGACGAGTAACTTACGTGAA**
KS_R3_mut1		**GACGTTGTAAAACGACGGGATCCCCGGAGTAGTGTTTCACGTAAGTTACTCGTCTTAC**
KS_F1_mut2	PCR assemply of mutant 2	**CCAGGAATTGGGGATCGGAATTCCCTTTACACTTTATGCTTCCGGCTCGTATGTTGTCAA**
KS_R1_mut2		**TTGTGCATTCGAAACTTACTCTATGTGCGACTTACAGAGGTATTGACAACATACGAGCCG**
KS_F2_mut2		**AGAGTAAGTTTCGAATGCACAATAGCGTACACTTGTACCGCGAACAAGCTGATCAGCCAT**
KS_R2_mut2		**TTAGCGATGTGTGTAAGGCTGCGCAAATTTCTCTATTAAATGGCTGATCAGCTTGTTC**
KS_F3_mut2		**GCCTTACACACATCGCTAAGATCGGCCACCGCCTGTAAGACGAGTAACTTACGTGAAACA**
KS_R3_mut2		**GACGTTGTAAAACGACGGGATCCCCGGAGTAGTGTTTCACGTAAGTTACTCGTCT**
KS_F1_mut3	PCR assemply of mutant 3	**CCAGGAATTGGGGATCGGAATTCCCTTTACACTTTATGCTTCCGGCTCGTATGTTGTCA**
KS_R1_mut3		**CGAAACTTACTCTATGTGCGACTTACAGAGGTATTGACAACATACGAGCCGGAAG**
KS_F2_mut3		**TCGCACATAGAGTAAGTTTCGAATGCACAATAGCCATCATTGTACGCCGAACAAGTCCG**
KS_R2_mut3		**GCGATGTGTGTAAGGCTGCGCAAATTTCTCTATTAAATGGCTGATCGGACTTGTTCGGCG**
KS_F3_mut3		**CAGCCTTACACACATCGCTAAGATCGAGCCACCGCCTGTAAGACGAGTAACTTACGTGAA**

Figure S1. Motifs found in the *mnmC* IGR with MEME.

Motifs #1, 2 and 5 are also in Fig. 1 in the main text. Motifs #3,4,7,8,9,11 and 15 are found in *Burkholderiaceae*, motifs #6 and 10 are found in *Pseudomonas*, motifs 12,13 and 14 are found in other γ-Proteobacteria.

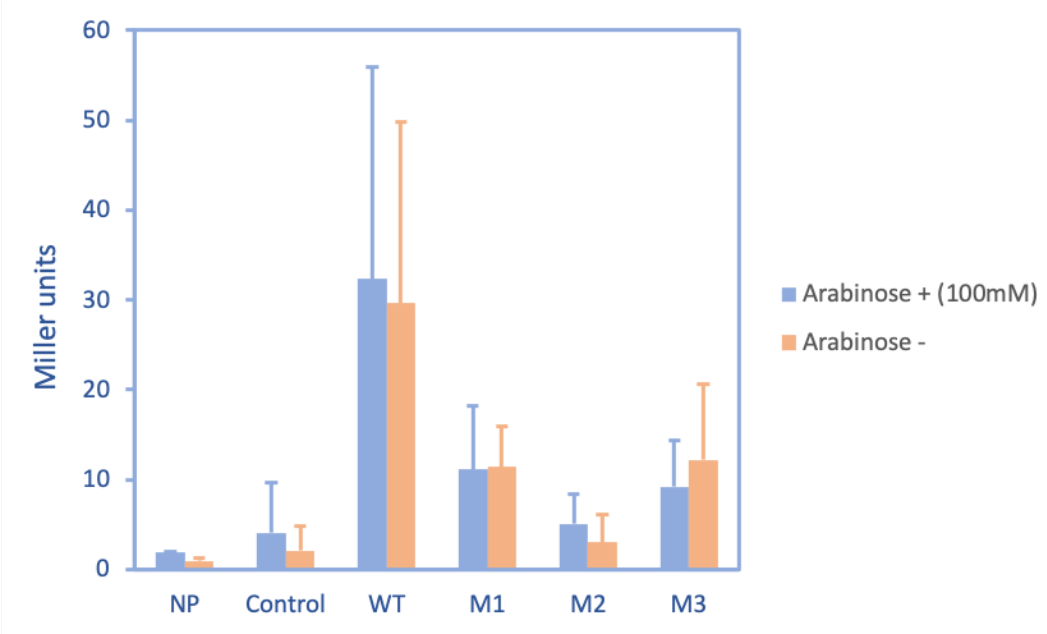

Figure S2. Characterization of the regulating potential of 28-motif region using *lacZ* reported gene

This test was performed with induction of lacZp with IPTG and overexpressing or not *mnmC* gene controlled by araBADp with arabinose to a final concentration at 100mM in the culture media.

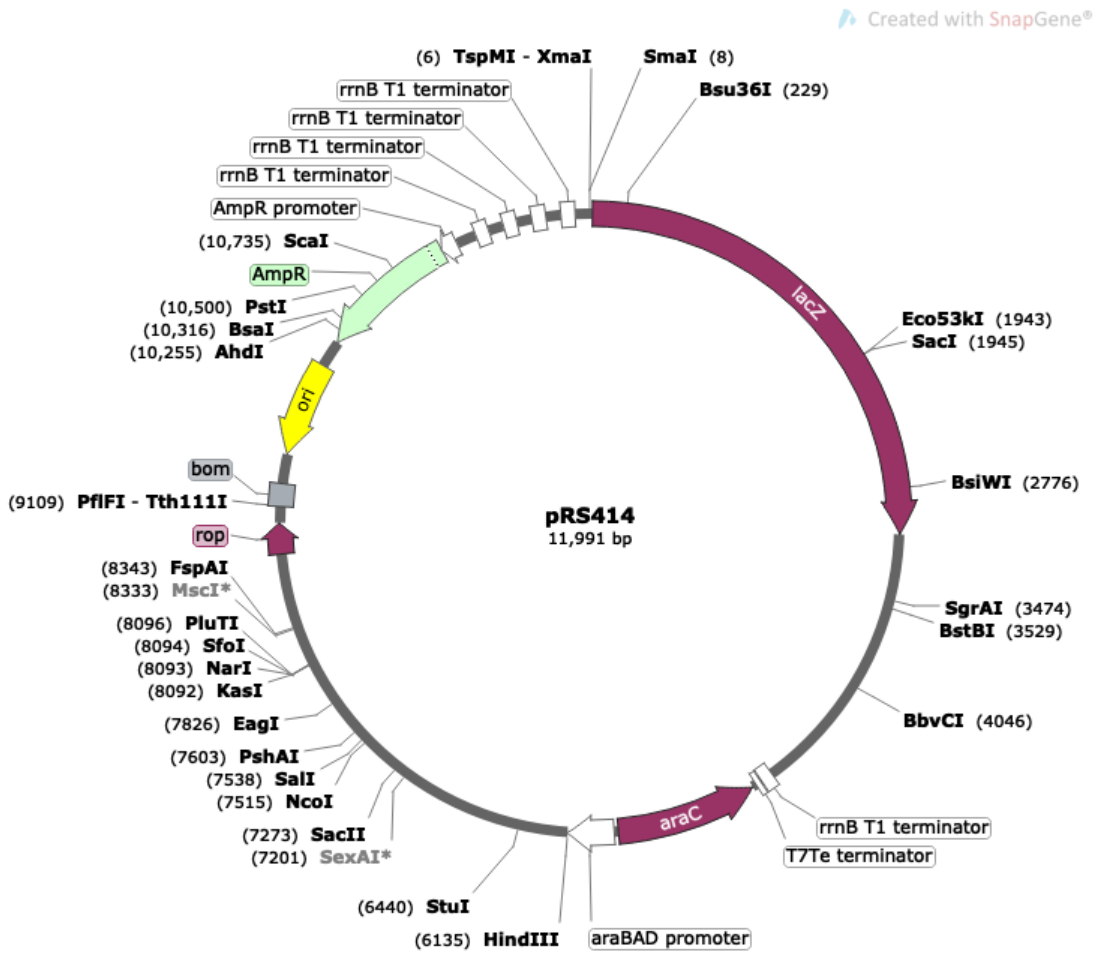

Figure S3: Map of pKS-Ara.

An inducible araBADp promoter with its araC gene have been cloned in pRS414 to allow arabinose-mediated induction of any chosen gene independently from lacZ. The rrnBT1 and T7Te terminators were added to prevent interference with the lacZ reporter gene.

This plasmid was constructed for reporter assays to evaluate the effect of a given gene on a cis-regulatory region. The fact that *lacZ* lacks a start codon allows to look at regulation mediated at the level of transcription or translation, and thus to evaluate both promoter regions or RNA motifs within the 5′-UTR.

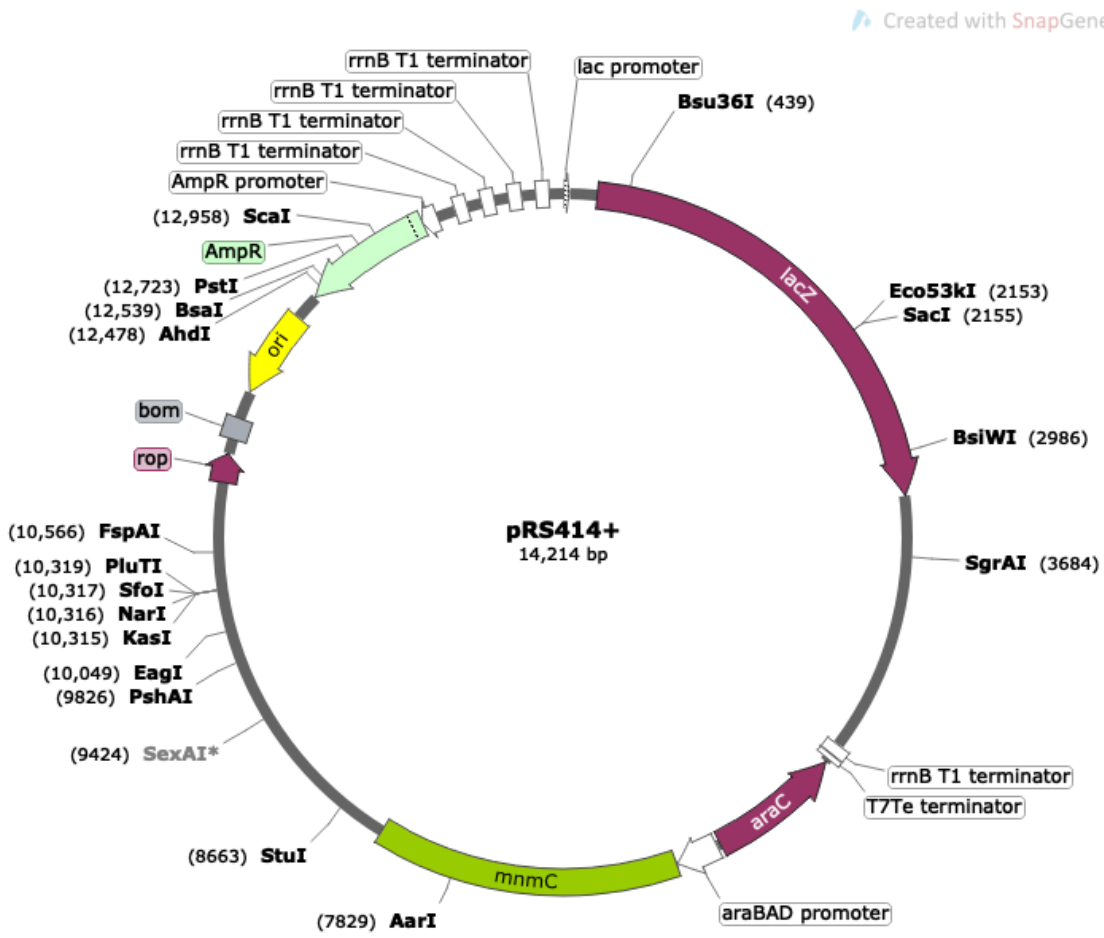

Figure S4: Map of the plasmids pKS- WT, Control, M1, M2, M3

This plasmid contains lacZp promoter, insertion regions (WT, Control, M1, M2, M3) and mnmC gene. This plasmid was used in the Miller assays.

Chapitre 2 : Nouveaux motifs régulateurs potentiels

Pour répondre à notre hypothèse sur le potentiel d'autorégulation, via leur région intergénique en 5′, des enzymes qui modifient l'ARN, nous avons utilisé la méthode publié dans (Naghdi *et al.*, 2017) (voir annexe1). Nous nous sommes intéressés particulièrement aux méthylases d'ARN, la méthylation étant une modification très répandue et ayant des implications fonctionnelles variées (voir sections 5 et 6 de l'introduction). Le but du projet étant de chercher de nouvelles structures d'ARN régulatrices en amont de méthylases d'ARN à tester au laboratoire pour évaluer la présence potentielle d'un tel mécanisme de régulation.

1. Matériel et méthode

1.1. RiboGap

Pour extraire toutes les séquences en amont de gènes reliés à la méthylation, plusieurs combinaisons de mots clés ont été ajoutées dans la section conditions dans RiboGap (Figure 7), afin d'avoir accès à toutes les séquences possibles. Les mots clés utilisés étaient : « methyl » et « RNA », « methyl » et « rRNA », « methyl » et « tRNA » et finalement « methyl », « methyltransferase » et « methylase ». Un seuil minimal de 25 nucléotides de taille a également été déterminé afin d'éliminer les séquences trop courtes moins susceptible de former des structures fonctionnelles. RiboGap génère un fichier tabulé (.csv) et un fichier fasta (.fa) des régions intergéniques correspondant aux critère décrits. Des milliers de séquences ont été récupérées pour chaque combinaison de mots clés.

1.2. GraphClust

Nous avons personnalisé certains paramètres lors de l'utilisation de GraphClust, notamment le paramètre BlastClust et le nombre d'itérations. Tel que discuté plus haut, BlastClust permet d'éliminer les séquences ayant une similarité dépassant un certain seuil (90% par défaut). Dans le cas de notre étude, nous avons fixé le paramètre à 97%, ce qui signifie que nous avons été plus permissif quant au degré de similarité que peuvent avoir les séquences entre elles, car nous avons remarqué qu'avec l'utilisation du paramètre par défaut (90%) la plupart des structures prédites par GraphClust étaient déjà connues. En conséquence, la fixation du paramètre à 97% a eu comme résultat la prédiction de *clusters* avec des alignements contenant plus de conservation de séquences que de co-variations, ce qui n'est pas nécessairement un désavantage car cela permettrait de

chercher des motifs moins évidents à trouver. Aussi, les résultats que nous avons obtenus démontrent que l'augmentation du filtre à 97% ne nous a pas empêché de trouver des motifs avec une distribution phylogénétique étendue. Le second paramètre modifié est le nombre d'itérations, que nous avons fixé à 15 au lieu de 2 par défaut, ce qui permet d'augmenter le nombre de prédictions. Le reste des paramètres a été utilisé par défaut. L'utilisation de GraphClust se fait en ligne de commande comme suit :

MASTER_GraphClust.pl --root run_test_1 --fasta my_seqs.fasta --config config.default_global –verbose examples/config.default_global.

1.3. Sélection des candidats

L'analyse des candidats obtenus par GraphClust s'est faite manuellement. Pour ce faire, nous avons déterminé des critères de sélection, à savoir: la conservation des séquences de l'alignement, la présence de co-variations et l'absence d'annotation dans les bases de données telles que Rfam ou RiboGap (Kalvari *et al*., 2018; Naghdi *et al*., 2017). Les structures des candidats sont également comparées aux motifs connus en utilisant l'outil en ligne CMCompare (les paramètres ont été laissés par default) (Eggenhofer *et al*., 2013).

1.4. Infernal

L'outil Infernal (*inference of RNA alignment*) permet de chercher dans des bases de données ADN des structures d'ARN et les similarités de séquences (Nawrocki & Eddy, 2013). Ainsi, pour chaque motif candidat sélectionné, nous avons utilisé l'outil Infernal pour chercher les instances dans le génome de toutes les bactéries. Infernal est un outil qui s'utilise en ligne de commande, la démarche et la méthode utilisées ont été décrite dans (El Korbi *et al*., 2014).

1.5. BPROM

Afin de s'assurer que les motifs sélectionnés ne chevauchent pas des régions promotrices, à l'image des résultats décrits au chapitre 2, nous avons analysé les séquences ADN de chaque motif en utilisant l'outil en ligne BPROM (*prediction of bacterial promoters* (Solovyev *et al*., 2011). Ceci a permis de filtrer davantage les résultats obtenus. En effet, dans le cas où un motif candidat chevauche un promoteur, celui-ci est éliminé du fait qu'il ne peut être un motif ARN, faute d'être transcrit, et qu'une prédiction de motif à ce niveau serait plus vraisemblablement due à la conservation du promoteur et non à la présence d'une structure ARN conservée.

2. Résultats

L'utilisation de différentes combinaisons de mots clés a donné lieu à différents résultats soit directement liés aux fonctions des gènes ou indirectement via leurs produits ou substrats. Exemple, l'utilisation de la combinaison « methyl », « methyltransferase » et « methylase » a permis d'extraire le plus grand nombre de séquence avec une grande diversité dans les gènes en aval, qui ne sont pas toujours liés à la fonction de méthylation de l'ARN, tels : **methyl**-accepting chimiotaxis protein McpB, C-5 cytosine-specific DNA **methyltransferase** et N-6-adenine-specific **methylase**. Tandis que les combinaisons « methyl » et « RNA », « methyl » et « rRNA », « methyl » et « tRNA » ont généré moins de résultats en termes de nombre séquences, mais dont les gènes en aval sont plus spécifiquement liés à la méthylation des ARN. Ainsi, nous avons séparé les résultats obtenus en deux catégories, à savoir : des motifs liés à des méthylases d'ARN et des motifs liés à des méthylases d'ADN. Dans chaque catégorie, les gènes en aval des motifs sont soit directement lié à la fonction de méthylation, c'est-à-dire codent pour des méthylases d'ARN ou d'ADN, ou bien indirectement liés, c'est-à-dire codent pour des protéines dont le substrat ou le produit final sont méthylés dans une voie de synthèse.

2.1. Motifs liés à des méthylases d'ARN

Dans cette catégorie, nous avons retenus trois motifs, à savoir : les motifs lasT, mnmG et rlmH. LasT et RlmH sont des méthylases d'ARN tandis que MnmG est une protéine impliquée dans une voie de synthèse d'une méthylation sur une uridine hyper-modifiée.

2.1.1. Motif mnmG

Le motif mnmG a été prédit par GraphClust (Figure 13A et 13B) dans la région 5′UTR du gène *mnmG,* aussi nommé *gidA*. Ce dernier code pour la protéine MnmG impliquée dans l'hypermodification (5-methylaminomethyl-2-thiouridine) d'une uridine à la position 34 (wobble) de certains ARNt, stabilisant l'appariement des bases U.G au niveau de la position wobble (Kurata *et al*., 2008). Le motif est bien conservé et présente plusieurs co-variations (Figure 13C et 13D). MnmG agit en dimère avec MnmE dans une voie de synthèse complexe impliquant d'autres protéines (voir chapitre 2). Le gène *mnmG* fait partie de l'opéron asnC-mioC-mnmG-rsmG localisé dans la région de l'origine de réplication chez *Escherichia coli* (Figure 13.E). AsnC est un régulateur transcriptionnel qui active notamment le gène *asnA* impliqué dans la synthèse de

l'asparagine, dont la concentration influence l'activité de la protéine AsnC, mais également le gène *mnmG* de façon post-transcriptionnelle (Kolling & Lother, 1985). MioC est une protéine liant la flavine mononucléotide (FMN) impliquée dans la régulation de la division cellulaire (Lies *et al*., 2015). RsmG est une méthyltransferase de l'ARNr S16 (Okamoto *et al*., 2007). Sous l'influence de plusieurs promoteurs, l'opéron présente plusieurs unités de transcriptions, à savoir : asnC, asnC-mioC, asnC-mioC-mnmG (Gielow *et al*., 1988), mioC (Mendoza-Vargas *et al*., 2009) , rsmG et mnmG- rsmG (Benitez-Paez *et al*., 2012).

Les résultats obtenus avec l'outil BPROM montrent un chevauchement du motif mnmG avec un possible promoteur (Figure 13.F), ce qui est normalement un critère d'élimination. Par ailleurs, en mettant le motif dans la perspective de sa région intergénique, nous avons remarqué que le promoteur annoté de *mnmG* (mnmGp) (Gielow *et al*., 1988) se trouvait en aval du motif ce qui rend une co-transcription avec le gène impossible avec ce promoteur. Toutefois, tel que discuté ci-haut, *mnmG* se trouve dans un opéron et il fait partie de plusieurs unités de transcription, notamment l'unité asnC-mioC-mnmG sous l'influence du promoteur asnCp. Ce dernier est également responsable de la transcription de asnC-mioC et asnC seul (Kolling & Lother, 1985). D'autre part, mioCp permet la transcription unique de *mioC* (Mendoza-Vargas *et al*., 2009). Ainsi, en tenant compte des combinaisons possibles des unités de transcription à partir du même promoteur, il n'est pas à exclure que le motif mnmG aie un rôle à jouer dans la régulation transcriptionnelle du gène *mnmG* comme ARN régulateur en cis. D'autre part, tel que discuté plus haut, le gène *mnmG* est régulé de manière post-transcriptionnelle par la protéine AsnC. Ainsi, considérant la présence du motif mnmG comme une potentielle structure ARN régulatrice, on peut émettre l'hypothèse que celle-ci jouerait un rôle dans la régulation de *mnmG* par AsnC. En effet AsnC est un régulateur post-transcriptionnel de *mnmG*, dont l'action sur l'ARNm pourrait se faire via le motif mnmG à l'image du mécanisme utilisé par les *leaders* de protéines ribosomales décrit dans l'introduction au point 4 et qui peuvent également affecter l'expression des gènes de leur opéron.

La localisation de *mnmG* étant adjacente à l'origine de réplication (oriC) chez *Escherichia coli*, sa transcription ainsi que celle du gène *mioC* furent considérés nécessaires à la réplication de l'ADN (Ogawa & Okazaki, 1991). Il a été démontré par la suite que l'effet de la transcription de ces deux gènes sur la réplication est minime et ne se produit que sous certaines conditions sous-optimales (Bates *et al*., 1997). Finalement, Lies et al. ont montré que *mnmG* et *mioC* influencent positivement

la division cellulaire, mais à l'échelle protéique car l'expression en *trans* de ces protéines peuvent complémenter le phénotype des KO lié à la division (Lies *et al*., 2015). Ainsi, considérant le fait que le motif mnmG est potentiellement un motif ARN se trouvant dans la région de l'origine de réplication, une influence lors de la transcription sur la réplication et la division cellulaire serait peu probable.

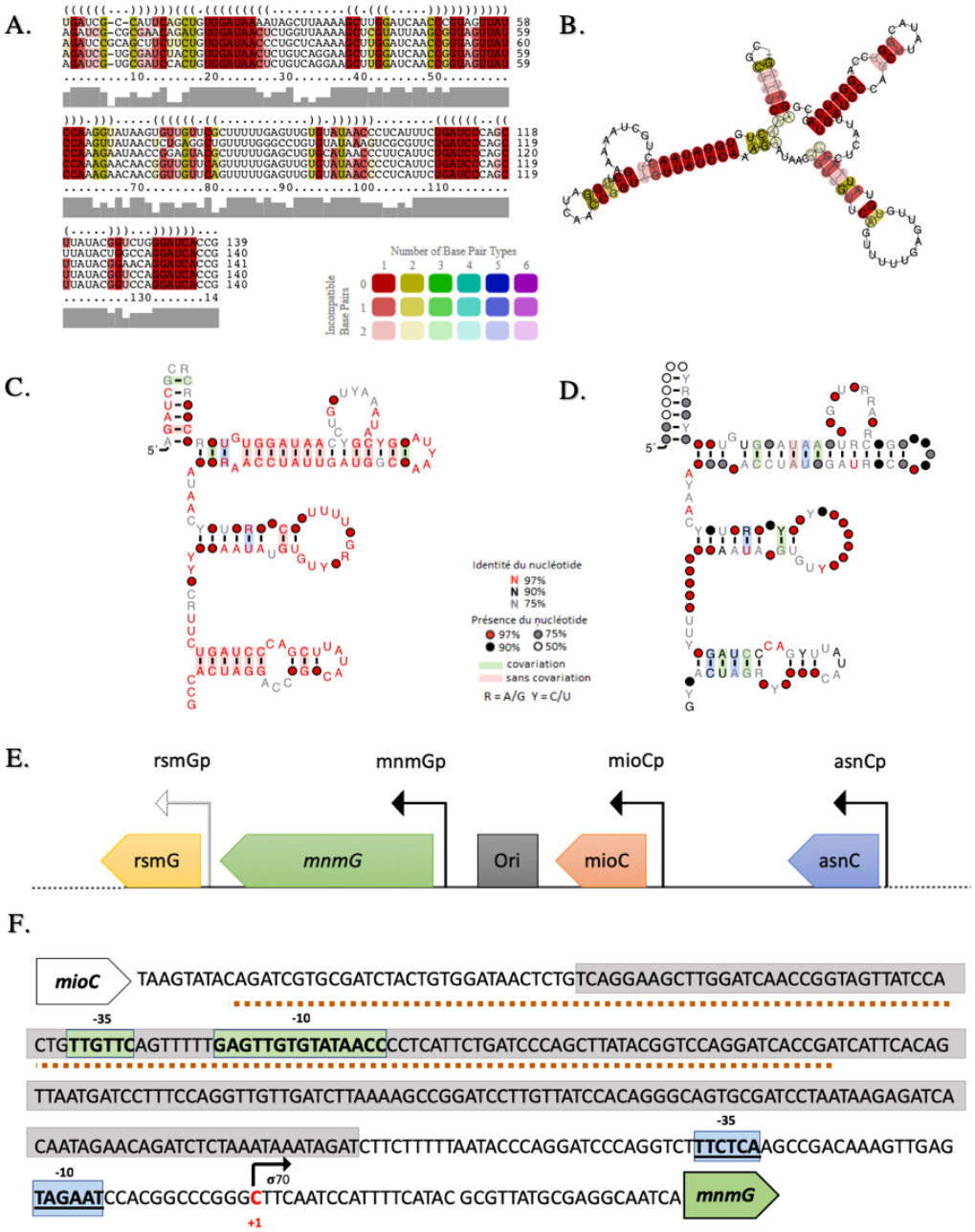

Figure 13. Motif mnmG

A) Alignement des séquences du motif prédit par GraphClust: rouge: paires de bases conservées, jaune: co-variations. B) Structure secondaire prédite par GraphClust, ou C) dessinée avec l'outil R2R de l'alignement en [A], et D) avec un alignement plus exhaustif obtenu avec l'outil Infernal. E) Région de l'origine de réplication chez *E. coli*. F) Région 5'UTR du gène *mnmG* : vert: promoteur prédit avec l'outil BPROM, bleu: promoteur annoté de *mnmG*, orange: motif mnmG, gris: origine de réplication du chromosome. La position du motif dans la région intergénique du gène *mnmG* montre un chevauchement avec un potentiel promoteur et avec l'origine de réplication du chromosome.

2.1.2. Motif lasT

Le motif lasT (Figure 14A et 14B) a été prédit dans la région 5′UTR du gène *lasT* (annoté aussi *yjtD*) qui est une méthyltransferase appartenant à la super famille SPOUT (Anantharaman *et al*., 2002). Le motif présente plusieurs co-variations et une structure bien conservée (Figure 14C). La structure représentée à la Figure 14D montre la que la seconde tige du motif n'est pas toujours présente, celle-ci pourrait correspondre à une tige accessoire. La présence de tiges accessoire est commune chez les ARN noncodants. Par ailleurs, On retrouve très peu d'informations sur le fonctionnement de l'enzyme et sa régulation, annoté comme méthyltransferase d'ARNr mais également d'ARNt (Purta *et al*., 2006). Le promoteur n'étant pas annoté, nous avons analysé la région intergénique du gène *lasT* avec l'outil BPROM. Un promoteur a été prédit en amont du motif lasT, ce qui permettrait une co-transcription avec le gène (Figure 14E). Ainsi, étant donné la conservation du motif et sa position dans l'UTR il serait un candidat intéressant à tester.

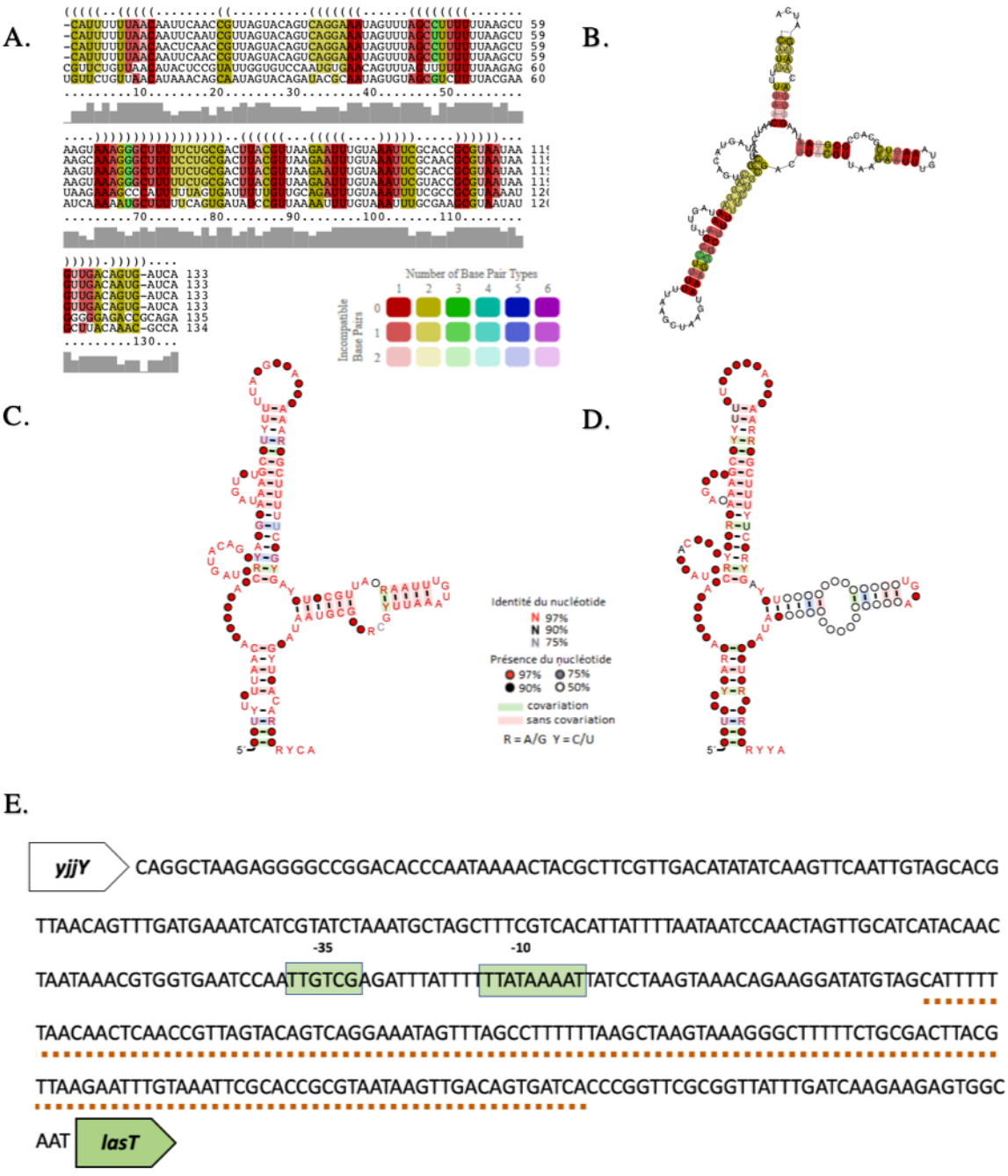

Figure 14. Motif lasT

A) Alignement des séquences du motif prédit par GraphClust : en rouge les séquences de paires de bases conservées, les autres couleurs représentent les co-variations. B) Structure secondaire prédite par GraphClust. C) Structure secondaire dessinée avec l'outil R2R de l'alignement en [A]. D) Structure secondaire dessinée avec l'outil R2R des résultats obtenus avec l'outil Infernal. E) Région 5'UTR du gène *lasT* : en Vert : promoteur prédit avec l'outil BPROM, en Orange la séquence qui correspond au motif lasT.

2.1.3. Motif rlmH

Le motif rlmH (Figure 15A et 15B) est prédit en amont du gène *rlmH* chez des bactéries à gram positif telles : *Bifidobacterium*, *Lactobacillus*, *Melissococcus* et *Acetobacterium*. Le motif est bien conservé et présente plusieurs co-variations (Figure 15C et 15D). Les résultats de prédiction avec PBROM ne montrent aucun chevauchement avec un promoteur. RlmH est une méthyltransferase de l'ARNr 23S, membre de la super famille SPOUT. Elle méthyle un résidu de pseudo-uridine à la position 1915. Une délétion de l'enzyme cause une légère réduction de la croissance, malgré ce modeste phénotype, ce gène est conservé chez les bactéries (Purta *et al*., 2008).

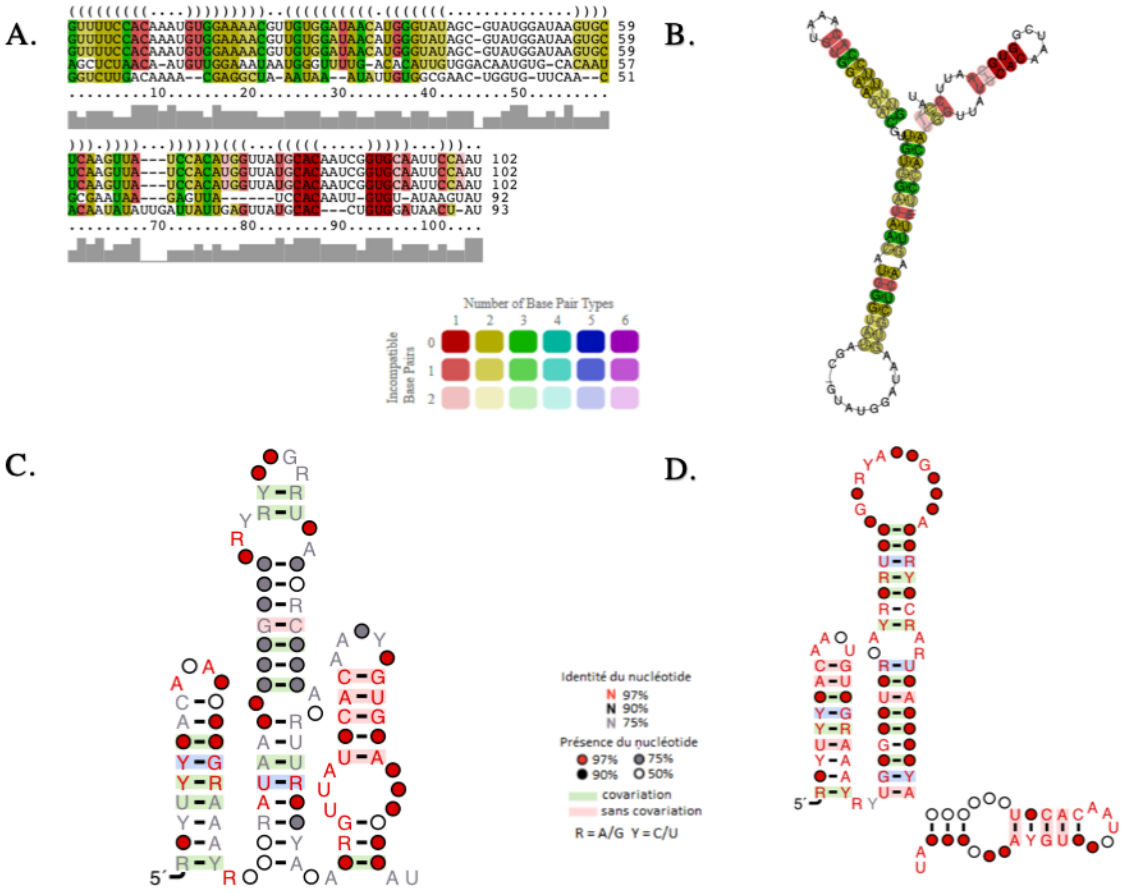

Figure 15. Motif rlmH

A) Alignement des séquences du motif prédit par GraphClust : en rouge les séquences de paires de bases conservées, les autres couleurs représentent les co-variations. B) Structure secondaire prédite par GraphClust. C) Structure secondaire dessinée avec l'outil R2R de l'alignement en [A]. D) Structure secondaire dessinée avec l'outil R2R des résultats obtenus avec l'outil Infernal.

2.2. Motifs liés à des méthylases d'ADN

2.2.1. Motifs 16 et 45

Les motifs 16 et 45 (Figure 16 et 17) ont été prédits par GraphClust dans le même UTR trouvé dans la majorité des cas entre les gènes encodant l'ADN topoisomérase III et une ADN méthyltransferase (Figure 17.E). Les résultats de recherche d'homologie d'Infernal montrent, à l'exception de quelques souches (Tableau 2), que les deux motifs se trouvent toujours ensembles dans l'UTR, incluant des alpha des gamma et des bêta-protéobactéries. Le motif 45 a été retrouvé avec 51 instances chez différentes souches, tandis que le motif 16 totalise 69 instances, pour un total de 71 instances pour les deux motifs. Ainsi, 19 fois sur 71 le motif 16 est retrouvé seul et 2 fois sur 71 le motif 45 est retrouvé seul (Tableau 3). Lorsque retrouvé seul, 16 fois sur 19 le motif 16 est retrouvé dans la région 3' d'une ADN topoisomérase sans l'ADN méthyltransferase, tandis que le motif 45 est retrouvé seul une fois sur deux dans la région 5' d'une méthyltransferase sans la topoisomérase.

BPROM n'a prédit aucun chevauchement avec un promoteur. Toutefois, cela n'exclut pas le fait que cette région puisse contenir effectivement un promoteur. Les motifs sont bien conservés et présentent beaucoup de co-variation. Cependant, étant reliés à une enzyme qui modifie l'ADN, et non l'ARN, et étant donnée la présence d'une topoisomérase en amont des motifs, le rôle que peuvent avoir ces deux motifs est ambigu. Les topoisomérases sont des enzymes impliquées dans le déroulement et le surenroulement de l'ADN notamment pendant la réplication et la transcription en ARN. À cause de sa structure en double hélice, l'ADN devient surenroulé en aval des fourches de réplication, la topoisomérase introduit des coupures dans le squelette ribose-phosphate de l'ADN pour alléger les tensions (Champoux, 2001). Ainsi, étant donné le rôle de la topoisomérase, il est possible que les motifs prédits ici soient retrouvés au niveau de l'ADN qui formerait des structures dues au « stress » du surenroulement. En effet, bien que ces deux motifs présentent des structures et un profil de co-variations très intéressant, le contexte génétique rend complexe l'attribution d'un rôle quelconque.

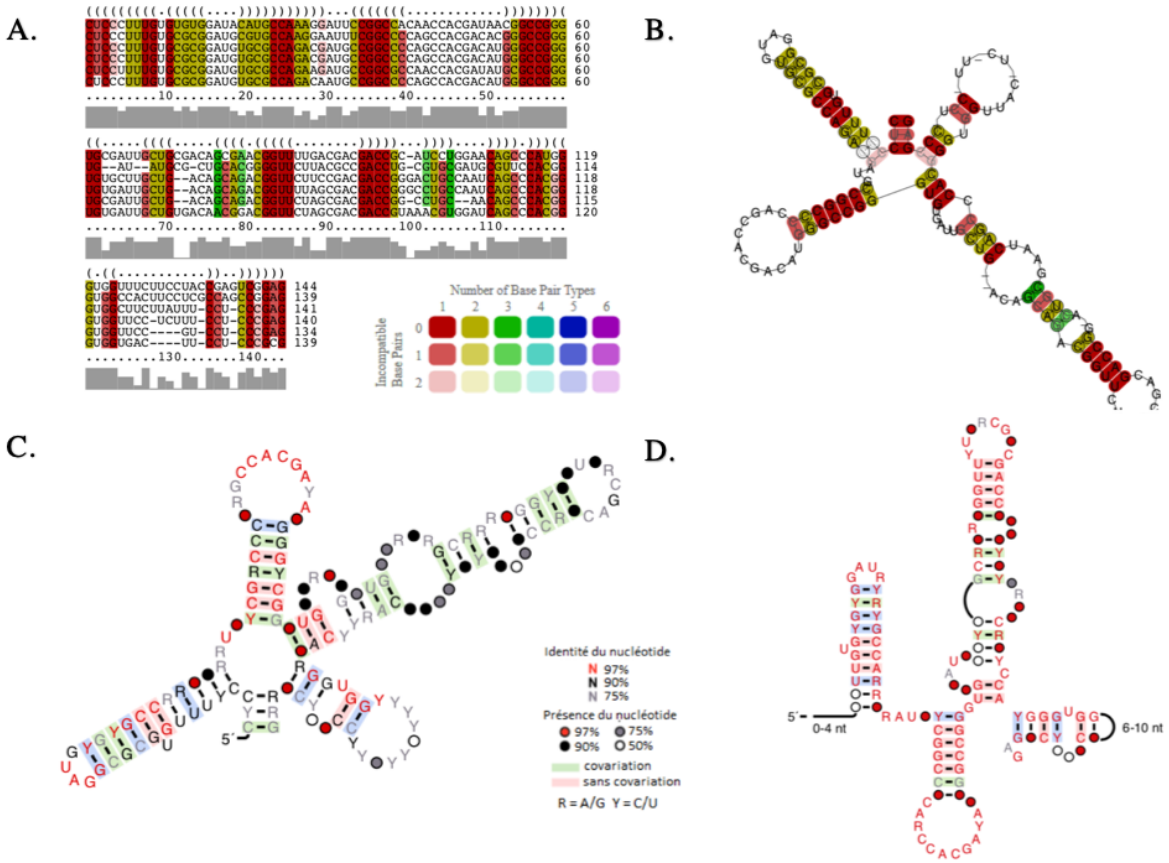

Figure 16. Motif 16

A) Alignement des séquences du motif prédit par GraphClust : en rouge les séquences de paires de bases conservées, les autres couleurs représentent les co-variations. B) Structure secondaire prédite par GraphClust. C) Structure secondaire dessinée avec l'outil R2R de l'alignement en [A]. D) Structure secondaire dessinée avec l'outil R2R des résultats obtenus avec l'outil Infernal.

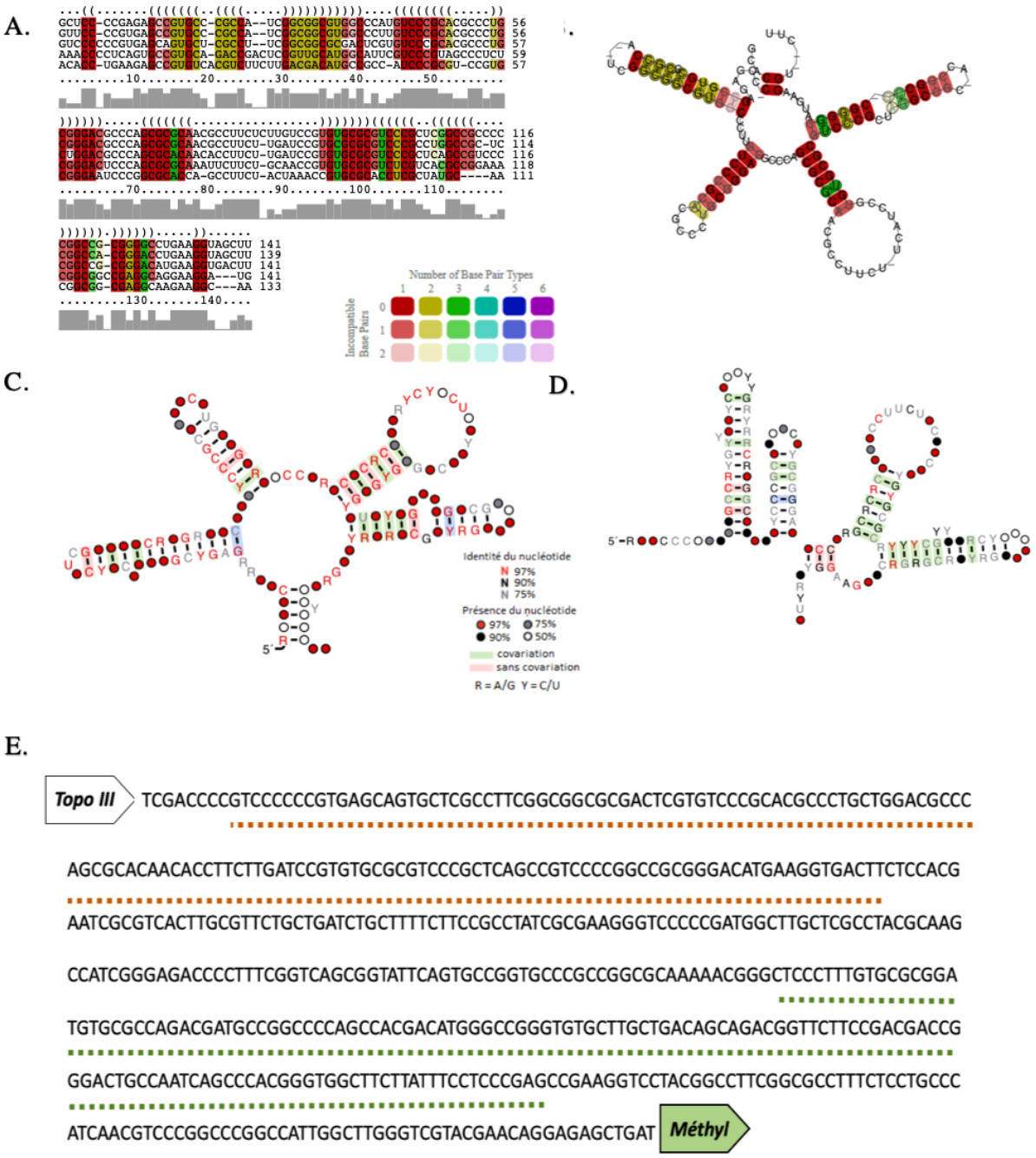

Figure 17. Motif 45

A) Alignement des séquences du motif prédit par GraphClust : en rouge les séquences de paires de bases conservées, les autres couleurs représentent les co-variations. B) Structure secondaire prédite par GraphClust. C) Structure secondaire dessiné avec l'outil R2R de l'alignement en [A]. D) Structure secondaire dessinée avec l'outil R2R des résultats obtenus avec l'outil Infernal. E) Région 5'UTR des gènes topoisomerase III et ADN méhyltransférase. En Orange la région correspondant au motif 45 et en Vert au motif 16 *Xanthomonas campestris pv. vesicatoria* str. 85-10.

Table 2. Liste des souches contenant le motif 16 sans le motif 45 dans leur génome ainsi que les gènes présents en 3' et 5' du motif

RÉFÉRENCE	ORGANISME	GÈNE EN 5'	GÈNE EN 3'
NC_003919.1	*Xanthomonas axonopodis pv. citri str.* 306	DNA methyltransferase	DNA topoisomerase III
NC_007492.2	*Pseudomonas fluorescens* Pf0-1	DUF3577 domain-containing protein	DNA topoisomerase III
NC_008752.1	*Acidovorax citrulli* AAC00-1	DNA methyltransferase	DNA topoisomerase III
NC_009512.1	*Pseudomonas putida* F1	DUF3577 domain-containing protein	DNA topoisomerase III
NC_010002.1	*Delftia acidovorans* SPH-1	Hypothetical protein	DNA topoisomerase III
NC_010501.1	*Pseudomonas putida* W619	DUF3577 domain-containing protein	DNA topoisomerase III
NC_012660.1	*Pseudomonas fluorescens* SBW25	DUF3577 domain-containing protein	DNA topoisomerase III
NC_015379.1	*Pseudomonas brassicacearum subsp.* brassicacearum NFM421	Hypothetical protein	DNA topoisomerase III
NC_015723.1	*Cupriavidus necator* N-1 chromosome 2	Pseudo integrase	AraC family transcriptional regulator
NC_015733.1	*Pseudomonas putida* S16	DUF3577 domain-containing protein	XRE family transcriptional regulator
NC_016830.1	*Pseudomonas fluorescens* F113	DUF3577 domain-containing protein	DNA topoisomerase III
NC_017986.1	*Pseudomonas putida* ND6	DUF3577 domain-containing protein	DNA topoisomerase III
NC_018220.1	*Pseudomonas putida* DOT-T1E chromosome	DUF3577 domain-containing protein	DNA topoisomerase III
NC_020800.1	*Xanthomonas axonopodis* Xac29-1	DUF3577 domain-containing protein	DNA topoisomerase III
NC_020815.1	*Xanthomona citri subsp. Citri* Aw12879	Hypothetical protein	DNA topoisomerase III
NC_021173.1	*Burkholderia thailandensis* MSMB121 chromosome 1	Hypothetical protein	DNA topoisomerase III
NC_021491.1	*Pseudomonas putida* H8234	DUF3577 domain-containing protein	DNA topoisomerase III
NC_023075.1	*Pseudomonas monteilii* SB3078	DUF3577 domain-containing protein	DNA topoisomerase III
NC_023076.1	*Pseudomonas monteilii* SB3101	DUF3577 domain-containing protein	DNA topoisomerase III

Table 3. Liste des souches contenant le motif 45 sans le motif 16 dans leur génome ainsi que les gènes présents en 3' et 5' du motif

RÉFÉRENCE	ORGANISME	GÈNE EN 5'	GÈNE EN 3'
NC_003155.4	*Streptomyces avermitilis* MA-4680 = NBRC 14893	class I SAM-dependent methyltransferase	Gluconokinase
NC_016803.1	*Desulfovibrio desulfuricans* ND132	cobalamin biosynthesis protein CbiG	glutamate-1-semialdehyde-2,1-aminomutase

Discussion

1. Nécessité d'innover dans la recherche de structures d'ARN fonctionnels

Dans notre premier article publié dans *Method*, nous avons proposé une nouvelle approche qui permet de contrer certaines difficultés associées à certaines méthodes existantes (présentées dans la section 7.1 de l'introduction). En effet, la conception de la base de données RiboGap est née d'un besoin d'accéder plus facilement aux séquences intergéniques pour l'étude des ARNnc régulateurs. Ces derniers se localisant pour la plupart dans les UTR chez les bactéries (Waters & Storz, 2009). Ainsi, en mettant à disposition une interface très facile à utiliser, l'accès aux séquences inter-géniques ainsi qu'aux éléments qui y sont annotés a été grandement simplifié. De plus, l'extraction de séquences liées à une fonction particulière est d'autant plus simple grâce à l'utilisation de mots clés. Cependant, une telle recherche est basée essentiellement sur l'annotation des gènes ce qui représente une limite majeure. En effet, de mauvaises annotations peuvent induire à de fausses conclusions. Aussi, l'utilisation de mots clés ne mène pas toujours à la fonction en question, tel que discuté dans la section résultats du chapitre 2, mais peut mener aux substrats ou aux produits des gènes, aboutissant ainsi à l'extraction de séquences en amont de gènes n'étant pas directement impliqués dans la fonction d'intérêt, le cas du motif mnmG (Figure 13) discuté dans la section résultats du chapitre 2. Par conséquent, une classification des résultats obtenus est nécessaire afin de pouvoir les interpréter et d'associer un rôle potentiel aux motifs prédits.

Nous avons intégré la base de données RiboGap dans un pipeline d'analyse avec l'outil GraphClust. L'avantage de l'outil GraphClust est qu'il n'est pas limité par le nombre de séquences à traiter ni par la taille de celles-ci, et parce qu'il est basé sur une approche d'analyse indépendante de l'alignement, discuté dans la section 7.3.2 de l'introduction, le temps de calcul est grandement réduit. Ainsi, des milliers de séquences peuvent être analysées en quelques heures seulement. Cependant, une telle approche peut causer la production d'alignement « forcés ». En effet, dans plusieurs des résultats que nous avons obtenus, nous avons remarqué que les séquences dans les alignements des *clusters* s'alignent très peu voire pas du tout, produisant de fausses co-variations donnant l'impression d'un candidat intéressant. Pour cette raison, il est important de vérifier l'alignement ainsi que la présence d'une conservation de séquences indicatrice d'une relation phylogénique entre celles-ci. De nombreux résultats obtenus avec GraphClust ont ainsi pu être éliminés, mais certains sont parfois plus ambigüs, comme le motif rlmH où la majorité des co-

variations seraient dues à deux séquences qui ne devraient probablement pas faire partie de l'alignement, malgré qu'une tige semble être conservée. D'autre part, la présence de co-variations est un des critères majeurs dans la sélection des motifs, mais basé sur ce critère, les ARNnc régulateurs les plus évidents ont déjà été trouvé. D'ailleurs, tel que discuté au chapitre 2, l'utilisation de GraphClust avec les paramètres par défaut a mené, dans la majorité des cas, à la prédiction de motifs déjà connus. Ainsi, pour chercher des motifs ARN moins « évidents », il est nécessaire d'innover dans les approches utilisées ainsi que les critères de sélection, par exemple: s'intéresser à la présence de tiges accessoires qui ne nuisent pas au maintien des structures dans les motifs d'ARN prédits (Figure 14, motif lasT).

2. Optimisation

Les premiers résultats que nous avons obtenus avec le pipeline ont fait l'objet de tests en laboratoire, notamment le motif 28 (ou mnmC) discuté dans les chapitres 1 et dans l'annexe 1. Ce dernier se présente dans une structure en trois tiges avec des boucles internes et externes (Figure 5 annexe 1). Selon les critères initialement établis (voir annexe 1), le motif mnmC représentait un très bon candidat à tester au laboratoire. Cependant, les tests *in vivo* (Figure 12) ont conclu à des résultats à l'inverse de ceux attendus. En effet, la présence du motif mnmC promeut l'expression du gène rapporteur, résultat supporté par les mutants du motif pour lesquels l'expression du gène rapporteur est significativement réduite. Ainsi, nous avons conclu que le motif serait une région promotrice contenant des activateurs transcriptionels, plutôt qu'une structure d'un ARN fonctionnel. Ces résultats nous ont mené à réévaluer le pipeline d'analyse, et à intégrer un outil de prédiction de promoteur, à savoir BPROM, pour corriger cette lacune. Le pipeline pourrait être reproposé tel que présenté à la Figure 18.

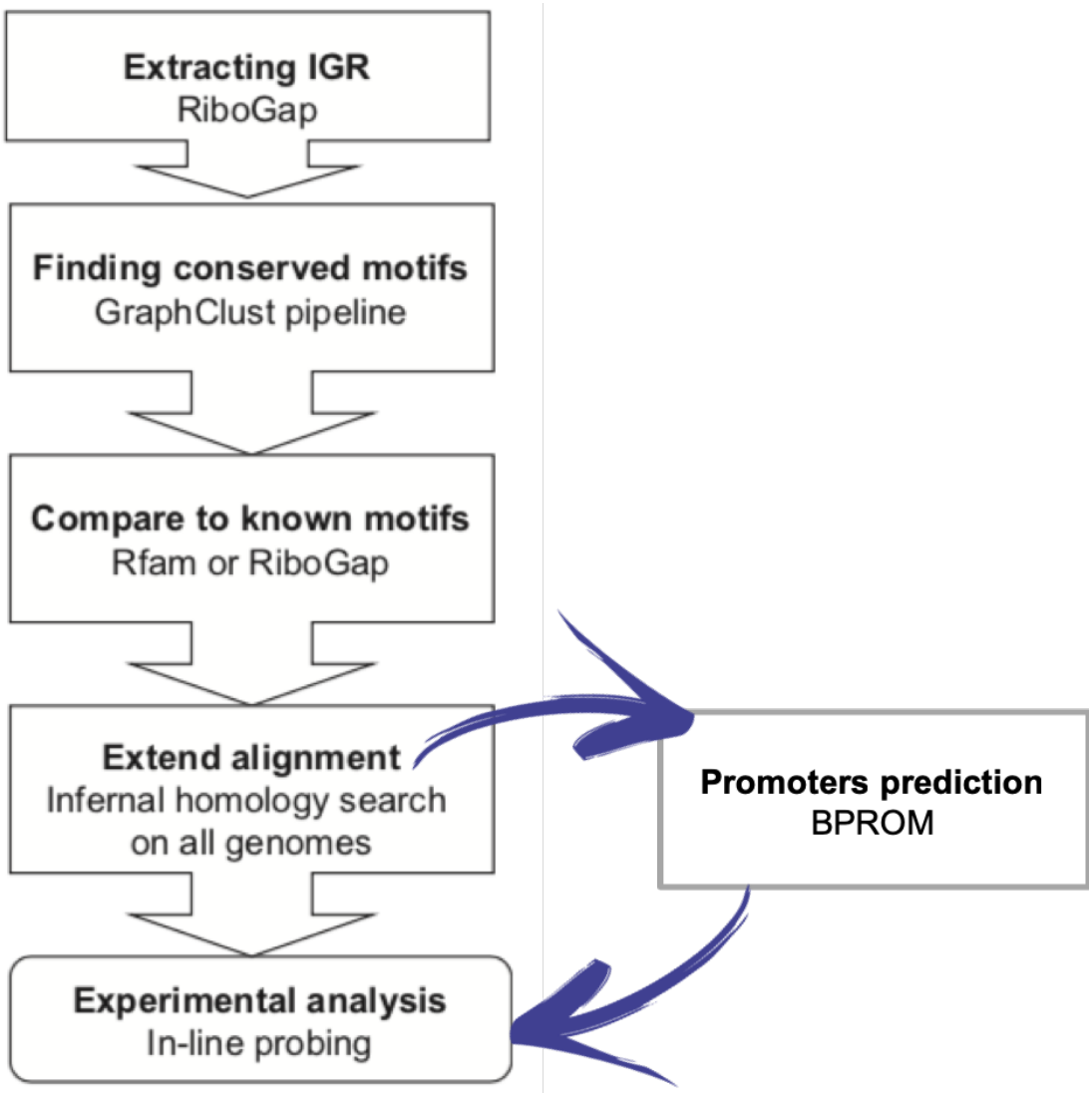

Figure 18. Représentation schématique du pipeline de découverte *de novo* d'ARNnc optimisé.

Adapté de (Naghdi *et al*., 2017).

3. Potentiels nouveaux motifs

En nous intéressant aux méthylases d'ARN avec l'hypothèse que celles-ci pourraient s'autoréguler via leurs régions intergéniques, nous avions pour intérêts d'étudier un des premiers exemples d'un ARNnc dont la régulation pourrait se faire via des enzymes modifiant l'ARN, ce qui constituerait un exemple d'étude des modifications post-transcriptionelles de façon plus large. Ainsi, bien que les travaux présentés ici n'ont pas permis de répondre à cette hypothèse, ils nous ont permis de sélectionner plusieurs candidats de nouveaux motifs intéressants avec une potentielle implication régulatrice. En effet, le motif mnmG en amont du gène *mnmG* se trouve dans la région de l'origine de réplication du chromosome (Figure 13E) et chevaucherait un promoteur selon les prédictions de BPROM (Figure 13F). Le contexte génétique du motif ainsi que la présence de plusieurs unités de transcription, pourraient indiquer une implication de la régulation de la transcription du gène *mnmG* malgré les résultats de BPROM, tel discuté au chapitre 2. Un mécanisme potentiel du motif est la régulation par la protéine AsnC via sa liaison directe au motif, tel discuté au chapitre 2. Parallèlement, les motifs lasT et rlmH pourraient potentiellement répondre à l'hypothèse initiale. En effet, les deux motifs se trouvent en amont de gènes codant des méthylases d'ARN et sont plutôt bien conservés avec beaucoup de co-variations. Finalement, les motifs 16 et 45 sont les motifs pouvant engendrer le plus de spéculation sur leurs rôles potentiels, étant donné que les gènes adjacents ont des fonctions associées avec l'ADN plutôt qu'avec l'ARN. Par exemple, tel que présenté dans la section résultats du chapitre 2, le motif 16 est retrouvé seul dans la région 3' d'une topoisomérase, sans l'ADN méthyltransferase, chez 19 souches différentes suggérant un lien potentiel avec la topoisomérase plutôt qu'avec la méthyltransferase. En effet, il pourrait apparaître logique de contrôler l'expression de la topoisomérase via des structures qui se formeraient transitoirement dans l'ADN selon le niveau de surenroulement de celui-ci. D'ailleurs, des études transcriptomiques chez *Pseudomonas aeruginosa* montrent une activité transcriptionnelle importante dans les régions 3′ de gènes de topoisomérase (Wurtzel *et al*., 2012). À l'inverse, certaines topoisomérase ont été montrées comme ayant également un effet sur l'ARN (DiGate & Marians, 1992), ce qui ouvrirait la porte à une possible boucle de rétro-action d'expression de la topoisomérase soit via une action sur l'ADN, soit sur l'ARN.

Conclusion

Les travaux présentés ici ont permis de tester une nouvelle méthode basée sur la fonction pour la recherche *de novo* d'ARNnc régulateurs. L'utilisation de la base de données RiboGap a permis un accès simplifié aux séquences intergéniques et des milliers de séquences ont pu être compilées. L'utilisation de la suite d'outil GraphClust a permis des analyses plus rapides d'un large ensemble de données sans être limité par la taille ni le nombre de séquences. Nous avons caractérisé des éléments régulateurs dans la région 5′UTR du gène *mnmC*. Par ailleurs, ce résultat nous a permis d'optimiser la méthode de recherche d'ARNnc grâce à l'outil BPROM qui permet de faire la prédiction des promoteurs et de prévoir ainsi un chevauchement avec les potentiels motifs prédits. Finalement, nous avons sélectionné une liste de motifs à tester au laboratoire comme motifs ARN régulateurs potentiels.

RÉFÉRENCES

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology* 215(3):403-410.

Anantharaman V, Koonin EV & Aravind L (2002) SPOUT: a class of methyltransferases that includes spoU and trmD RNA methylase superfamilies, and novel superfamilies of predicted prokaryotic RNA methylases. *J Mol Microbiol Biotechnol* 4(1):71-75.

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL & Mori H (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2:2006 0008.

Backofen R, Amman F, Costa F, Findeiss S, Richter AS & Stadler PF (2014a) Bioinformatics of prokaryotic RNAs. *RNA Biol* 11(5):470-483.

Backofen R, Amman F, Costa F, Findeiss S, Richter AS & Stadler PF (2014b) Bioinformatics of prokaryotic RNAs. *RNA biology* 11(5).

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW & Noble WS (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* 37(Web Server issue):W202-208.

Bailey TL & Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2:28-36.

Bailey TL, Johnson J, Grant CE & Noble WS (2015) The MEME Suite. *Nucleic Acids Res* 43(W1):W39-49.

Barrick JE & Breaker RR (2007) The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome Biol* 8(11):R239.

Bastet L, Turcotte P, Wade JT & Lafontaine DA (2018) Maestro of regulation: Riboswitches orchestrate gene expression at the levels of translation, transcription and mRNA decay. *RNA Biol* 15(6):679-682.

Bates DB, Boye E, Asai T & Kogoma T (1997) The absence of effect of gid or mioC transcription on the initiation of chromosomal replication in Escherichia coli. *Proc Natl Acad Sci U S A* 94(23):12497-12502.

Benitez-Paez A, Villarroya M & Armengod ME (2012) Regulation of expression and catalytic activity of Escherichia coli RsmG methyltransferase. *Rna* 18(4):795-806.

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J & Sayers EW (2015) GenBank. *Nucleic acids research* 43(Database issue):D30-35.

Bernhart SH, Hofacker IL, Will S, Gruber AR & Stadler PF (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC bioinformatics* 9:474.

Boccaletto P, Machnicka MA, Purta E, Piatkowski P, Baginski B, Wirecki TK, de Crecy-Lagard V, Ross R, Limbach PA, Kotter A, Helm M & Bujnicki JM (2018) MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res* 46(D1):D303-D307.

Breaker RR (2011a) Prospects for Riboswitch Discovery and Analysis. *Mol Cell* 43(6):867-879.

Breaker RR (2011b) Prospects for riboswitch discovery and analysis. *Molecular cell* 43(6):867-879.

Breaker RR (2018) Riboswitches and Translation Control. *Cold Spring Harb Perspect Biol* 10.1101/cshperspect.a032797.

Breaker RR, Atilho RM, Malkowski SN, Nelson JW & Sherlock ME (2017) The Biology of Free Guanidine As Revealed by Riboswitches. *Biochemistry* 10.1021/acs.biochem.6b01269.

Bujnicki JM, Oudjama Y, Roovers M, Owczarek S, Caillet J & Droogmans L (2004) Identification of a bifunctional enzyme MnmC involved in the biosynthesis of a hypermodified uridine in the wobble position of tRNA. *Rna* 10(8):1236-1242.
Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K & Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.
Cases I & de Lorenzo V (2005) Promoters in the environment: Transcriptional regulation in its natural context. *Nat Rev Microbiol* 3(2):105-118.
Cases I, de Lorenzo V & Ouzounis CA (2003) Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol* 11(6):248-253.
Champoux JJ (2001) DNA topoisomerases: structure, function, and mechanism. *Annu Rev Biochem* 70:369-413.
Chen H, Bjerknes M, Kumar R & Jay E (1994) Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. *Nucleic acids research* 22(23):4953-4957.
Crick FH, Barnett L, Brenner S & Watts-Tobin RJ (1961) General nature of the genetic code for proteins. *Nature* 192:1227-1232.
Crooks GE, Hon G, Chandonia JM & Brenner SE (2004) WebLogo: a sequence logo generator. *Genome research* 14(6):1188-1190.
Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Giron CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Kahari AK, Keenan S, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Overduin B, Parker A, Patricio M, Perry E, Pignatelli M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Aken BL, Birney E, Harrow J, Kinsella R, Muffato M, Ruffier M, Searle SM, Spudich G, Trevanion SJ, Yates A, Zerbino DR & Flicek P (2015) Ensembl 2015. *Nucleic acids research* 43(Database issue):D662-669.
Curry KA & Tomich CS (1988) Effect of ribosome binding site on gene expression in Escherichia coli. *DNA* 7(3):173-179.
de Lorenzo V, Giovannini F, Herrero M & Neilands JB (1988) Metal ion regulation of gene expression. Fur repressor-operator interaction at the promoter region of the aerobactin system of pColV-K30. *Journal of molecular biology* 203(4):875-884.
DiGate RJ & Marians KJ (1992) Escherichia coli topoisomerase III-catalyzed cleavage of RNA. *J Biol Chem* 267(29):20532-20535.
Dirusso CC, Heimert TL & Metzger AK (1992) Characterization of Fadr, a Global Transcriptional Regulator of Fatty-Acid Metabolism in Escherichia-Coli - Interaction with the Fadb Promoter Is Prevented by Long-Chain Fatty Acyl Coenzyme-As. *Journal of Biological Chemistry* 267(12):8685-8691.
Eggenhofer F, Hofacker IL & Siederdissen CHZ (2013) CMCompare webserver: comparing RNA families via covariance models. *Nucleic Acids Res* 41(W1):W499-W503.
El Korbi A, Ouellet J, Naghdi MR & Perreault J (2014) Finding instances of riboswitches and ribozymes by homology search of structured RNA with Infernal. *Methods Mol Biol* 1103:113-126.
Farewell A, Diez AA, DiRusso CC & Nystrom T (1996) Role of the Escherichia coli FadR regulator in stasis survival and growth phase-dependent expression of the uspA, fad, and fab genes. *Journal of Bacteriology* 178(22):6443-6450.

Farnham PJ & Platt T (1981) Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription in vitro. *Nucleic acids research* 9(3):563-577.
Feng Y & Cronan JE (2011) Complex binding of the FabR repressor of bacterial unsaturated fatty acid biosynthesis to its cognate promoters. *Mol Microbiol* 80(1):195-218.
Fox GE & Woese CR (1975) 5S RNA secondary structure. *Nature* 256(5517):505-507.
Gibson DG, Smith HO, Hutchison CA, 3rd, Venter JC & Merryman C (2010) Chemical synthesis of the mouse mitochondrial genome. *Nat Methods* 7(11):901-903.
Gielow A, Kucherer C, Kolling R & Messer W (1988) Transcription in the region of the replication origin, oriC, of Escherichia coli: termination of asnC transcripts. *Mol Gen Genet* 214(3):474-481.
Gollnick P, Babitzke P, Antson A & Yanofsky C (2005) Complexity in regulation of tryptophan biosynthesis in Bacillus subtilis. *Annual review of genetics* 39:47-68.
Gossringer M & Hartmann RK (2012) 3'-UTRs as a source of regulatory RNAs in bacteria. *The EMBO journal* 31(20):3958-3960.
Grave KD & Costa F (2010) Molecular graph augmentation with rings and functional groups. *J Chem Inf Model* 50(9):1660-1668.
Griffiths-Jones S (2005) RALEE--RNA ALignment editor in Emacs. *Bioinformatics* 21(2):257-259.
Gruber AR, Findeiss S, Washietl S, Hofacker IL & Stadler PF (2010) RNAz 2.0: improved noncoding RNA detection. *Pac Symp Biocomput* :69-79.
Hayashi K, Morooka N, Yamamoto Y, Fujita K, Isono K, Choi S, Ohtsubo E, Baba T, Wanner BL, Mori H & Horiuchi T (2006) Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. *Mol Syst Biol* 2:2006 0007.
Heyne S, Costa F, Rose D & Backofen R (2012a) GraphClust: alignment-free structural clustering of local RNA secondary structures. *Bioinformatics* 28(12):I224-I232.
Heyne S, Costa F, Rose D & Backofen R (2012b) GraphClust: alignment-free structural clustering of local RNA secondary structures. *Bioinformatics* 28(12):i224-232.
Hochsmann M, Toller T, Giegerich R & Kurtz S (2003) Local similarity in RNA secondary structures. *Proc IEEE Comput Soc Bioinform Conf* 2:159-168.
Janssen S & Giegerich R (2015) The RNA shapes studio. *Bioinformatics* 31(3):423-425.
Jean-Pierre F, Perreault J & Deziel E (2015) Complex autoregulation of the post-transcriptional regulator RsmA in Pseudomonas aeruginosa. *Microbiology* 161(9):1889-1896.
Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, Bateman A, Finn RD & Petrov AI (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res* 46(D1):D335-D342.
Kawano M, Reynolds AA, Miranda-Rios J & Storz G (2005) Detection of 5'- and 3'-UTR-derived small RNAs and cis-encoded antisense RNAs in Escherichia coli. *Nucleic Acids Res* 33(3):1040-1050.
Kim J & Almo SC (2013) Structural basis for hypermodification of the wobble uridine in tRNA by bifunctional enzyme MnmC. *BMC Struct Biol* 13:5.
Kim JN, Roth A & Breaker RR (2007) Guanine riboswitch variants from Mesoplasma florum selectively recognize 2'-deoxyguanosine. *Proceedings of the National Academy of Sciences of the United States of America* 104(41):16092-16097.

Kitamura A, Sengoku T, Nishimoto M, Yokoyama S & Bessho Y (2011) Crystal structure of the bifunctional tRNA modification enzyme MnmC from Escherichia coli. *Protein Sci* 20(7):1105-1113.

Kolling R, Gielow A, Seufert W, Kucherer C & Messer W (1988) AsnC, a multifunctional regulator of genes located around the replication origin of Escherichia coli, oriC. *Mol Gen Genet* 212(1):99-104.

Kolling R & Lother H (1985) AsnC: an autogenously regulated activator of asparagine synthetase A transcription in Escherichia coli. *J Bacteriol* 164(1):310-315.

Kozak M (1983) Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. *Microbiol Rev* 47(1):1-45.

Kurata S, Weixlbaumer A, Ohtsuki T, Shimazaki T, Wada T, Kirino Y, Takai K, Watanabe K, Ramakrishnan V & Suzuki T (2008) Modified uridines with C5-methylene substituents at the first position of the tRNA anticodon stabilize U.G wobble pairing during decoding. *J Biol Chem* 283(27):18801-18811.

Lapouge K, Schubert M, Allain FH & Haas D (2008) Gac/Rsm signal transduction pathway of gamma-proteobacteria: from RNA recognition to regulation of social behaviour. *Mol Microbiol* 67(2):241-253.

Lies M, Visser BJ, Joshi MC, Magnan D & Bates D (2015) MioC and GidA proteins promote cell division in E. coli. *Front Microbiol* 6:516.

Lodato PB, Hsieh PK, Belasco JG & Kaper JB (2012) The ribosome binding site of a mini-ORF protects a T3SS mRNA from degradation by RNase E. *Molecular microbiology* 86(5):1167-1182.

Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF & Hofacker IL (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol* 6:26.

Marbaniang CN & Vogel J (2016) Emerging roles of RNA modifications in bacteria. *Curr Opin Microbiol* 30:50-57.

Marrakchi H, Zhang YM & Rock CO (2002) Mechanistic diversity and regulation of Type II fatty acid synthesis. *Biochem Soc T* 30:1050-1055.

Masse E, Majdalani N & Gottesman S (2003) Regulatory roles for small RNAs in bacteria. *Current opinion in microbiology* 6(2):120-124.

Mathy N, Benard L, Pellegrini O, Daou R, Wen T & Condon C (2007) 5'-to-3' exoribonuclease activity in bacteria: role of RNase J1 in rRNA maturation and 5' stability of mRNA. *Cell* 129(4):681-692.

Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V, Salgado H, Juarez K, Contreras-Moreira B, Huerta AM, Collado-Vides J & Morett E (2009) Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli. *PLoS One* 4(10):e7526.

Meyer MM (2016) The role of mRNA structure in bacterial translational regulation. *Wiley interdisciplinary reviews. RNA* 10.1002/wrna.1370.

Meyer MM (2017) The role of mRNA structure in bacterial translational regulation. *Wiley Interdiscip Rev RNA* 8(1).

Meyer MM, Ames TD, Smith DP, Weinberg Z, Schwalbach MS, Giovannoni SJ & Breaker RR (2009) Identification of candidate structured RNAs in the marine organism 'Candidatus Pelagibacter ubique'. *BMC genomics* 10:268.

Mironov AS, Gusarov I, Rafikov R, Lopez LE, Shatalin K, Kreneva RA, Perumov DA & Nudler E (2002) Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* 111(5):747-756.

Mizuno T (1984) Regulation of gene expression by a small RNA transcript (micRNA). *Tanpakushitsu kakusan koso. Protein, nucleic acid, enzyme* 29(11):908-913.

Montange RK & Batey RT (2006) Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. *Nature* 441(7097):1172-1175.

Morita MT, Tanaka Y, Kodama TS, Kyogoku Y, Yanagi H & Yura T (1999) Translational induction of heat shock transcription factor sigma32: evidence for a built-in RNA thermosensor. *Genes & development* 13(6):655-665.

Motorin Y & Helm M (2011) RNA nucleotide methylation. *Wiley Interdiscip Rev RNA* 2(5):611-631.

Moukadiri I, Garzon MJ, Bjork GR & Armengod ME (2014) The output of the tRNA modification pathways controlled by the Escherichia coli MnmEG and MnmC enzymes depends on the growth conditions and the tRNA species. *Nucleic Acids Res* 42(4):2602-2623.

Moukadiri I, Prado S, Piera J, Velazquez-Campoy A, Bjork GR & Armengod ME (2009) Evolutionarily conserved proteins MnmE and GidA catalyze the formation of two methyluridine derivatives at tRNA wobble positions. *Nucleic Acids Res* 37(21):7177-7193.

My L, Ghandour Achkar N, Viala JP & Bouveret E (2015) Reassessment of the Genetic Regulation of Fatty Acid Synthesis in Escherichia coli: Global Positive Control by the Dual Functional Regulator FadR. *J Bacteriol* 197(11):1862-1872.

Nachtergaele S & He C (2017) The emerging biology of RNA post-transcriptional modifications. *RNA Biol* 14(2):156-163.

Naghdi MR, Smail K, Wang JX, Wade F, Breaker RR & Perreault J (2017) Search for 5'-leader regulatory RNA structures based on gene annotation aided by the RiboGap database. *Methods* 117:3-13.

Nahvi A, Sudarsan N, Ebert MS, Zou X, Brown KL & Breaker RR (2002) Genetic control by a metabolite binding mRNA. *Chem Biol* 9(9):1043.

Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J & Finn RD (2015) Rfam 12.0: updates to the RNA families database. *Nucleic acids research* 43(Database issue):D130-137.

Nawrocki EP & Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29(22):2933-2935.

Nawrocki EP, Kolbe DL & Eddy SR (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics* 25(10):1335-1337.

Ogawa T & Okazaki T (1991) Concurrent transcription from the gid and mioC promoters activates replication of an Escherichia coli minichromosome. *Mol Gen Genet* 230(1-2):193-200.

Okamoto S, Tamaru A, Nakajima C, Nishimura K, Tanaka Y, Tokuyama S, Suzuki Y & Ochi K (2007) Loss of a conserved 7-methylguanosine modification in 16S rRNA confers low-level streptomycin resistance in bacteria. *Mol Microbiol* 63(4):1096-1106.

Okuda S & Yoshizawa AC (2011) ODB: a database for operon organizations, 2011 update. *Nucleic acids research* 39(Database issue):D552-555.

Ovcharenko A & Rentmeister A (2018) Emerging approaches for detection of methylation sites in RNA. *Open Biol* 8(9).
Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W & Haussler D (2006) Identification and classification of conserved RNA secondary structures in the human genome. *Plos Comput Biol* 2(4):e33.
Poggio S, Domeinzain C, Osorio A & Camarena L (2002) The nitrogen assimilation control (Nac) protein represses asnC and asnA transcription in Escherichia coli. *FEMS Microbiol Lett* 206(2):151-156.
Purta E, Kaminska KH, Kasprzak JM, Bujnicki JM & Douthwaite S (2008) YbeA is the m3Psi methyltransferase RlmH that targets nucleotide 1915 in 23S rRNA. *Rna* 14(10):2234-2244.
Purta E, van Vliet F, Tkaczuk KL, Dunin-Horkawicz S, Mori H, Droogmans L & Bujnicki JM (2006) The yfhQ gene of Escherichia coli encodes a tRNA:Cm32/Um32 methyltransferase. *BMC Mol Biol* 7:23.
Qiao H, Lu N, Du E, Yao L, Xiao H, Lu S & Qi Y (2011) Rare codons in uORFs of baculovirus p13 gene modulates downstream gene expression. *Virus research* 155(1):249-253.
Ravnum S & Andersson DI (2001) An adenosyl-cobalamin (coenzyme-B12)-repressed translational enhancer in the cob mRNA of Salmonella typhimurium. *Mol Microbiol* 39(6):1585-1594.
Regulski EE & Breaker RR (2008) In-line probing analysis of riboswitches. *Methods Mol Biol* 419:53-67.
Regulski EE, Moy RH, Weinberg Z, Barrick JE, Yao Z, Ruzzo WL & Breaker RR (2008) A widespread riboswitch candidate that controls bacterial genes involved in molybdenum cofactor and tungsten cofactor metabolism. *Molecular microbiology* 68(4):918-932.
Repoila F & Darfeuille F (2009) Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects. *Biol Cell* 101(2):117-131.
Romeo T, Vakulskas CA & Babitzke P (2013) Post-transcriptional regulation on a global scale: form and function of Csr/Rsm systems. *Environmental microbiology* 15(2):313-324.
Roovers M, Oudjama Y, Kaminska KH, Purta E, Caillet J, Droogmans L & Bujnicki JM (2008) Sequence-structure-function analysis of the bifunctional enzyme MnmC that catalyses the last two steps in the biosynthesis of hypermodified nucleoside mnm5s2U in tRNA. *Proteins* 71(4):2076-2085.
Ruzzo WL & Gorodkin J (2014) De novo discovery of structured ncRNA motifs in genomic sequences. *Methods Mol Biol* 1097:303-318.
Salgado H, Martinez-Flores I, Lopez-Fuentes A, Garcia-Sotelo JS, Porron-Sotelo L, Solano H, Muniz-Rascado L & Collado-Vides J (2012) Extracting regulatory networks of Escherichia coli from RegulonDB. *Methods Mol Biol* 804:179-195.
Seetin MG & Mathews DH (2012) RNA structure prediction: an overview of methods. *Methods Mol Biol* 905:99-122.
Serganov A & Nudler E (2013) A Decade of Riboswitches. *Cell* 152(1-2):17-24.
Sergeeva OV, Bogdanov AA & Sergiev PV (2015) What do we know about ribosomal RNA methylation in Escherichia coli? *Biochimie* 117:110-118.
Sherlock ME & Breaker RR (2017) Biochemical Validation of a Third Guanidine Riboswitch Class in Bacteria. *Biochemistry* 10.1021/acs.biochem.6b01271.
Sherlock ME, Malkowski SN & Breaker RR (2017) Biochemical Validation of a Second Guanidine Riboswitch Class in Bacteria. *Biochemistry* 10.1021/acs.biochem.6b01270.

Siebert S & Backofen R (2005) MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics* 21(16):3352-3359.
Simons RW, Houman F & Kleckner N (1987) Improved single and multicopy lac-based cloning vectors for protein and operon fusions. *Gene* 53(1):85-96.
Solovyev V, Salamov A, Seledtsov I, Vorobyev D & Bachinsky A (2011) Automatic Annotation of Bacterial Community Sequences and Application to Infections Diagnostic. *Bioinformatics 2011* :346-+.
Tatusova T, Ciufo S, Fedorov B, O'Neill K & Tolstoy I (2015) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic acids research* 43(7):3872.
Tian S, Yesselman JD, Cordero P & Das R (2015) Primerize: automated primer assembly for transcribing non-coding RNA domains. *Nucleic acids research* 43(W1):W522-526.
Valverde C, Lindell M, Wagner EG & Haas D (2004) A repeated GGA motif is critical for the activity and stability of the riboregulator RsmY of Pseudomonas fluorescens. *J Biol Chem* 279(24):25066-25074.
Wang JX & Breaker RR (2008) Riboswitches that sense S-adenosylmethionine and S-adenosylhomocysteine. *Biochem Cell Biol* 86(2):157-168.
Wang X & He C (2014) Dynamic RNA modifications in posttranscriptional regulation. *Mol Cell* 56(1):5-12.
Waters LS & Storz G (2009) Regulatory RNAs in bacteria. *Cell* 136(4):615-628.
Weinberg Z, Barrick JE, Yao Z, Roth A, Kim JN, Gore J, Wang JX, Lee ER, Block KF, Sudarsan N, Neph S, Tompa M, Ruzzo WL & Breaker RR (2007) Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res* 35(14):4809-4819.
Weinberg Z & Breaker RR (2011) R2R--software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC bioinformatics* 12:3.
Weinberg Z, Wang JX, Bogue J, Yang JY, Corbino K, Moy RH & Breaker RR (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol* 11(3).
Will S, Joshi T, Hofacker IL, Stadler PF & Backofen R (2012) LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs. *Rna* 18(5):900-914.
Will S, Reiche K, Hofacker IL, Stadler PF & Backofen R (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS computational biology* 3(4):e65.
Wilson KS & von Hippel PH (1995) Transcription termination at intrinsic terminators: the role of the RNA hairpin. *Proceedings of the National Academy of Sciences of the United States of America* 92(19):8793-8797.
Winkler W, Nahvi A & Breaker RR (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* 419(6910):952-956.
Winkler WC, Nahvi A, Sudarsan N, Barrick JE & Breaker RR (2003) An mRNA structure that controls gene expression by binding S-adenosylmethionine. *Nature structural biology* 10(9):701-707.
Wurtzel O, Yoder-Himes DR, Han K, Dandekar AA, Edelheit S, Greenberg EP, Sorek R & Lory S (2012) The single-nucleotide resolution transcriptome of Pseudomonas aeruginosa grown in body temperature. *PLoS Pathog* 8(9):e1002945.
Yao Z, Weinberg Z & Ruzzo WL (2006a) CMfinder--a covariance model based RNA motif finding algorithm. *Bioinformatics* 22(4):445-452.

Yao ZZ, Weinberg Z & Ruzzo WL (2006b) CMfinder - a covariance model based RNA motif finding algorithm. *Bioinformatics* 22(4):445-452.

Zhang YM, Marrakchi H & Rock CO (2002) The FabR (YijC) transcription factor regulates unsaturated fatty acid biosynthesis in Escherichia coli. *Journal of Biological Chemistry* 277(18):15558-15565.

Zhu K, Zhang YM & Rock CO (2009) Transcriptional Regulation of Membrane Lipid Homeostasis in Escherichia coli. *Journal of Biological Chemistry* 284(50):34880-34888.

ANNEXE

Annexe 1 : Recherche de structures régulatrices d'ARN dans les régions 5′ basé sur l'annotation des gènes en utilisant la base de données RiboGap

Search for 5′-leader regulatory RNA structures based on gene annotation aided by the RiboGap database

Mohammad Reza Naghdi[1], Katia Smail[1], Joy X. Wang[2], Fallou Wade[1], Ronald R. Breaker[2] and Jonathan Perreault[1]

[1] : INRS - Institut Armand-Frappier, 531 boul des Prairies, Laval (Québec), H7V 1B7, Canada.

[2] : Department of Molecular, Cellular and Developmental Biology and the Howard Hughes Medical Institute, Yale University, P.O. Box 208103, New Haven, CT 06520-8103.

Correspondance : Jonathan Perreault, adresse courriel : jonathan.perreault@iaf.inrs.ca
INRS-Institut Armand-Frappier, 531 boulevard des Prairies, Laval (Québec), H7V 1B7, Canada.
Tel : 1-450-687-5010 (ext 4411), Fax : 1-450-686-5301.

Contribution des auteurs

Mohammad Reza Naghdi a conçu et implémenté la base de données RiboGap. Il a aussi intégré RiboGap dans le cadre d'un *pipeline* de découverte d'ARNnc avec la suite GraphClust. Il a utilisé cette approche pour trouver un ARNnc qu'il a utilisé comme exemple pour expliquer les approches expérimentales. Il a rédigé la plus grande partie de l'article. Katia Smail a utilisé le *pipeline,* pour la découverte d'ARNnc en lien avec la méthylation, et a contribué à son optimisation ; elle a généré les résultats relatifs au motif *methyl-28* et rédigé une partie de l'article. Fallou Wade a aidé à mettre au point le système de notification et les pages d'aide de la base de données. Joy X. Wang a contribué à certaines des idées de départ pour la base de données RiboGap, de même qu'à la préparation d'une version préliminaire, elle était pour cela en partie guidée par Ronald R. Breaker. Ronald Breaker a aussi contribué à la rédaction du manuscrit. Jonathan Perreault a participé à plusieurs idées concernant la base de données, de même que son utilisation pour plusieurs types de questions et a participé à la rédaction de l'article.

Résumé

L'étude d'ARN noncodants (ARNnc) et leur importance pour la régulation des gènes nous ont amenés à développer des outils bio-informatiques permettant de poursuivre la découverte de nouveaux ARNnc. La découverte *de novo* d'ARNnc présente plusieurs défis, d'abord en raison de la difficulté de récupérer un grand nombre de séquences pour des activités géniques données, et en second lieu en raison des exigences de calcul exponentielles requises pour la génomique comparative à grande échelle. Récemment, plusieurs outils de prévision de la structure secondaire des ARN conservés ont été développés, mais nombre d'entre eux ne sont pas conçus pour découvrir de nouveaux ARNnc ou sont trop lents pour permettre des analyses à grande échelle. Nous présentons ici différentes approches utilisant la base de données RiboGap pour trouver des ARNnc connus et pour découvrir des motifs de séquence simples dotés de rôles de régulation. Cette base de données peut également être utilisée pour extraire facilement des séquences intergéniques de bactéries et d'archées afin de trouver des structures d'ARN conservées en amont de gènes donnés. Nous montrons également comment étendre l'analyse afin de choisir les meilleurs ARNnc candidats pour la validation expérimentale.

Abstract

The discovery of noncoding RNAs (ncRNAs) and their importance for gene regulation led us to develop bioinformatics tools to pursue the discovery of novel ncRNAs. Finding ncRNAs de novo is challenging, first due to the difficulty of retrieving large numbers of sequences for given gene activities, and second due to exponential demands on calculation needed for comparative genomics on a large scale. Recently, several tools for the prediction of conserved RNA secondary structure were developed, but many of them are not designed to uncover new ncRNAs, or are too slow for conducting analyses on a large scale. Here we present various approaches using the database RiboGap as a primary tool for finding known ncRNAs and for uncovering simple sequence motifs with regulatory roles. This database also can be used to easily extract intergenic sequences of eubacteria and archaea to find conserved RNA structures upstream of given genes. We also show how to extend analysis further to choose the best candidate ncRNAs for experimental validation.

Keywords: ncRNA; bioinformatics; GraphClust; Infernal; Rfam; RNA secondary structure; riboswitch; methyltransferase; TRAP; CsrA; RsmA

Introduction

In the past decade, the availability of a wealth of sequence information was successfully exploited with bioinformatics tools for the discovery of noncoding RNAs (ncRNAs) in bacteria (Backofen *et al*., 2014b). Because bacterial genomes are dense in information, using comparative genomics to examine intergenic regions (IGRs) has yielded numerous new functional RNA structures. IGRs are sequences that bridge gaps between protein-coding sequences or open reading frames (ORFs). Sequence elements that regulate bacterial gene expression are mostly found in IGRs (Breaker, 2011b; Gossringer & Hartmann, 2012), regardless of whether they are cis or trans-acting (Serganov & Nudler, 2013). Among the most common ncRNAs known in bacteria are the so-called small RNAs (sRNAs) which are independently transcribed and act in trans by binding mRNA targets, typically to repress expression (Masse *et al*., 2003; Mizuno, 1984). There are also many types of cis-acting elements such as ribosomal protein leaders (Meyer, 2016), thermoregulators (Morita *et al*., 1999) and riboswitches (Nahvi *et al*., 2002; Winkler *et al*., 2003). One way to discover de novo ncRNAs is to look for RNA structures in the IGR upstream of genes with similar activity. While most IGR sequences are poorly conserved compared to coding sequence, conserved sequence and structure can be observed in IGRs where a cis-regulatory RNA element is present. Riboswitches are especially well-suited for such searches because their ability to specifically recognize metabolites to exert gene regulation requires a highly conserved structure. To embark on a campaign to discover novel ncRNAs by using comparative sequence and structure analysis, it is helpful to pool IGRs associated with genes of related functions. However, finding and extracting all the intergenic sequences for a specific function from general databases like GenBank (Benson *et al*., 2015) or Ensembl (Cunningham *et al*., 2015) can be laborious and requires programming skills. There is no simple way to extract intergenic sequences without any prior knowledge of gene positions. Obtaining IGRs associated with all the genes with similar or related annotated activities is even more difficult. Moreover, for large numbers of representatives (more than 500 sequences) and variable length IGRs, prediction of secondary structure is time-consuming and challenging for most available software packages (Seetin & Mathews, 2012).

Since bacterial IGR sequences have important regulatory elements such as promoters, noncoding RNAs and terminators, a tool to easily fetch known information about IGRs would be useful. Here we describe how this can be done with RiboGap by simply choosing the fields you wish to display and the keywords or parameters corresponding to the query. This provides an easy means to look

for known RNA structures or sequence motifs associated with genes that are grouped by different annotated functions, taxonomic ensembles or even phenotypic traits, such as concerted regulation under different growth conditions. Moreover, fetching the IGR sequences with RiboGap also facilitate searches for ncRNA elements, notably by comparative genomics.

In this report, we describe such a comparative genomic pipeline which uses RiboGap along with GraphClust (Heyne *et al.*, 2012b), a software package for secondary structure alignment prediction that can manage large numbers of sequences. Finally, we describe a method to validate putative ncRNAs with an experimental approach for riboswitch validation, in-line probing (Regulski *et al.*, 2008). The proposed methodology also relies on other software packages to improve analysis, mainly Infernal, for homology searches, as well as R2R and Emacs (with the Ralee plug-in) for RNA structure and alignment visualization, respectively.

Materials and methods

1. Bioinformatics

a. RiboGap

RiboGap is a database (http://ribogap.iaf.inrs.ca) which helps find IGRs from the 5′-UTRs as well as from the 3′-UTR for one or several genes without any previous knowledge of gene positions. RiboGap also provides information on the presence of known noncoding RNAs and Rho-independent transcription terminators (Farnham & Platt, 1981; Wilson & von Hippel, 1995) in IGRs. Moreover, RiboGap uses operon data from the Operon DataBase: ODB (Okuda & Yoshizawa, 2011) including known and conserved bacterial operons (Tatusova *et al.*, 2015), as well as operon predictions based on adjacent genes with the same orientation. RiboGap can carry out keyword searches by using pattern matching. It also offers the possibility to use genome coordinates from hits of the RNA homology search suite Infernal (Nawrocki & Eddy, 2013) as an input to look for corresponding operons. A simple schematic presentation of RiboGap tables illustrates the potential connections that can be made between different types of data (Figure 1). Despite its great flexibility and ability to permit the equivalent of complex SQL queries, RiboGap has a simple interface (Figure 2).

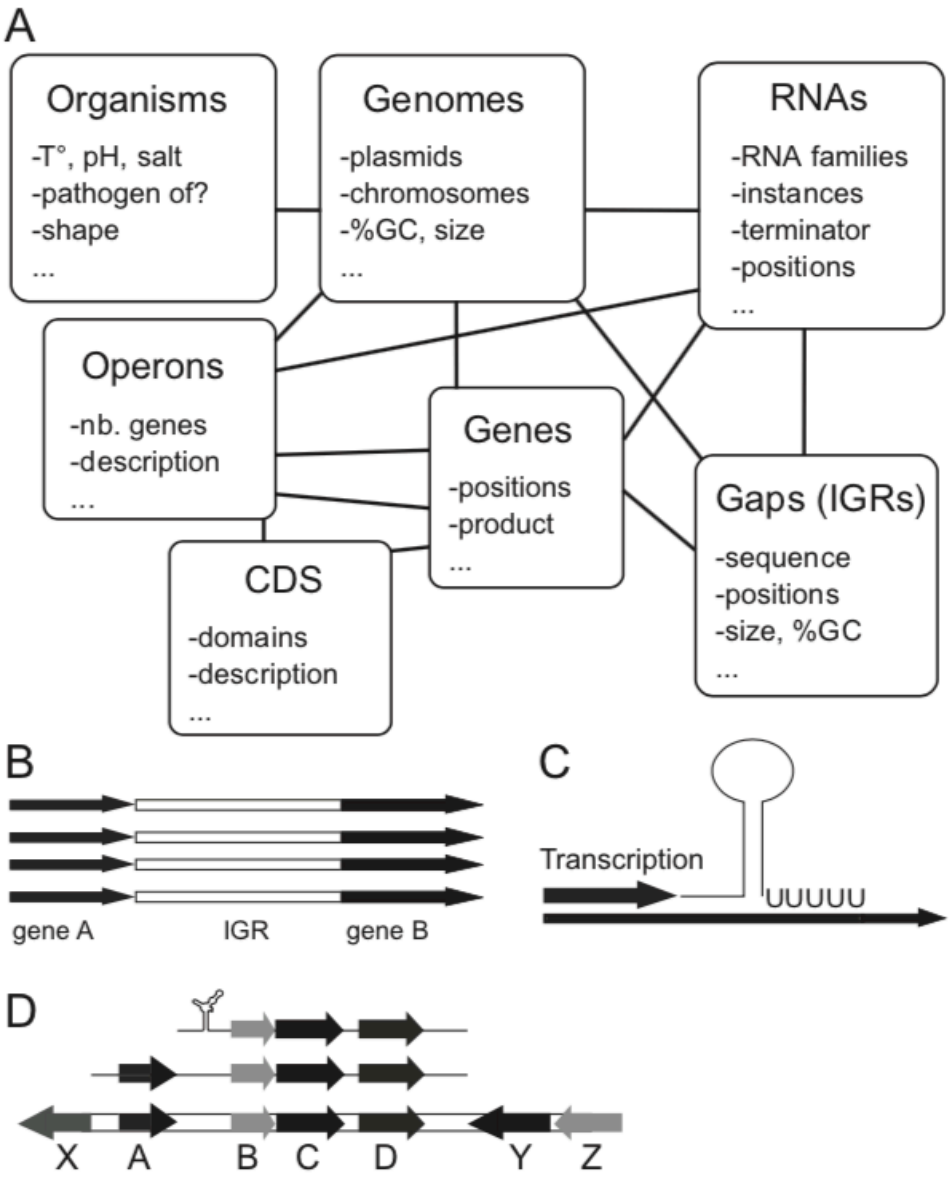

Figure 1. Schematic representation of tables and major applications of RiboGap.
(A) Interconnected seven main tables which contain information on chromosomes, genes, proteins, ncRNAs, intergenic sequences and operons, through which defined queries can be defined. (B) Extraction of intergenic sequences from both the 5´-UTR and the 3´-UTR, which is one of the major applications of RiboGap. (C) De novo discovery of Rho independent terminator positions in specific ncRNAs or terminator evaluation for certain intergenic regions. (D) Prediction of genes and operons controlled by specific regulatory elements like riboswitches.

b. GraphClust

GraphClust requires several algorithms to be installed before using it: LocARNA version 1.7 or more recent (Will *et al*., 2007), Vienna RNApackage version 2.0 or more recent (Lorenz *et al*., 2011), RNAz version 2.1 (Gruber *et al*., 2010), Infernal version 1.0.2 (Nawrocki *et al*., 2009), CMfinder version 0.2 (Yao *et al*., 2006a), RNAshapes version 2.0.6 (Janssen & Giegerich, 2015) and BLASTClust from NCBI (Camacho *et al*., 2009).

c. Optional software

The text editor Emacs helps visualize the RNA alignment with the plugin Ralee (Griffiths-Jones, 2005). R2R is a program useful for displaying RNA secondary structures by summarizing alignment information which can be installed locally (Weinberg & Breaker, 2011). MEME, which can be used locally or on a web server, finds conserved motifs among multiple sequences (Bailey *et al*., 2009).

d. Finding known RNA structures

While most proteins are annotated, only a few genomes have a thorough annotation of ncRNAs. Indeed, aside from tRNAs, rRNAs, SRP RNA and tmRNA, RNAs are typically not part of standard annotation pipelines. As such, sRNAs and ncRNAs found in UTR regions need to be found in other databases such as Rfam, which can be especially cumbersome when one wants to search for ncRNAs associated with particular genes. To do this with RiboGap, the user simply needs to choose keywords and fields of interest like the "description" field in the section "rna_family" and position information from the "rna_known" section, combined with "gene product" from the "cds" (coding sequence) section, as well as any other information the user wishes to display, such as the species name obtained with the "description" field from "fragment" (Figure 2). For example, a user who would be interested in the function and regulation of the uca gene could use the keyword "urea carboxylase" in the field "gene product" to look for known RNAs upstream of genes related to the same function, but not necessarily associated with uca. The result of this query shows an association of ykkC-yxkD leader and mini-ykkC RNAs with genes annotated as urea carboxylases (Supplementary Table S1). These RNAs have recently been renamed guanidine-I and guanidine-II riboswitches, respectively (Breaker *et al*., 2017; Sherlock & Breaker, 2017; Sherlock *et al*., 2017). Interestingly, a few Rho-independent terminators found in the same UTRs overlap the riboswitch position, providing a clear hypothesis for the expression platform for these representatives (Figure 3.A).

Inversely, it is possible to query for genes putatively regulated by given riboswitches. For instance, to find the genes that the SAM/SAH riboswitch potentially regulates, we can examine the gene located immediately downstream of the motif, or evaluate multiple genes if there appears to be an operon (Montange & Batey, 2006; Wang & Breaker, 2008). Additional details are available in supplementary materials in section S1.1.2 and supplementary table S2.

e. Finding sequence motifs

i. Known sequence motifs

In addition to relatively complex RNA structures, leader regions of mRNAs can harbor simple regulatory sequence motifs bound by proteins. For instance, the Trp RNA-binding Attenuation Protein (TRAP) (Gollnick *et al*., 2005) represses the synthesis pathway of tryptophan when its concentration is high. This repression is achieved when the 11 units of the undecameric TRAP complex bind as many copies of a three nucleotide-motif starting by U or G and followed by A and G. These are separated by two or three nucleotides. The complete motif can be represented as ((U/G)AG(N)2-3)11. Such a motif can easily be searched for in RiboGap with a regular expression (REGEXP, refer to the help page of RiboGap for more information) :

[TG]AG[ACGT]{2,3}[TG]AG[ACGT]{2,3}[TG]AG[ACGT]{2,3}[TG]AG[ACGT]{2,3}[TG]AG[ACGT]{2,3}[TG]AG[ACGT]{2,3}[TG]AG[ACGT]{2,3}[TG]AG[ACGT]{2,3}[TG]AG[ACGT]{2,3}[TG]AG[ACGT]{2,3}[TG]AG[ACGT]{2,3}[TG]AG[ACGT]{2,3}[TG]AG[ACGT]

Alternatively, to allow for a slight divergence from the known and well-characterized motif, we can look for fewer repeats and allow the insertion of a long sequence, which could hypothetically be looped out to permit the binding of the TRAP complex with multiple copies of the short RNA sequence:

[TG]AG[ACGT]{2,3}[TG]AG[ACGT]{2,3}[TG]AG[ACGT]{2,3}[TG]AG[ACGT]{2,100}[TG]AG[ACGT]{2,3}[TG]AG[ACGT]{2,3}[TG]AG[ACGT]{2,3}[TG]AG[ACGT]

Using this motif to search for TRAP-mediated regulation, we found many of the expected instances of known TRAP binding sites, such as the one upstream of trpE in Bacillus subtilis. We also found an interesting hit upstream of trpB (encoding a subunit of the tryptophan synthase) in the archaea Methanosarcina acetivorans (Figure 3.B). Because the distribution of TRAP is almost exclusively limited to firmicutes, this may look like a spurious hit. However, the presence of a TRAP analog in the genome of this archaea suggests that this example is legitimate and that a horizontal transfer

event has occurred. There were, however, numerous hits unlikely to be true TRAP binding sites since the corresponding genomes do not encode the TRAP protein and because the genes downstream do not encode tryptophan-related functions. The complete set of hits is provided in Table S3 and highlights other interesting hits not documented before, such as representatives upstream of genes encoding chorismate-binding-like protein.

As an additional example, the protein RsmA/CsrA is known to bind two stem-loops with the sequence "GGA" in their loops (Lapouge *et al*., 2008; Valverde *et al*., 2004). It was recently shown that both in Escherichia coli and Pseudomonas aeruginosa, the protein regulates itself by binding to its own mRNA, which also harbors the typical binding motif (Jean-Pierre *et al*., 2015; Romeo *et al*., 2013). We assumed this would be the case for more species, and thus we searched upstream of csrA or rsmA for instances of the motif. To do so, we allowed appropriate space for two adjacent stem-loops with a "GGA", with the second one overlapping the ribosome binding site:

GGA[ACGT]{4,50}GGA[ACGT]{3,10}[ACGT]$

This search query corresponds to "GGA" followed by 4 to 50 nucleotides and a second "GGA" with 4 to 11 nucleotides at the exact end of the IGR (indicated by "$") and thus directly upstream of the start codon. The proportion of csrA/rsmA IGRs that had this motif was higher than for any other sets of genes we examined: 51% as opposed to 14% on average for other genes (supplementary Table S4). Again, this highlights how simple queries in RiboGap can help to quickly test a hypothesis or to generate data that warrants further analysis. The complete collection of queries and resulting data are available in supplementary materials section S1.2.2 (including an example on how to easily generate a query in MySQL language).

The same principle was used to find potential unannotated small ORFs in IGRs. Given that the optimal ribosome binding site of all proteobacteria is AGGAGG, we looked for this motif followed by a spacer and an AUG start codon, followed by a certain number of "codons" and a stop codon. This query was made more complex by the fact that the coding sequence needs to be a multiple of three, but this can easily be circumvented by copying many times the basic REGEXP pattern:

AGGAGG[ACGT]{6,12}ATG[ACGT] (TAA|TAG|TGA)|

AGGAGG[ACGT]{6,12}ATG[ACGT]{6}(TAA|TAG|TGA)|

AGGAGG[ACGT]{6,12}ATG[ACGT]{9}(TAA|TAG|TGA)|

and so on to 30 nucleotides (10 codons):

|AGGAGG[ACGT]{6,12}ATG[ACGT]{30}(TAA|TAG|TGA)…

Where (TAA|TAG|TGA) corresponds to the three stop codons. This statement thus allowed us to find putative small ORFs in so-called 5′ UTRs encoding between 2 and 29 amino acids. Over 3,000 hits were found in proteobacteria (Figure 3.C and Supplementary Table S5), more details available in supplementary material, section S1.2.3. It has been previously described that such mini-ORFs can affect mRNA stability (Lodato *et al*., 2012; Mathy *et al*., 2007) or efficiency of translation (Qiao *et al*., 2011) suggesting that, through diverse mechanisms, many of these ~3,000 hits are likely to be involved in regulating expression of the downstream gene. Even if REGEXP requires some knowledge of pattern matching for MySQL, the provided examples can be used as templates for many types of searches.

ii. New sequence motifs

Most regulatory RNAs beyond the simple repeats of a TRAP binding motif usually include secondary structural elements. Nevertheless, it is possible to search for unknown motifs in sequences with readily available tools such as MEME. Extracting IGR sequences from genomes is made particularly easy by the RiboGap database. As an example of such a motif search, we conducted a simple query to fetch the "5′ gap" sequences of genes for which the "product" (in the cds section) had the word "iron", such as for the gene product "iron-enterobactin transporter subunit". This was done in the E. coli str. K-12 substr. MG1655 genome (Supplementary Fig. S1) and the sequences provided by RiboGap were submitted to MEME (Bailey & Elkan, 1994). We found a conserved motif corresponding to the Fur-box (Figure 3.D), which is a 19-base-pair inverted DNA repeat known to allow control of gene expression by binding Fur protein according to iron concentration (de Lorenzo *et al*., 1988). While this is not an RNA motif, this search, which took only a few minutes, still yielded an important conserved motif, the Fur-box, from these functionally related sequences (Supplementary Fig. S2). Details are available in Supplementary material, supplementary section S1.2.4.

f. New structures

i. Obtaining intergenic sequences with RiboGap

To discover new ncRNA structures with the pipeline illustrated in Figure 4, the IGR sequences for a chosen gene function or collection need to be extracted. For this purpose, RiboGap is a useful database (available at: ribogap.iaf.inrs.ca). Two fields of the table cds (coding sequence) should be selected: gene and product (Supplementary Fig. S3). Note that more fields can be selected in every step as desired. Second, select the DNA fragment and description fields from Chromosome

information to get the plasmid/chromosome accession numbers and the bacteria strain names, respectively. Third, the IGR sequences should be selected from the gap5 table (Supplementary Fig. S3), together with all the fields from gap5. Finally, the search should be narrowed down by using the conditions section. Any keyword can be used for the product or gene fields in the cds sub-section with either "REGEXP" or "find some pattern" (Supplementary Fig. S4). The keywords "methyl" and "RNA" are used to find genes with functions related to RNA methylation, such as genes encoding "tRNA methylases" or "16S rRNA methyltransferases" (Supplementary Fig. S4). The other example uses pattern matching with the REGEXP option. This can be useful for searches that require more complex keywords, which is the case for cation-related genes, such as magnesium, iron or calcium transporters/exporters/efflux pump (Supplementary Fig. S5).

The rationale for our focus on "methyl" and "RNA" was that RNA methylases could potentially interact and modify their own mRNA to repress their expression in a classical negative feedback loop. Using RiboGap we extracted IGRs in front of genes annotated as RNA methylases (e.g., tRNA (mnm(5)s(2)U34-methyltransferase). We retrieved 8150 sequences and used this dataset with GraphClust.

In general, to search for ncRNAs in IGRs, one additional condition should be added (Supplementary Figure S4). In the gap5 sub-section, a size bigger than (>=) 25 should be used to filter small IGRs that are unlikely to harbor ncRNAs and otherwise would greatly increase the noise and false positives in the search for conserved structured RNAs. Indeed, in many species, the vast majority of IGRs are smaller than 25 nucleotides, while most ncRNAs are observed in IGRs of at least 100 nucleotides (Meyer *et al*., 2009). A cut-off of 100 nucleotides would thus be appropriate, but by being very conservative with a 25-nucleotide lower limit, we still eliminate a large number of IGRs, while retaining the possibility of finding a particularly small structured RNA in noncoding sequences between genes of the same operon. Otherwise, 25 base-pairs are not even enough to accommodate a typical promoter.

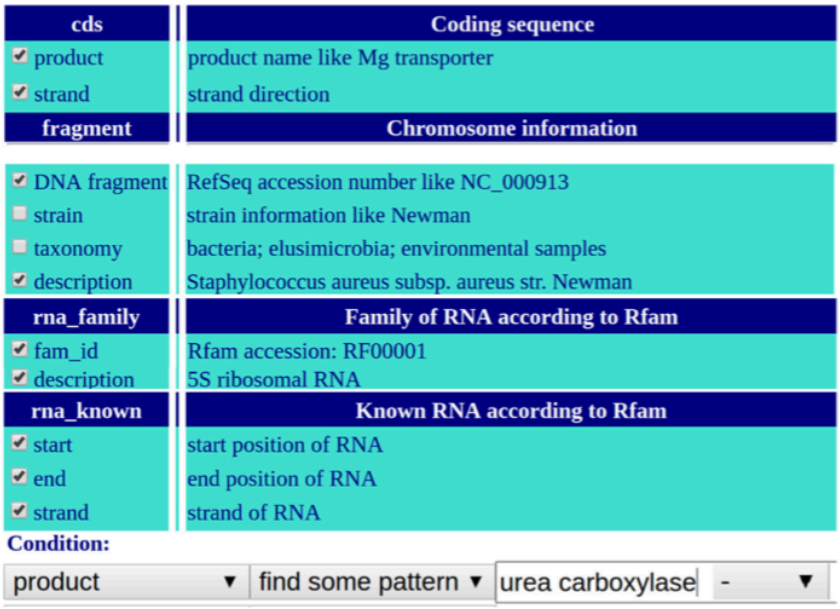

Figure 2. Screenshot of RiboGap interface from the Advanced search page. Some of the empty fields have been cropped to show the full query.

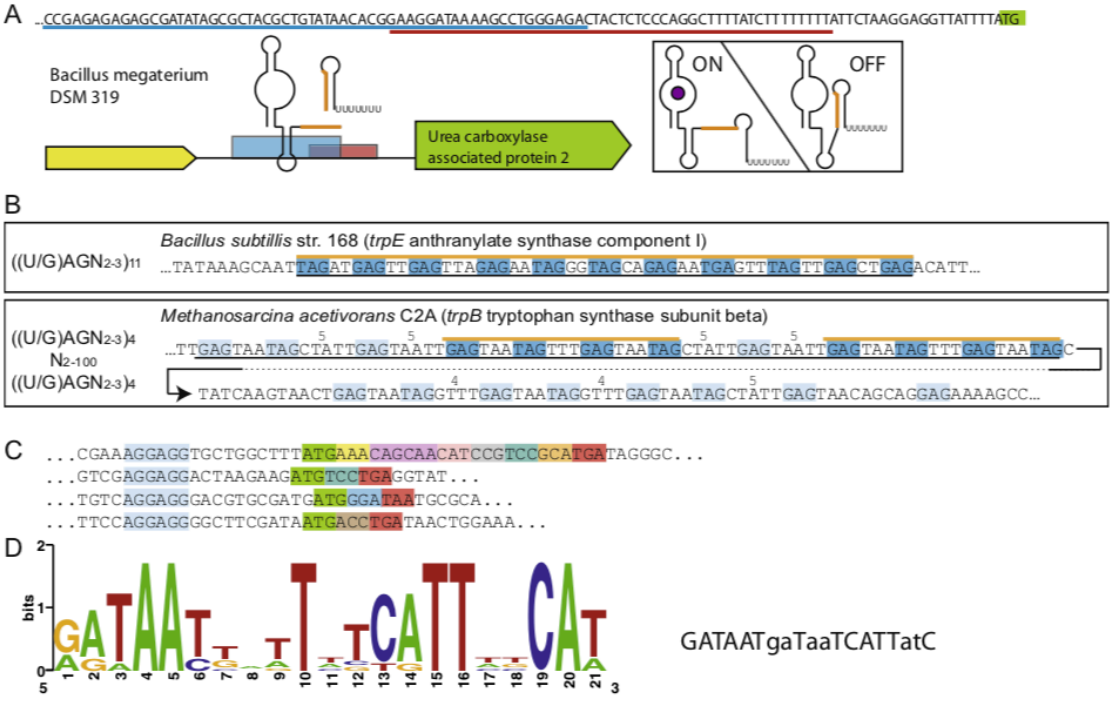

Figure 3. Results from RiboGap queries

(A) The expression platform of a ykkC-yxkD (guanidine-I) riboswitch is observed from the overlap of the Rho-independent terminator (sequence underlined in red) and aptamer (partial sequence, underlined in blue). Genetic context, sequence overlap of structures is shown in orange. Hypothetical ON and OFF states are pictured on the right. (B) Examples of TRAP binding motifs found with the patterns used for the search on the left. Top is the canonical pattern, with a hit corresponding to a confirmed TRAP-regulatory site. Bottom, the pattern used for the search was less stringent and revealed a putative archaeon TRAP-like binding site. Black lines under the sequences indicate the putative full TRAP motif. Blue boxes and orange line correspond to binding sites found by the search, pale blue to other potential binding sites. Spacers larger than the typical 2-3 bases are indicated above the sequence. (C) Examples of mini ORFs found by the RiboGap query. Mini-ORF Shine-Dalgarno boxed in blue, start codon in green, stop codon in red and other codons in different colors. (D) Conserved "iron-box" found by MEME and drawn by sequence logo (Crooks *et al*., 2004) on the left, Fur-box consensus on the right. Positions matching the "iron-box" consensus are in bold.

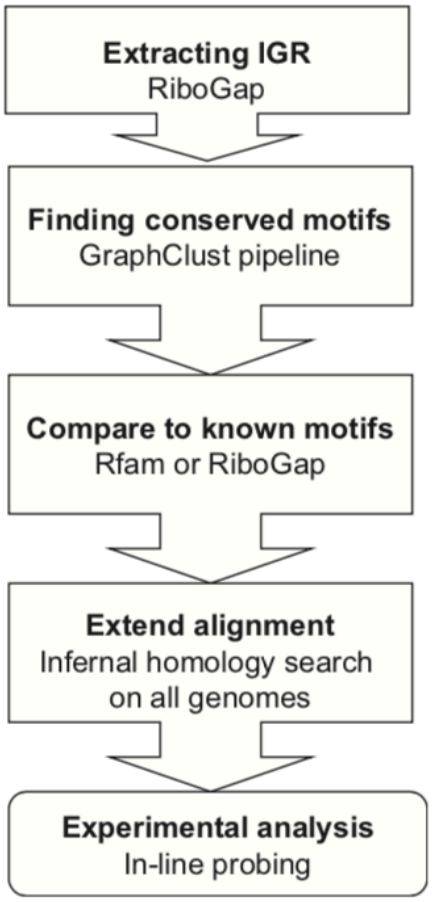

Figure 4. Schematic presentation of the pipeline for novel RNA structures discovery.

ii. Finding conserved elements and predicting RNA secondary structures

In this critical step, the GraphClust package is used to compare all the sequences of the FASTA file generated by the query to group in clusters the sequences that have homology to their primary sequence and secondary structure. This package generates many alignments, as well as figures and graphs that are used to help discriminate ncRNAs from false positives. To predict new structured RNAs, the FASTA file should be provided as input for GraphClust. GraphClust uses the BLASTClust (Altschul *et al*., 1990) package to remove sequences that are too closely related. The default filter screens at 90%, which permits it to automatically discard sequences that are more than 90% related. This ensures that conservation is due to function, rather than close phylogenetic relationship. It also favors alignments with more covariation.

To run GraphClust, some preinstalled software is required, and more information is provided with the GraphClust package. Several parameters can be modified, but the default command line for GraphClust is:

MASTER_GraphClust.pl --root run_test_1 --fasta my_seqs.fasta --config config.default_global –verbose

Other default parameters include a window size of 150 nucleotides, minimum sequence length of 100 nucleotides and two iterations. These parameters provided excellent ncRNA candidates containing secondary structures very well supported by covariation. To verify whether the candidate ncRNA may already be known, it is useful to do the same query, but with more information from tables such as RNA_family and RNA_known for the query (Supplementary Fig. S6) as described in section 3.1. These fields provide information on ncRNAs from the Rfam database (Nawrocki *et al*., 2015), which is particularly useful for identifying known RNAs and avoiding the "discovery" of an already well-known ncRNAs. By selecting both the boxes of Known RNA and RNA family sub-sections (Supplementary Fig. S6) all the intergenic sequences containing known RNAs will be shown. If no results are returned, it means that there is no known RNA assigned to the IGR. For our "RNA" and "methyl" query, all the good candidates obtained with the described settings corresponded to known ncRNAs: tRNA, RNase P, cobalamin riboswitch and SAM riboswitch (SAM example in Supplementary Fig. S8). Indeed, the more widespread a ncRNA is, and the more covariation it has, the more it is likely that it has already been found. We thus executed the GraphClust command using 97% to filter out only very closely

related sequences with BLASTClust and using 15 iterations to get more candidates (details on how to use different parameters in Supplementary Data section S2).

g. How to analyze results

i. 2.1.7.1 Analyzing the structure alignments

By running GraphClust with default parameters, 20 clusters of sequences with predicted RNA secondary structures can be obtained. Two folders are interesting for analysis, CLUSTER and RESULT. These folders contain numerous files, and so we will mention only the most useful files for candidate analysis. In the CLUSTER folder, there are three subfolders, but the most important is the MODEL subfolder where the sequence alignments and predicted secondary structures can be found. MODEL has the clusters with high scores. The RESULT folder contains the final refined results from CLUSTER but sometimes omits useful information by discarding sequences from other intermediate alignments. For this reason, it can be useful to examine the model.tree.aln.ps, model.tree.aln.alirna.ps, best_subtree.aln.ps and model.tree.aln.rel_plot .pdf files as well (Supplementary Fig. S8).

To analyze the predicted structure, several criteria should be taken into account: sequence alignment quality, sequence conservation, covariation and structure likelihood. Since the sequences were aligned based on structure conservation, it is important to look at sequence alignments both in predicted models by CMfinder (Yao *et al.*, 2006b) and LocARNA (Will *et al.*, 2007). It is worthwhile to visualize alignments with more than 10 sequences with Emacs in Ralee-mode (Griffiths-Jones, 2005) with the cmfinder.stk file, this allows easier inspection of base-pairs for which mis-pairs occur in a few sequences of the alignment (Supplementary Fig. S9). Indeed, LocARNA settings cause mispairs to dim color coding of base-pairs, which can lead to the base-pair not being highlighted at all in large and diverse alignments. Note that alignments with at least three sequences can be considered as candidates, depending on the criteria met.

The first criterion is sequence conservation. An alignment with too much conservation is not ideal because in such cases, apparent structure conservation could result from nearly identical sequences from phylogenetically close strains (Supplementary Fig. S10). However, some level of sequence conservation is expected (Supplementary Fig. S11), otherwise, spurious conserved structures can result from alignments of unrelated stem-loops for instance. Depending on the type of ncRNA, double-stranded regions are often more conserved because mutations that disrupt structure are not tolerated. However, in the case of ncRNAs with a very wide phylogenetic distribution, positions

in stems can vary considerably and in such cases, critical single stranded regions can have a higher level of conservation, which can be a good indicator for quality of sequence alignment.

The second criterion is covariation of sequences. Since RNA structures are conserved evolutionarily, then mutations at positions where base-pairing need to be preserved should be compensated by a mutation of the opposite nucleotide that restores the base-pairing to maintain RNA structure (Fox & Woese, 1975). More covariation in the alignment indicates a higher probability of a structured functional RNA. Covariation is the strongest indicator that helps support and confirm ncRNA candidates. While more likely to occur by chance, compatible variations are often observed in alignments of structured RNAs and also support the model. Compatible variations correspond to instances when only one position of the base-pair varies, but still permits base-pairing. For example, in the case of an A-U base-pair compared to a G-U base-pair, only the purine changes, but the U stays the same while still permitting a base-pair at that position. Several examples are provided to illustrate how to distinguish good (Supplementary Fig. S12), ambiguous (Supplementary Fig. S13), partially good (Supplementary Fig. S14 and S15), and poor sequence alignments or covariation (Supplementary Fig. S16 and 17).

The third criterion is the reliability of structures and alignments. To evaluate this, we use the files model.tree.aln.ps, bestsubtree.ps and model.tree.aln.rel_plot.pdf. The latter indicates the reliability of the predicted structure at given positions and is generated by LocARNA-P (Will *et al*., 2012). This file shows the degree of reliability of sequence alignment and structure. The more it tends towards "1", the higher the probability of forming the corresponding base-pair at each given position. Note that this prediction is made from the model.tree.aln.ps file, which is created by LocARNA. (Supplementary Fig S18-S21). The file model.tree.aln.ps is an initial model prediction of ncRNA. LocARNA uses this file as a guide for making secondary structure predictions for more sequences. The file of bestsubtree.ps groups the best sequences for the predicted model. Reliability of the predicted structure for model.tree.aln.ps is evaluated by LocARNA-P and presented in the tree.aln.rel_plot.pdf file. This file indicates the boundary of ncRNAs as well as the reliability of the structure and sequence alignment. The boundaries of the ncRNA are very useful to locate more precisely the ncRNA position in the IGR.

As a result of this procedure, we selected the "RNA-methyl-28" (cluster number 28) from the RESULTS folder as an example of a potential candidate for further analysis (Figure 5). The alignment has many positions with compatible variation and all the sequences clearly belong to

the alignment, as illustrated by the highly-conserved stem around position 30 (Figure 5 and Supplementary Fig. S22). Note that RNA-methyl-28 has no hits in Rfam, indicating that this putative ncRNA is likely a novel RNA family.

ii. Analyzing potential riboswitches

To selectively bind metabolites, riboswitches must fold in defined binding pockets which are determined by precise secondary and tertiary structures that are very well conserved. As a consequence, riboswitches often have a "conservation and structural signature". Many functional RNAs can have several elements in their structure: loops, bulges or multistem junctions. In general, functional RNAs such as riboswitches or ribozymes have several of these structural elements with conserved nucleotides (examples of a candidate unlikely to be a riboswitch, Supplementary Fig S23; of the Mg2+-II riboswitch [Mg-sensor], S24-S27; and of the Mg2+-I riboswitch [ykoK], S28-S31). When they are phylogenetically widespread, the base-pairing regions often harbor a lot of covariation and thus have low sequence conservation. In contrast, nucleotides in single-stranded regions involved in ligand binding can be highly conserved, and are typically in multistem junctions. These features can help to distinguish between different types of ncRNAs and to choose candidate clusters more likely to be riboswitches for further analysis.

iii. Additional considerations

The proximity of the candidate structure to a potential expression platform can also be a good indication that it acts as a regulatory RNA. This can easily be evaluated with the positions of the putative ncRNA within the IGR (which are the numbers in the name of the sequence in the alignment). Subtracting the end position from the size of the IGR gives the number of nucleotides separating the RNA structure from the start codon. Hence, if this number is less than 10, chances are that the structure overlaps the Shine-Dalgarno sequence (Chen *et al*., 1994; Curry & Tomich, 1988) and thus blocks translation, at least in some conditions. Additional criteria include whether the structure is already known (see section 3.1.) or if it is close to a Rho-independent transcription terminator, which would also suggest a potential mode of regulation, through transcription termination in this case. In the case of the candidate RNA-methyl-28, we evaluated the distance of the RNA structure from the start codon. In all the sequences from the alignment, the short distance separating the ncRNA from the start codon (1-2 nucleotides) implies that the Shine-Dalgarno is sequestered in the RNA structure. This further supports a regulatory function for this putative RNA (Supplementary Fig. S32).

iv. Performing a global homology search of candidate motifs by Infernal

Once a potential candidate is found, GraphClust does a search for each model to extend from the initial "cluster alignment" through all sequences provided as input. However, the aligned structures can be used to evaluate the distribution of the motif in the whole bacterial genomes as well as in additional databases if desired. To do so, the stockholm file, (model.cmfinder.stk) in the CLUSTER/MODEL folder of the candidate should be chosen to do a cmbuild for Infernal. The procedure has been described in more detail previously (El Korbi *et al*., 2014). Briefly, the steps to be executed are as follow: cmbuild and cmcalibrate to prepare the infernal covariance model, cmsearch to do a homology search in the chosen target database, and cmalign to consolidate the hits in a new alignment. If the ncRNA candidate is real, the new alignment is likely to provide more support for the conserved structure, with more covariation notably. It may also highlight which portions of the initially predicted structure are more important, since the other portions might not be present in all sequences of the post infernal alignment.

We applied this analysis to the RNA-methyl-28 motif. Using Infernal, we did a global homology search to evaluate the distribution of the motif in all bacterial genomes (NCBI bacteria genomes version 2015). We found 264 hits in different bacteria (Supplementary information and Table S6). We used R2R to draw the RNA-methyl-28 motif based on information from all 264 hits found with Infernal, and found that the RNA structure consensus derived from the 264 hits is similar to the initial GraphClust structure-alignment (Figure 5). One small stem appears as highly conserved, but not the other stems, which are nevertheless supported by additional covariation in this extended alignment, as opposed to merely a compatible variation in the initial alignment. Note that these two types of base-pair variations are not distinguished by the analyses of GraphClust, but can be observed directly in the alignments or with the help of a R2R summarized consensus (Figure 5). The results from any cmsearch can be uploaded with the tabfile format in RiboGap to get information on each hit.

h. Methode in vitro

i. In-line probing for experimental validation

Once a potential ncRNA candidate is chosen, it has to be experimentally validated to prove that it is a genuine ncRNA and to decipher its function. The sequence to be analyzed can be chosen by comparing the SCORE of cmsearch, which is indicated in the cluster.all.fa file in the RESULT folder. The top score indicates that the sequence fits well with the predicted model. This does not mean that sequences with lower scores are not good. For instance, if the motif is already known,

the lower scoring sequences could actually represent distal variants worth investigating. Indeed, the deoxyguanosine riboswitch was revealed by looking at unusual hits for guanine riboswitches (Kim *et al.*, 2007). Based on this we also decided to look at outliers of Mg2+-II riboswitch alignments (Supplementary data section S4.1).

i. PCR to construct the template for RNA production

The first step to start in vitro assays is the preparation of a PCR template. For this purpose, the complete intergenic sequence can be selected from RiboGap and positions of the candidate ncRNA should be determined within this IGR. About 20 to 50 nucleotides can be added to each extremity of the candidate sequence, i.e. the portion of the sequence corresponding to the alignment, which is smaller than the full IGR. Alternatively, if the transcription start site is known, it can be used to determine the 5′ end. Several PCR templates can be constructed for the in vitro assays. If genomic DNA is readily available, preparing the transcription template simply requires addition of a T7 RNA polymerase promoter to a DNA primer used to amplify the fragment of interest. For RNA-methyl-28, the template for transcription could be amplified from E. coli JM109 genomic DNA with the following forward 5′-TTCTAATACGACTCACTATAGGTAAGTTTCG AATGCACAATA-3′ and reverse 5′- TAAGTTACTCGTCTTACAGG-3′ primers. However, when doing global genomics screens, it is common to end up with sequences from species for which assembly PCR is a more practical option (example of a Mg2+-II riboswitch in Supplementary material section S5.2). Overlapping oligonucleotides can be designed with primerize (https://primerize.stanford.edu/) (Tian *et al.*, 2015). Two or three G nucleotides should be added in 5′ of the sequence for efficient RNA transcription. Note that to do assembly PCR, the concentration of first and last oligonucleotide should be 1 μM and the other oligonucleotides concentration should be 0.1 μM.

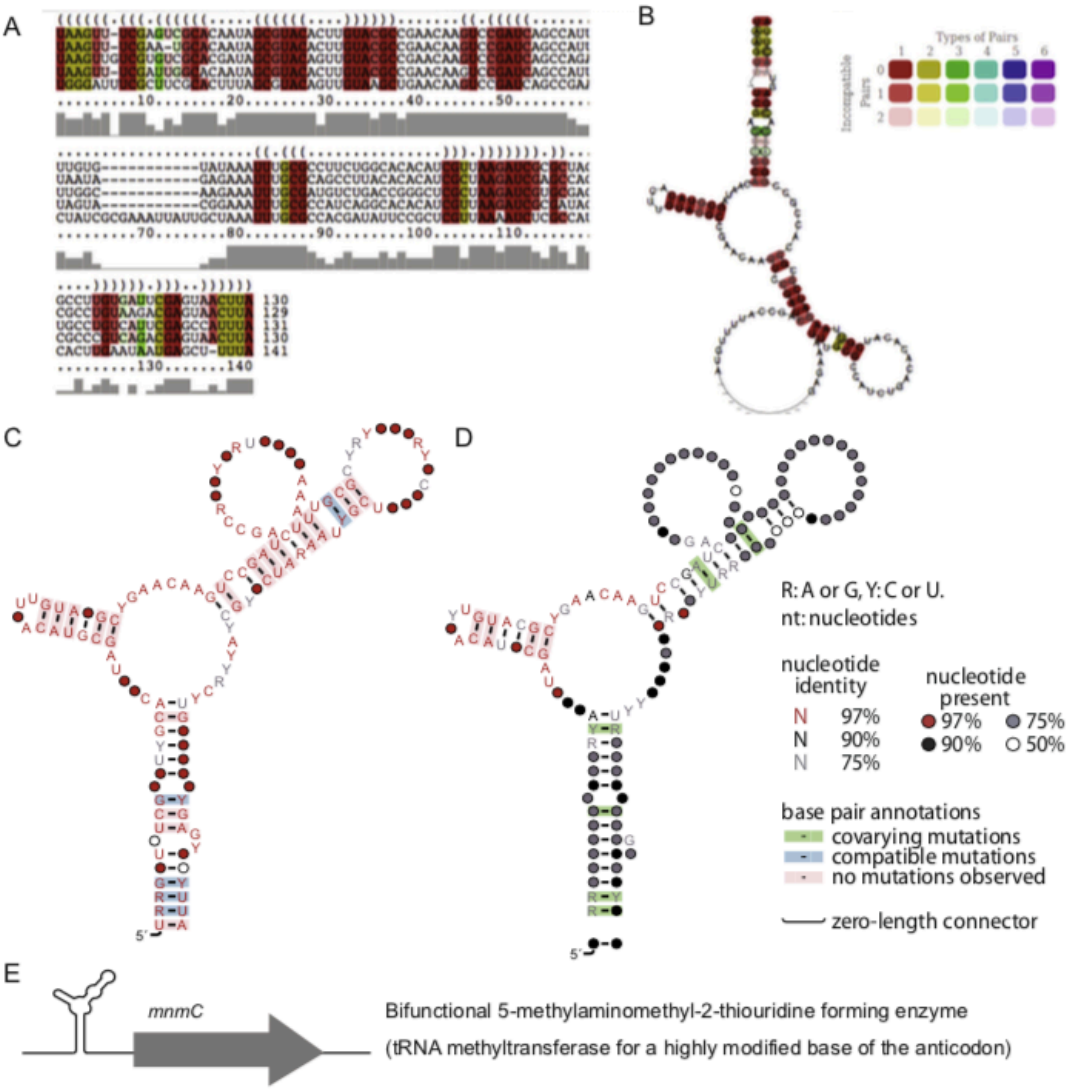

Figure 5. RNA methylation-related candidate "RNA-methyl-28".
(A) Sequence alignment of predicted structure for the candidate. (B) Prediction of secondary structure by GraphClust for the candidate. GraphClust uses the RNAalifold (Bernhart *et al.*, 2008) option of the ViennaRNA Package (Lorenz *et al.*, 2011) for motif drawing. (C) Consensus secondary structures for the RNA-methyl-28 candidate drawn by R2R (Weinberg & Breaker, 2011), a software package that draws consensus motifs and annotations in one figure: secondary structure, sequence conservation, covariation and compatible variation. (D) R2R structure of the Infernal-extended alignment. (E) Genetic context of RNA-methyl-28.

j. RNA transcription

After producing double-stranded DNA templates, RNA can be transcribed from DNA. There are several commercial kits for this purpose. We use 10 μL DNA template (~10 pmoles) with 10 μL of 5X transcription buffer, 15 μL 10 mM rNTPs, 1 μL 0.1 U/μL pyrophosphatase (Roche), 1μL 40 U/μL RNase inhibitor (Roche) and 2 μL T7 RNA polymerase (10 U/μL final concentration) in a final volume of 50 μL. After incubating at 37°C for 2 h and degrading the template with 1 μL 2 U/μL RQ1 DNase, the RNA product is purified by denaturing 6% PAGE for 2 h. RNA is then eluted and dissolved in 21 μL RNase-free water. 1 µL of this sample is used to determine the concentration of RNA by using a Nanodrop spectrophotometer.

k. Dephosphorylation and Labeling

Dephosphorylation of RNA is performed by following the manufacturer's instructions for Antarctic phosphatase (NEB). To label RNA, 2 μL radioactive ATP (γ-32-P), 3 to 10 pmoles of dephosphorylated RNA, 1 μL of 10 U/μL polynucleotide T4 kinase and PNK buffer (NEB) in 20 μL is incubated at 37ºC for 1 h. The labeled product is purified on denaturing 6% PAGE.

l. Determination of candidate RNA structure and potential modulation by in-line probing

To determine the activity of an RNA suspected to be a riboswitch, RNA is incubated in conditions favoring a structure-dependent degradation pattern with in-line probing. In these conditions, different concentrations of ligand can be assayed to test ligand binding and get data necessary for KD calculation. In-line reactions are carried out for 40 h at room temperature. Standard in-line reactions are 50mM Tris pH 8.3, 100 mM KCl and 20 mM MgCl2, but in the case of metal ion ligands, the in-line reaction can be carried with varying concentrations of Mg2+. To be able to determine RNA structure, two types of ladder are prepared. Up to three times as much of labeled RNA can be digested by T1 enzyme and alkaline digestion. T1 reactions are carried out with the radiolabeled RNA and 1.5 μL of T1 RNAse 1 U/μL in T1 solution incubated at 56°C for 5 minutes. Alkaline digestion is conducted with the RNA being incubated in 20 μL of 1X alkaline solution at 90° C for 1 minute and 20 seconds. The samples are then run on 10% denaturing PAGE for approximately 3 hours at 70 W, exposed with phosphorimaging screens and scanned by a Typhoon FL9500. The technique has been described in more detail by Regulski and Breaker (Regulski & Breaker, 2008).

In-line probing was carried out on a construct of the IGR upstream of mnmC in E. coli str. JM109 to confirm the structure of the RNA. We prepared a control in-line probing reaction containing a spontaneous digested RNA without metabolite, a no-reaction sample of undigested RNA, RNA subjected to partial digestion by RNase T1, and a partial alkaline digestion (Na2CO3). The labeled RNA was incubated for 35 hours and then the pattern of RNA degradation was examined by denaturing 10% PAGE (Figure 15). The small highly conserved stem is not supported by the in-line probing data in this construct. This could be due to inappropriate structure prediction, especially considering that the stem is not supported by covariation. Alternatively, since riboswitches and many types of regulatory RNA elements have at least two mutually exclusive structures, in-line probing (Figure 15.B) may not represent the conserved RNA motif candidate RNA-methyl-28, but rather an alternative structure. An additional example for the Mg2+-II riboswitch is provided (Supplementary material section S4.1).

m. In-line probing solutions

Buffer mixtures for RNA preparation and analysis are detailed below. 5X transcription buffer: 400 mM HEPES-KOH (pH 7.5 at 23°C), 120 mM MgCl2, 10 mM spermidine, 200 mM DTT. 2X in-line probing buffer:100 mM Tris (pH 8.3 at 23°C), 200 mM KCl). 10X alkaline solution: 1M Na2CO3 at pH 9.2. 2X RNase T1 digest buffer: sodium citrate 75 mM at pH 5.0 and 30% formamide to denature RNA for a more uniform digestion. To stop in-line probing reactions and also to visualize the migration of RNA by denaturing (8 M urea) polyacrylamide gel electrophoresis (PAGE), an equal volume of 2X gel loading buffer (95% formamide, 10 mM EDTA [pH 8], 0.05 % (w/v) bromophenol blue, 0.05% (w/v) xylene cyanol) was added.

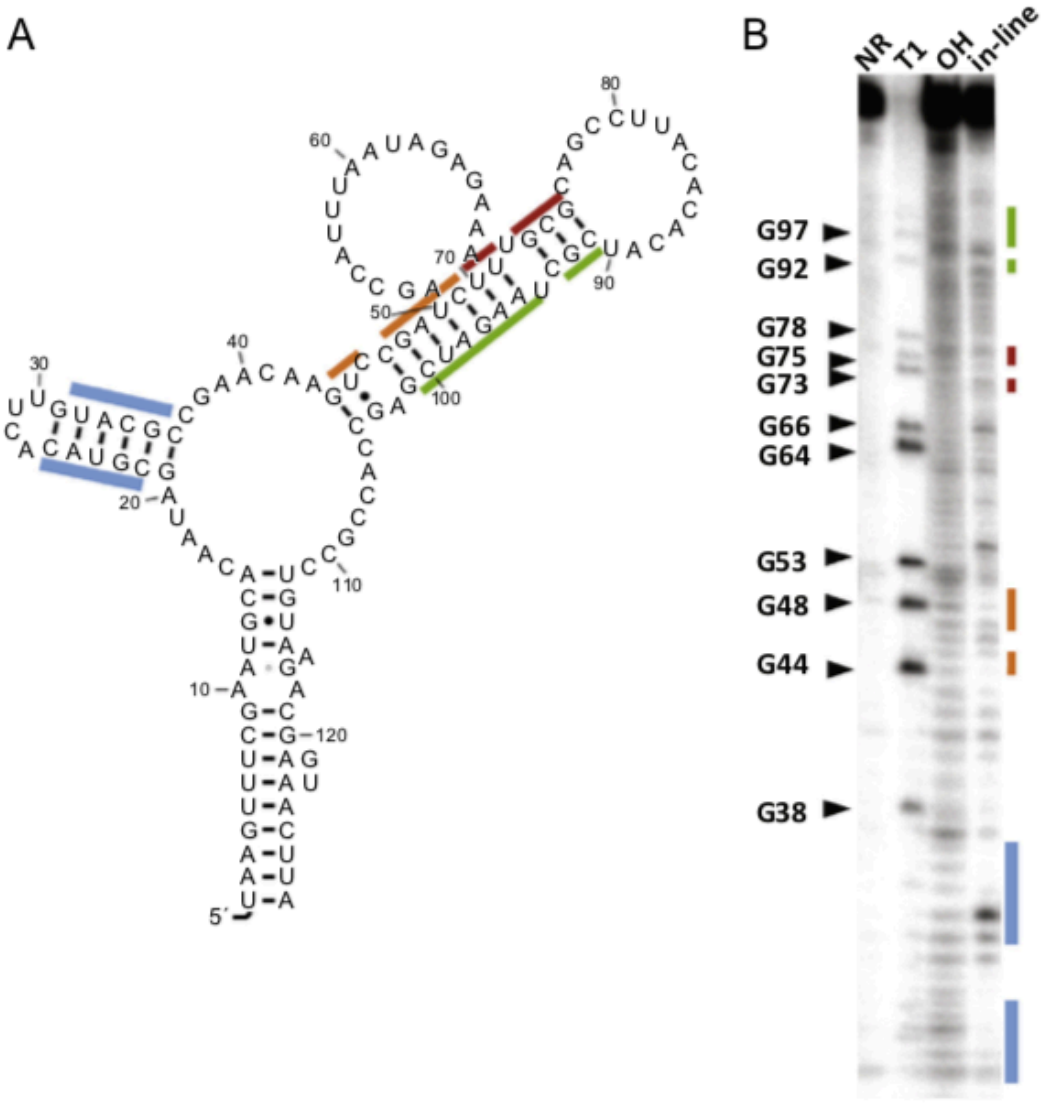

Figure 6. In-line probing of the RNA-methyl-28 candidate.
(A) Secondary structure of the sequence used for the in-line probing experiment. (B) In-line probing of the RNA illustrated in B. RNA was incubated for 40 h. The colored lines along the gel correspond to regions indicated with the same colors on the secondary structure representation.

Discussion

Targeted comparative genomics, where sequences upstream of homologous proteins are aligned to look for conserved RNA structures, have been fruitful in the past (Weinberg *et al*., 2007). However, these studies were undertaken with data extracted from databases such as "NCBI Nucleotide" using scripts or homemade programs and could only be performed by someone with programming skills. Here we demonstrate the accessibility and usefulness of RiboGap in extracting and exploring intergenic sequences for ncRNA. Beyond its ease of use, RiboGap extends the types of sequence ensembles the user can make by allowing function-based queries, rather than protein domain-based queries, providing additional data useful for downstream analysis. RiboGap can be used on a regular basis by most genomics researchers interested in obtaining results from simple or complex queries. Since RiboGap centralizes data from many different databases, it permits users to optimize their research by querying the combined data from various original databases. This cuts down on the laborious compilation and parsing of multiple sets of data required prior to analysis, associated with drawing information from regular databases. The extraction of intergenic sequences is an important part of the pipeline which can lead to de novo predictions of ncRNAs with GraphClust. Even if the latter is relatively efficient to compare sequences on a large scale, comparing all IGRs of all sequenced bacteria and archaea would require considerable computing resources, as opposed to targeting limited datasets provided by RiboGap. This targeted strategy focuses on sets of genes which have a higher likelihood of harboring regulating ncRNAs. Alternatively, the users can choose from multiple sets of functions to explore less obvious associations which may link more subtle regulatory mechanisms related to the ncRNA candidate structures. In both cases, using RiboGap can greatly reduce computing time as well as the number of candidate ncRNAs to evaluate, which is even more time consuming. Perhaps even more critical, choosing which candidate ncRNA to study from the large number of putative ncRNAs requires the analysis of countless alignments and structures, many of which might be interesting, but most of which would not be.

While RiboGap is a powerful tool for extracting intergenic sequences associated with chosen gene functions, it is limited by gene annotations. Poorly annotated genes can either prevent the user from getting a set of sequences corresponding to the chosen function, or lead to the prediction of ncRNAs associated with another unrelated function. Here, we show an example of the latter. Intergenic sequences associated with genes annotated as "urea carboxylases" were searched for

the presence of known ncRNAs. As expected, this led us to find ykkC and mini-ykkC guanidine riboswitches (Breaker *et al*., 2017; Sherlock & Breaker, 2017; Sherlock *et al*., 2017). In this case, the annotation of uca is likely wrong due to a lack of knowledge regarding guanidine biology, resulting in missannotation of uca as encoding urea decarboxylase enzyme, rather than a guanidine decarboxylase.

Our initial screen of magnesium-related genes only identified members of one of the two known Mg2+ riboswitches. We thus adjusted parameters of BLASTClust from 90% to 98% to find the Mg2+-I (ykoK) riboswitch as we had previously done with our RNA-methyl-28 search. Analysis of the GraphClust results should be performed with precaution as none of the currently available software can comprehensively predict the existing RNA secondary structures. Yet, because GraphClust uses both CMfinder and LocARNA, it benefits from different covariation model predictions and alignments of secondary structure instead of merely sequence. Finally, one should not forget that even though some very powerful tools are available for ncRNA prediction, the inspection of the candidate motifs is necessary to appropriately evaluate them and decide which ones to prioritize for experimental validation.

Acknowledgements

J.P. thanks support from Natural Sciences and Engineering Council of Canada (NSERC) [418240 to J.P.]. J.P. is a junior 1 FRQS research scholar. R.R.B. is supported by the NIH (GM022778) and by the Howard Hughes Medical Institute. The authors wish to thank J. Lajoie, V. Korniakova, R. Walsh and E. Boutet for helpful discussions. Computations and data extraction were made on the supercomputer Mammouth parallèle 2, managed by Calcul Québec and Compute Canada (funded by CFI, NanoQuébec, RMGA and FRQ-NT).

Supplementary material

https://ars.els-cdn.com/content/image/1-s2.0-S1046202316303164-mmc1.pdf