

EXTENSION DES FONCTIONNALITÉS DE LA BASE DE DONNÉES RIBOGAP POUR LA DÉCOUVERTE DE NOUVELLES STRUCTURES D'ARN NONCODANTS

Par
Samia Djerroud

Mémoire ou thèse présenté(e) pour l'obtention du grade de
Maître ès Sciences (M.Sc.)
en sciences de Maîtrise en Microbiologie Appliquée

Jury d'évaluation

Président du jury et
examineur interne

Pr Frédéric Veyrier

Examineur externe

Pr Yoann Augagneur
Université de Rennes 1, France

Directeur de recherche

Pr Jonathan Perreault

REMERCIEMENTS

Je tiens à remercier toutes les personnes qui ont contribué à l'aboutissement de mon projet de recherche. Dans un premier temps, j'aimerais remercier mon directeur de recherche le professeur Jonathan Perreault de m'avoir accepté dans son laboratoire, d'avoir été un précieux guide et de m'avoir donné tous les outils nécessaires qui m'ont permis d'être une bonne bio-informaticienne.

J'en profite également pour remercier mes collègues et tous les membres de l'Institut National de la Recherche Scientifique dont le service administratif et informatique pour leur support et leur aide à mon intégration complète.

Je tiens à souligner toute ma reconnaissance pour Mr Hui Zhong Lu pour tout son support durant mon utilisation du serveur Calcul Québec. Il a été d'une aide précieuse et d'un support inconditionnel.

D'un autre côté, j'aimerais remercier mes très chers parents. Leur soutien et leur amour inconditionnels m'ont accompagné durant toutes ces longues années et m'ont permis d'être la personne que je suis aujourd'hui. Je remercie également mon frère, ma sœur et mon mari pour leurs encouragements.

Enfin, j'aimerais exprimer ma gratitude pour mes amis et ma famille qui m'ont soutenu durant les dures épreuves que j'ai rencontrées. Un grand merci à tous !

RÉSUMÉ

Les *riboswitchs* sont des éléments régulateurs localisés en 5' UTR de l'ARNm des gènes qu'ils contrôlent. Pour découvrir de nouveaux *riboswitchs* et autres ARN noncodants (ARNnc), nous avons opté pour une approche bio-informatique car des méthodes expérimentales, bien qu'essentielles pour confirmer les découvertes, seraient très coûteuses et longues si elles devaient être appliquées à l'analyse de toutes les régions intergéniques. D'où la nécessité d'utiliser la bio-informatique en s'appuyant sur la génomique comparative comme étape primaire pour trier les différents candidats.

La méthode bio-informatique consiste en première étape à extraire toutes les régions intergéniques en lien avec un ligand choisi. Pour ce faire j'ai conçu deux bases de données : RiboGap v2 pour les génomes complets (génomes séquencés entièrement) et RiboGap v2.1 pour les génomes incomplets (séquencés partiellement). Ces deux bases de données regroupent les séquences codantes, régions intergéniques, ARNnc, terminateurs Rho dépendants, terminateurs Rho indépendants et prédictions de potentiels promoteurs à partir des génomes complets et incomplets. La deuxième étape de la méthode bio-informatique consiste à exécuter la suite GraphClust sur les régions extraites qui permet la détection des similitudes entre les structures secondaires d'ARN en se basant sur l'alignement des structures des séquences d'ARN. Nous avons choisi le second messenger ppGpp comme cible prometteuse de *riboswitchs* hypothétiques. Les séquences en amont de gènes associés à cette molécule signal ont donc été extraites pour notre étude de génomique comparative.

Mots-clés : séquences codantes, régions intergéniques, terminateurs, promoteurs, opérons, ARN noncodants, *riboswitchs*, GraphClust, génomique comparative.

ABSTRACT

Riboswitches are regulatory elements located in the 5' UTR of the mRNA of the genes they control. To discover new riboswitches and other non-coding RNAs (ncRNAs), we opted for a bioinformatics approach because an experimental method, although essential to confirm the discoveries, would be prohibitively expensive and long if used to analyze all the intergenic regions. Hence the need to use comparative genomics in bioinformatics as a primary step to sort out the different candidates.

The bioinformatics method consists in a first step to extract all the intergenic regions linked to a chosen ligand. For this purpose, I designed two databases: RiboGap v2 for complete genomes (fully sequenced genomes) and RiboGap v2.1 for incomplete genomes (partially sequenced). These two databases group coding sequences, intergenic regions, ncRNAs, Rho-dependent terminators, Rho-independent terminators and putative promoter predictions from complete and incomplete genomes. The second step of the bioinformatics method is to run the GraphClust suite on the extracted regions which allows the detection of similarities between RNA secondary structures based on the alignment of RNA sequence structures. We chose the second messenger ppGpp as a promising target of hypothetical riboswitches. Upstream sequences of genes associated with this signal molecule were therefore extracted for our comparative genomics study.

Keywords: coding sequences, intergenic regions, terminators, promoters, operons, noncoding RNAs, riboswitches, GraphClust, comparative genomics.

TABLE DES MATIÈRES

REMERCIEMENTS	III
RÉSUMÉ	V
ABSTRACT	VII
TABLE DES MATIÈRES	IX
LISTE DES FIGURES.....	XIII
LISTE DES TABLEAUX.....	XV
LISTE DES ABRÉVIATIONS.....	XVII
1 INTRODUCTION.....	19
1.1 GENOMES	19
1.1.1 Gènes / Protéines	20
1.1.2 Régions intergéniques	21
1.1.3 Opérons	21
1.1.4 Promoteurs.....	21
1.1.5 Termineur.....	22
1.2 ARN NONCODANTS	23
1.2.1 Petit ARN (sARN).....	23
1.2.2 ARN ribosomiaux (ARNr).....	24
1.2.3 ARN de transfert (ARNt)	24
1.2.4 Riboswitchs	25
1.3 LA BASE DE DONNEES « RIBOGAP »	29
1.3.1 Sources des données	31
1.3.2 Interface web de RiboGap	32
1.3.3 Utilisation de RiboGap	34
2 PROBLEMATIQUE.....	36
3 HYPOTHESE.....	36
4 OBJECTIFS.....	37
5 RIBOGAP: A RELATIONAL DATABASE FOR PROKARYOTE GENOMICS.....	38
RIBOGAP : UNE BASE DE DONNEES RELATIONNELLE POUR LES PROCARYOTES GENOMIQUES	38
5.1 RESUME.....	38
5.2 ABSTRACT.....	39
5.3 INTRODUCTION	39

5.4	MATERIALS AND METHODS	41
5.4.1	<i>Genomic data download</i>	41
5.4.2	<i>Noncoding RNA detection</i>	41
5.4.3	<i>tRNA and rRNA determination</i>	42
5.4.4	<i>Promoter predictions</i>	42
5.4.5	<i>Terminator annotations</i>	43
5.4.6	<i>Functional gene annotation</i>	44
5.5	RESULTS AND DISCUSSION.....	44
5.5.1	<i>Data in RiboGap v2</i>	44
5.5.2	<i>Use of RiboGap v2</i>	45
5.6	CONCLUSION.....	48
5.7	ABBREVIATION.....	49
5.8	ACKNOWLEDGEMENTS	49
5.9	FUNDING.....	50
5.10	CONTRIBUTIONS.....	50
5.11	REFERENCES	50
5.12	FIGURES	54
5.13	TABLES.....	59
6	DECOUVERTE DE NOUVEAUX ARN NONCODANTS ASSOCIES A PPGPP.....	62
6.1	MATERIELS ET METHODES.....	62
6.1.1	<i>RiboGap</i>	63
6.1.2	<i>GraphClust</i>	65
6.1.3	<i>Sélection de clusters candidats</i>	68
6.1.4	<i>Infernal</i>	68
6.2	RESULTATS.....	69
7	DISCUSSION GÉNÉRALE	74
7.1	UTILITE DE RIBOGAP	74
7.2	OUTILS DE PREDICTION UTILISES DANS LA MISE A JOUR DE LA BASE DE DONNEES	75
7.3	AUGMENTATION DU NOMBRE DE CANDIDATS.....	76
7.4	ABSENCE DU <i>RIBOSWITCH</i> YKKC-YXKD DE LA SOUS-CLASSE PPGPP.....	77
7.5	LA METHODOLOGIE IN-LINE PROBING	77
8	CONCLUSION	78
9	BIBLIOGRAPHIE.....	79
10	ANNEXE I : DIAGRAMME COMPLET DE RIBOGAP	85
11	ANNEXE II : LES GROUPES (CLUSTERS) RESULTATS.....	87
11.1	RIBOGAP V1	87

11.1.1	<i>Cluster 11</i>	87
11.1.2	<i>Cluster 1</i>	88
11.1.3	<i>Cluster 8</i>	89
11.1.4	<i>Cluster 14</i>	90
11.2	RIBOGAP V2.....	92
11.2.1	<i>GraphClust</i>	92
11.2.2	<i>Recherche d'homologie avec Infernal</i>	129
11.3	RIBOGAP V2.1.....	138
11.3.1	<i>GraphClust</i>	138
11.3.2	<i>Recherche d'homologues avec Infernal</i>	149
12	ANNEXE III : DROITS DE REPUBLICATIONS DES IMAGES	153
13	ANNEXE IV : MATERIEL SUPPLEMENTAIRE DE L'ARTICLE « RIBOGAP : A RELATIONAL DATABASE FOR PROKARYOTE GENOMICS »	160
13.1	TABLE DES MATIERES	160
13.2	THE FIRST VERSION OF THE DATABASE « RIBOGAP »	161
13.3	IMPROVEMENT OF THE DATABASE « RIBOGAP »	162
13.3.1	<i>Genomes</i>	162
13.3.2	<i>Coding sequences</i>	167
13.3.3	<i>Intergenic regions</i>	171
13.3.4	<i>Noncoding RNA</i>	172
13.3.5	<i>tRNA</i>	174
13.3.6	<i>rRNA</i>	175
13.3.7	<i>Promoters</i>	176
13.3.8	<i>Terminators</i>	177
13.3.9	<i>Conserved proteins</i>	179
13.4	DATA IN RIBOGAP V2.....	182
13.4.1	<i>RiboGap v2's diagram</i>	182
13.4.2	<i>Compilation of promoter predictions</i>	183
13.5	THE USE OF THE DATABASE.....	187
13.5.1	<i>Find a promoter, a terminator and a riboswitch in the same intergenic region</i>	187
13.5.2	<i>Search the expression platforms of all riboswitches</i>	191
13.5.3	<i>Discover new small RNAs</i>	200
13.5.4	<i>Motif finder : G-quadruplex example</i>	202
13.5.5	<i>Intergenic sequences with cis regulatory RNAs</i>	205
14	ANNEXE V: L'ARTICLE « A SURVEY OF CIS REGULATORY NON-CODING RNA INVOLVED IN BACTERIAL VIRULENCE»	210
14.1	RÉSUMÉ.....	210

LISTE DES FIGURES

FIGURE 1–STRUCTURE D'UNE PARTIE DU GENOME PROCARYOTE	20
FIGURE 2 - DE L'ADN A LA PROTEINE	20
FIGURE 3 - LES BOITES -35 ET -10 D'UN PROMOTEUR EN PRESENCE DE L'ARN POLYMERASE.....	21
FIGURE 4 - STRUCTURE D'UN TERMINATEUR RHO INDEPENDANT	22
FIGURE 5 -STRUCTURE D'UN TERMINATEUR RHO DEPENDANT	23
FIGURE 6 : REGULATION GENETIQUE PAR LES PETITS ARN PAR LE MECANISME HFQ-DEPENDANT	24
FIGURE 7 - STRUCTURE DE L'ARNT.....	25
FIGURE 8 - MECANISMES DE REGULATION DES <i>RIBOSWITCHS</i>	26
FIGURE 9 - LE <i>RIBOSWITCH</i> YKKC-YXKD.	28
FIGURE 10 : MECANISME D'ACTION DE PP _{GPP}	29
FIGURE 11 : SCHEMA SIMPLIFIE DE LA BASE DE DONNEES DE RIBOGAP-VERSION1.....	30
FIGURE 12 - TABLES DE L'INTERFACE WEB DE RIBOGAP V1	33
FIGURE 13: SECTION DES CONDITIONS EMISES PAR L'UTILISATEUR.....	34
FIGURE 14 - RIBOGAP'S V2 AND V2.1 SIMPLIFIED DIAGRAM (FIG 1 IN THE ARTICLE)	54
FIGURE 15- ILLUSTRATION OF DIFFERENT REGULATORY ELEMENTS (FIG 2 IN THE ARTICLE)	55
FIGURE 16 - DISTANCES BETWEEN RIBOSWITCHES AND EXPRESSION PLATFORMS (FIG 3 IN THE ARTICLE)	56
FIGURE 17 - REGULATORY ELEMENT CONTEXT OF A SMALL RNA (FIG 4 IN THE ARTICLE).....	58
FIGURE 18 : DIAGRAMME DU PIPELINE DE <i>CLUSTERING</i> COMPLET. LES PHASES EXECUTEES EN PARALLELE SONT REPRESENTEES DANS DES BOITES EMPILEES	66
FIGURE 19 - ALIGNEMENT STRUCTUREL DU CLUSTER 45	70
FIGURE 20 - STRUCTURE SECONDAIRE PREDITE POUR LE CLUSTER 45.....	71
FIGURE 21 - CONTEXTE GENETIQUE DE L'UNE DES SEQUENCES INTERGENIQUES DU CLUSTER 45	72
FIGURE 22 - CLUSTER RESULTANT DES SEQUENCES HOMOLOGUES DU CLUSTER 45	73
FIGURE 23 - <i>CLUSTER</i> RESULTANT DES SEQUENCES HOMOLOGUES DU <i>CLUSTER</i> 45	73
FIGURE 24 - ALIGNEMENT STRUCTUREL DU CLUSTER 11	87
FIGURE 25 - ALIGNEMENT STRUCTUREL DU CLUSTER 1.....	88
FIGURE 26 - ALIGNEMENT STRUCTUREL DU CLUSTER 8.....	89
FIGURE 27–ALIGNEMENT STRUCTUREL DU CLUSTER 14	90
FIGURE 28 - STRUCTURES SECONDAIRES DES <i>CLUSTERS</i> 1, 8, 11 ET14 GENEREES AVEC R2R	91
FIGURE 29 - EXEMPLES DE <i>CLUSTERS</i> INTERESSANTS POUR LA REQUETE 1	114
FIGURE 30 - STRUCTURES SECONDAIRES OBTENUES PAR GRAPHCLUST	128
FIGURE 31 - STRUCTURES SECONDAIRES OBTENUES PAR GRAPHCLUST	133
FIGURE 32 - EXEMPLES DE STRUCTURES SECONDAIRES DES SEQUENCES DE LA REQUETE 2	138

FIGURE 33 - STRUCTURE SECONDAIRES DE QUELQUES <i>CLUSTERS</i> SELECTIONNEES	144
FIGURE 34 - STRUCTURES SECONDAIRES DE CANDIDATS INTERESSANTS	148
FIGURE 35–FIRST VERSION OF RIBOGAP TABLES AND ATTRIBUTES (FIG. S1 IN THE SUPPLEMENTARY DATA)....	161
FIGURE 36 - THE WEB INTERFACE OF THE DATABASE (FIG. S2 IN THE SUPPLEMENTARY DATA).....	162
FIGURE 37 - THE CODING SEQUENCE EXTRACTED FROM A GENBANK FILE (FIG. S3 IN THE SUPPLEMENTARY DATA)	167
FIGURE 38 - AN EXAMPLE OF A FALSE POSITIVE GIVEN WHILE USING BLAST. (FIG. S4 IN THE SUPPLEMENTARY DATA)	174
FIGURE 39 - TRNASCAN OUTPUT (FIG. S5 IN THE SUPPLEMENTARY DATA)	175
FIGURE 40 - RRNA POSITIONS FOUND IN GENBANK (FIG. S6 IN THE SUPPLEMENTARY DATA)	175
FIGURE 41 - FINDING PROMOTERS INSIDE INTERGENIC SEQUENCES (FIG. S7 IN THE SUPPLEMENTARY DATA). ..	176
FIGURE 42 - RNIE OUTPUT (FIG. S8 IN THE SUPPLEMENTARY DATA).....	178
FIGURE 43 - DIFFERENT DATABASES THAT COMPOSE INTERPROSCAN (FIG. S9 IN THE SUPPLEMENTARY DATA).180	
FIGURE 44 - DETAILED DIAGRAM OF RIBOGAP V2 (FIG. S10 IN THE SUPPLEMENTARY DATA).....	182
FIGURE 45 - SELECTION OF PROMOTERS FOR THE GENOME TAXONOMY "PROTEOBACTERIA" (FIG. S11 IN THE SUPPLEMENTARY DATA).....	183
FIGURE 46 - EVALUATION OF PROMOTER CONSERVATION CAN HELP VALIDATE PREDICTED -35 AND -10 BOXES.A (FIG. S12 IN THE SUPPLEMENTARY DATA)	186
FIGURE 47 - STEP1: SELECT IGRs WITH TERMINATORS AND PROMOTERS (FIG. S13 IN THE SUPPLEMENTARY DATA)	188
FIGURE 48 - STEP2: SELECT IGRs WITH RIBOSWITCHES (FIG. S14 IN THE SUPPLEMENTARY DATA)	188
FIGURE 49 - STEP1: SELECT IGRs WITH RIBOSWITCHES (FIG. S15 IN THE SUPPLEMENTARY DATA)	191
FIGURE 50 - STEP2: SELECT ALL THE NCRNA (NO TERMINATORS) (FIG. S16 IN THE SUPPLEMENTARY DATA)....	192
FIGURE 51 - SELECT IGRs WITH TERMINATOR AND PROMOTER (FIG. S17 IN THE SUPPLEMENTARY DATA)	200
FIGURE 52 - USING THE WEB INTERFACE TO SELECT CODING SEQUENCES THAT HAVE G4 MOTIF WITH REGEXP (FIG. S18 IN THE SUPPLEMENTARY DATA)	203
FIGURE 53 - SELECT THERMOREGULATOR AND "PERFECT SHINE-DALGARNO" FROM THE WEB INTERFACE (FIG. S19 IN THE SUPPLEMENTARY DATA)	209

LISTE DES TABLEAUX

TABLEAU 1 - DESCRIPTION DES DIFFERENTES TABLES DE RIBOGAPV1	31
TABLEAU 2 - COMPARISON OF AVAILABLE DATA IN RIBOGAP V1, V2 AND V2.1 (TABLE 1 IN THE ARTICLE)	59
TABLEAU 3 - DISTANCES BETWEEN RIBOSWITCH APTAMERS AND PUTATIVE EXPRESSION PLATFORMS (TABLE 2 IN THE ARTICLE)	60
TABLEAU 4 - INTERGENIC SEQUENCES WITH CIS-REGULATORY RNAs WHICH HAVE A PERFECT SHINE-DALGARNO CLOSE TO SD (TABLE 3 IN THE ARTICLE)	61
TABLEAU 5 - COMPARAISON DU NOMBRE DE CANDIDATS TROUVES ENTRE LES TROIS VERSIONS DES BASES DE DONNEES	74
TABLEAU 6- DESCRIPTION OF A GENBANK FORMAT (TABLE S1 IN THE SUPPLEMENTARY DATA)	164
TABLEAU 7 - CODING SEQUENCES EXTRACTED (TABLE S2 IN THE SUPPLEMENTARY DATA).....	168
TABLEAU 8 - DESCRIPTION OF bTSSFINDER RESULT (TABLE S3 IN THE SUPPLEMENTARY DATA).	177
TABLEAU 9-EXAMPLE OF AN OUTPUT OF RHO TERMPREDICT (TABLE S4 IN THE SUPPLEMENTARY DATA).	179
TABLEAU 10 - EXAMPLE OF GENES THAT CODE FOR “HYPOTHETICAL PROTEINS” (TABLE S5 IN THE SUPPLEMENTARY DATA).	180
TABLEAU 11 - DESCRIPTION OF INTERPROSCAN OUTPUT (TABLE S6 IN THE SUPPLEMENTARY DATA).....	181
TABLEAU 12 - DIFFERENT MYSQL QUERIES THAT CALCULATE THE NUMBER OF PROMOTERS, GENES AND OPERONS ACCORDING TO TAXONOMY (TABLE S7 IN THE SUPPLEMENTARY DATA).....	183
TABLEAU 13 - CALCULATION OF PROMOTERS PER GENES AND OPERONS ACCORDING TO TAXONOMY (TABLE S8 IN THE SUPPLEMENTARY DATA).	184
TABLEAU 14 - ANALYSIS OF THE EXISTENCE OF A RIBOSWITCH, PROMOTER AND A TERMINATOR IN AN INTERGENIC REGION (TABLE S10 IN THE SUPPLEMENTARY DATA).....	190
TABLEAU 15 - MYSQL QUERIES AND PROGRAMS THAT CALCULATE DIFFERENT DISTANCES (TABLE S11 IN THE SUPPLEMENTARY DATA)	193
TABLEAU 16 - OCCURRENCE OF TERMINATORS BETWEEN TWO PROMOTERS (TABLE S12 IN THE SUPPLEMENTARY DATA)	201
TABLEAU 17 - COMPARISON OF G-QUADRUPLEX MOTIF SEARCH IN RIBOGAP WITH G4RNA ANALYSIS. THE MOTIF BOX IS HIGHLIGHTED IN YELLOW (TABLE S14 IN THE SUPPLEMENTARY DATA)	204
TABLEAU 18 - DIFFERENT MYSQL QUERIES THAT CALCULATE THE NUMBER OF INTERGENIC SEQUENCES (TABLE S15 IN THE SUPPLEMENTARY DATA)	205

LISTE DES ABRÉVIATIONS

A	Adénosine
ADN	Acide Désoxyribonucléique
ARN	Acide RiboNucléique
ARNm	ARN messenger
ARNnc	ARN noncodant
ARNr	ARN ribosomal
ARNt	ARN transfer
ATP	Adenosine Triphosphate
C	Cytosine
CDD	<i>Conserved Domains</i> (Domaines conservés)
CDS	<i>Coding Sequence</i> (Séquence codante)
CSV	<i>Comma – Separated Values</i> (valeurs séparées par des virgules)
G	Guanine
GTP	Guanosine triphosphate
IGR	<i>Intergenic Region</i> (Région Intergénique)
NCBI	<i>National Center Biotechnology Information</i>
ODB	<i>Operon DataBase</i> (Base de données des opérons)
RDT	<i>Rho-Dependent Terminators</i> (Termineurs Rho Dépendants)
Rfam	<i>RNA Family DataBase</i> (Base de données des familles d'ARN)
RIT	<i>Rho-Independent Terminators</i> (Termineurs Rho Indépendants)
SAM	S-adenosyl Methionine
sARN	<i>Small RNA</i> (petit ARN)
SD	<i>Shain-Dalgarno</i>
T	Thymine
TPP	Thiamine Pyrophosphate
UML	<i>Unified Modeling Language</i> (Langage de Modélisation Unifié)
U	<i>Uridine</i>
VF	<i>Virulence factors</i> (Facteurs de Virulence)

1 INTRODUCTION

Les procaryotes sont des microorganismes unicellulaires dont la structure ne contient pas de noyau contrairement aux eucaryotes (Krieg, 2005). La séquence d'ADN appelée génome se trouve ainsi directement dans le cytoplasme, habituellement sous forme circulaire, qui peut être composé de chromosomes et parfois aussi de plasmides (ADN extra-chromosomique). Les procaryotes sont divisés en deux domaines de la vie: les bactéries (ex: *Escherichia coli* IA139) et les archées (ex: *Methanococcus maripaludis* C7) (DeLong & Pace, 2001).

La bio-informatique est une science multidisciplinaire (l'informatique, les mathématiques et la biologie) et son rôle indispensable revient au fait que les biologistes chercheurs génèrent une quantité considérable de nouvelles données portant sur les génomes, les gènes, les organismes, les ARN, leurs interactions, leurs structures et leur évolution. D'où le besoin d'utiliser des approches informatiques telles que les bases de données permettant la modélisation, la manipulation et le stockage de ces données souvent avec des associations très complexes (Colston *et al.*, 2014).

Un autre rôle de la bio-informatique est l'analyse et la prédiction d'éléments génétiques qui présentent des structures ou des motifs conservés, telles que : les séquences protéiques, les gènes et les ARN régulateurs (Gu & Bourne, 2009).

1.1 Génomes

Chez les procaryotes, le génome est constitué d'un ou plusieurs chromosomes formés de deux brins d'ADN complémentaires et ayant habituellement une structure circulaire dans le cytoplasme, de plus petites molécules d'ADN circulaire, appelées plasmides, peuvent aussi être présentes. Un chromosome est composé de gènes et de régions intergéniques qui peuvent avoir des promoteurs et des terminateurs (Tatusova *et al.*, 2016). La figure 1 montre la constitution d'une séquence d'ADN chez les procaryotes.

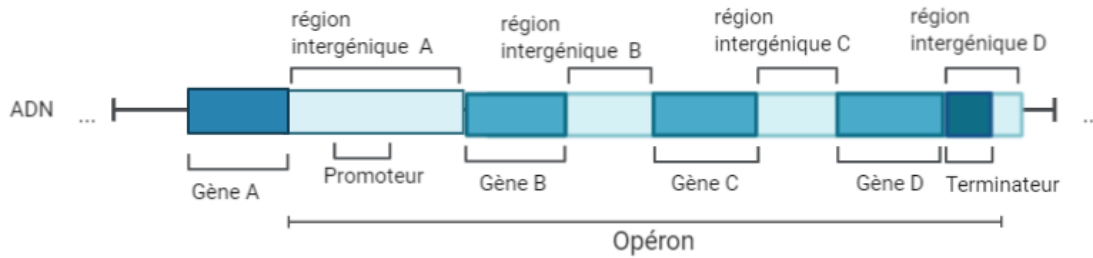


Figure 1–Structure d’une partie du génome procaryote

De nombreux laboratoires de recherche contribuent à l’identification des séquences des génomes grâce au séquençage et à l’assemblage qui consiste à déterminer l’enchaînement des quatre nucléotides : Adénine (A), Thymines (T), Guanine (G) et Cytidine (C) de la séquence d’ADN. L’assemblage du génome entier est parfois difficile, voire impossible, d’où l’existence du terme « niveau d’assemblage » qui permet de savoir si le génome a été entièrement assemblé, on parle donc de génome complet, ou s’il a été partiellement assemblé, on parle dans ce cas de Contigs ou de Scaffolds. Pour faciliter la compréhension des annotations, nous les appellerons génome incomplet dans ce manuscrit (Pruitt *et al.*, 2007).

1.1.1 Gènes / Protéines

Les gènes sont des séquences codantes transcrites en ARNm puis traduits en séquences d’acides aminés où chaque codon est converti en un acide aminé selon le code génétique (Figure 2). La séquence protéique prend ensuite une structure tridimensionnelle spécifique lui conférant sa fonction (Carter & Houlihan, 2001).

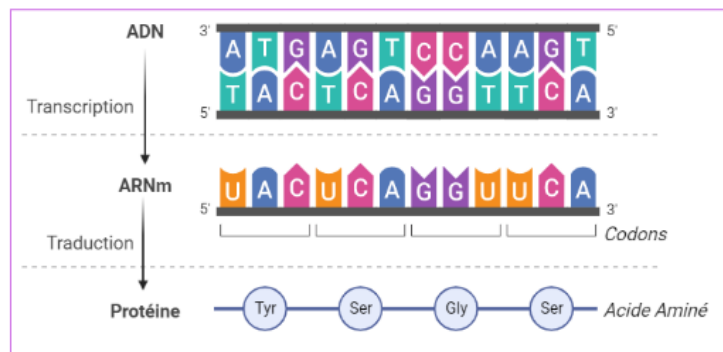


Figure 2 - De l'ADN à la protéine

(Tirée de BioRender.com)

1.1.2 Régions intergéniques

Pour les besoins de ce mémoire, nous définissons les régions intergéniques comme des séquences se trouvant entre les gènes codant des protéines, ces régions ne sont donc pas traduites en protéines. Ces séquences dont l'importance fonctionnelle varie, peuvent comprendre des éléments responsables de contrôler la transcription des gènes, tels que : des promoteurs, des opérateurs et des terminateurs. Les régions intergéniques peuvent aussi être transcrites en ARN noncodants (ARNnc) (Tatusova *et al.*, 2016).

1.1.3 Opérons

Les opérons constituent un groupe de gènes qui opèrent sous le contrôle du même promoteur. Dans la base de données *Operon DataBase* (ODB) accessible à partir du lien : <https://operondb.jp/>. On y retrouve les opérons connus de la littérature et les opérons conservés (Okuda & Yoshizawa, 2010).

1.1.4 Promoteurs

Un promoteur est une séquence qui se trouve directement en amont des gènes, où l'ARN polymérase arrive à se fixer pour initier la transcription. Chez les procaryotes, le promoteur comprend deux séquences courtes aux positions -10 et -35 en amont du site de début de transcription (Ayoubi & Van De Yen, 1996) comme on peut le voir dans la figure 3.

- La boîte -35 du promoteur permet de stabiliser la liaison de l'ARN polymérase au promoteur
- La boîte -10 permet à l'ARN polymérase d'identifier le site d'initiation de la transcription

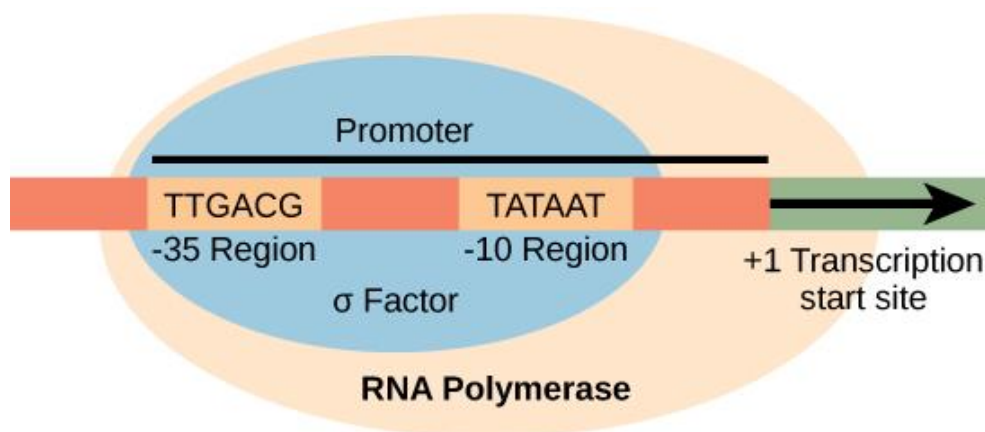


Figure 3 - les boîtes -35 et -10 d'un promoteur en présence de l'ARN polymérase

(La figure est tirée de CNX OpenStax - http://cnx.org/contents/GFy_h8cu@10.53:rZudN6XP@2/Introduction)

1.1.5 Termineur

Le termineur est une séquence qui marque typiquement la fin de la transcription d'un gène ou d'un opéron. Sa transcription produit un ARN qui adopte une structure secondaire en forme d'une épingle à cheveux (ou tige-boucle) qui est un appariement intra-chaîne qui déstabilise l'ARN-polymérase jusqu'à dissociation.

On retrouve deux types de termineurs : les Rho indépendants et les Rho dépendants (Alberts *et al.*, 2008).

- La figure 4 indique la structure d'un termineur Rho indépendant qui prend la forme d'une épingle à cheveux riche en paires de bases G-C, suivie d'une séquence poly-U permettant une libération plus facile de l'ARN polymérase. En effet, la formation d'une structure en tige boucle stable, dans le canal de sortie de l'ARN au sein de l'ARN polymérase, déstabilise l'hybride ADN/ARN déjà affaibli par la forte concentration en paires G-C. (Farnham & Platt, 1981).

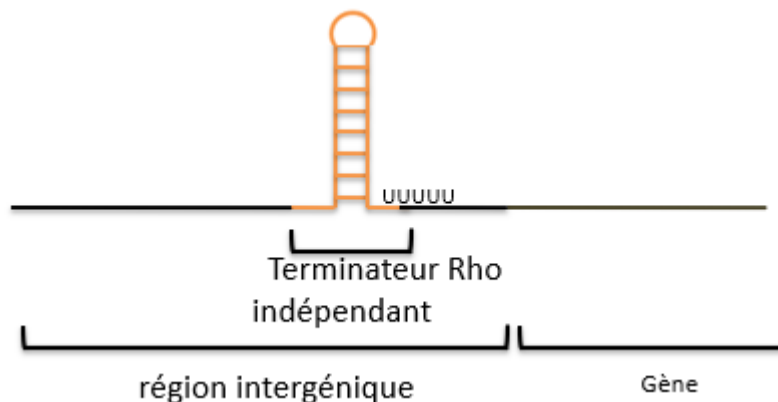


Figure 4 - Structure d'un termineur Rho indépendant

- La figure 5 est un termineur Rho dépendant qui aurait une structure en épingle à cheveux plus courte et qui n'est pas riche en paires de bases G-C et qui est non-suivie d'une séquence poly-U. Il y a donc nécessité du facteur rho qui a une affinité pour les ARN en court de synthèse, le parcourant de 5' vers 3' jusqu'à trouver l'ARN-polymérase. Le facteur rho est ATP dépendante, dont l'hydrolyse permettra la dissociation du complexe(Ciampi, 2006).

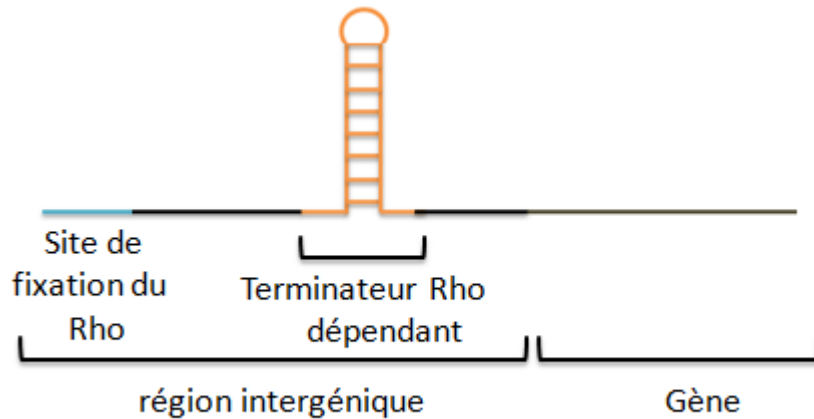


Figure 5 -Structure d'un terminateur Rho dépendant

1.2 ARN noncodants

Les ARNnc sont des molécules transcrites, mais non traduites. Leurs structures secondaires déterminent leur fonction qui est souvent la régulation de gènes (Tran *et al.*, 2009). Il en existe plusieurs types : les petits ARN, les ARN de transfert, les ARN ribosomiaux, les thermorégulateurs, les *riboswitchs*, les *leaders*, etc. Quelques exemples d'intérêt pour ce mémoire sont décrits dans cette section.

1.2.1 Petit ARN (sARN)

Les sARN se lient sur un site d'une séquence d'ARNm spécifique, typiquement pour empêcher le ribosome de s'y associer et inhiber ainsi l'expression du gène, soit en clivant cet ARNm, soit en bloquant physiquement l'accès au ribosome (Argaman *et al.*, 2001). Cette liaison est montrée dans la figure 6.

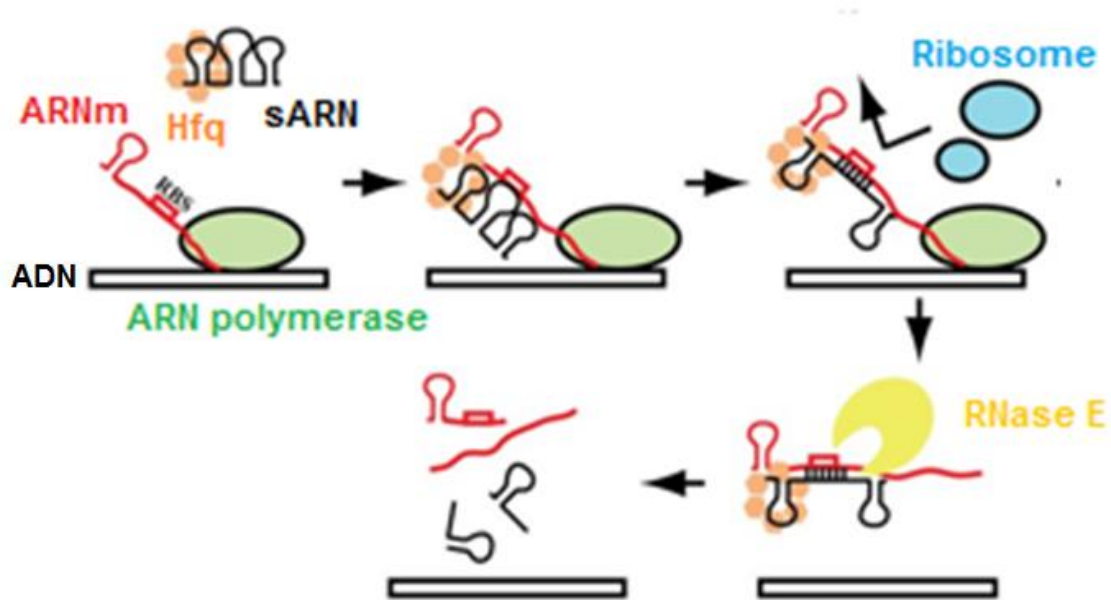


Figure 6 : Régulation génétique par les petits ARN par le mécanisme Hfq-dépendant

Tirée de (De Lay *et al.*, 2013)

Les petits ARN sont de forme simple brin associé ou non à une protéine Hfq. Le complexe ribonucléoprotéique reconnaît son transcrite cible, un ARNm, via une complémentarité partielle avec le petit ARN, ce qui peut bloquer l'accès du ribosome à son site de liaison et/ou favoriser la dégradation de l'ARNm suite au recrutement de la ribonucléase E. D'où la fonction de régulation des petits ARN en empêchant la traduction de la protéine codée par ces ARNm (De Lay *et al.*, 2013; Downward, 2004).

1.2.2 ARN ribosomaux (ARNr)

Les ARNr existent dans les deux sous-unités des ribosomes (petite et grande sous-unités). Ces ARNr combinés avec des protéines agissent en tant que site de synthèse des protéines. Ces structures complexes se déplacent le long de la molécule d'ARNm pendant la traduction pour polymériser graduellement la séquence d'acides aminés constituant les protéines (Noller, 1984).

1.2.3 ARN de transfert (ARNt)

Les ARNt interviennent dans l'étape de traduction, où leur fonctionnement principal est le transfert des acides aminés individuels aux ribosomes trouvant. Le site de fixation de l'acide aminé et

l'anticodon sont représentés dans la figure 7. La séquence des anticodons permet de déterminer l'identité de l'ARNt, c'est-à-dire de l'acide aminé porté par celui-ci.

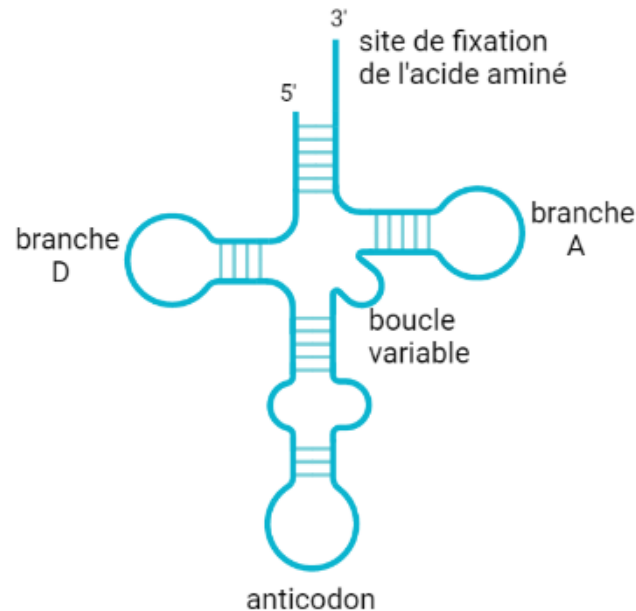


Figure 7 - Structure de l'ARNt

(Tirée de BioRender.com)

1.2.4 Riboswitchs

Un *riboswitch* est un ARNnc régulateur présent dans la région intergénique sur la partie en amont du gène qu'il régule (Breaker, 2011). Il comporte deux parties. La partie aptamère peut se lier directement à un ligand qui est une petite molécule (métabolite ou ion inorganique) déclenchant ainsi une modification de structure de la partie « plateforme d'expression ». La modification de structure de la plateforme d'expression déclenche soit la formation d'un terminateur/anti-terminateur ce qui influence l'expression du gène se trouvant en aval en inhibant/activant la transcription de la région codante, soit la séquestration/libération de la séquence de liaison au ribosome et donc la traduction de la protéine codée par ce gène. D'autres mécanismes de régulation existent, mais ce sont les principaux. La figure 8 résume quelques exemples de régulations effectuées par les *riboswitchs*.

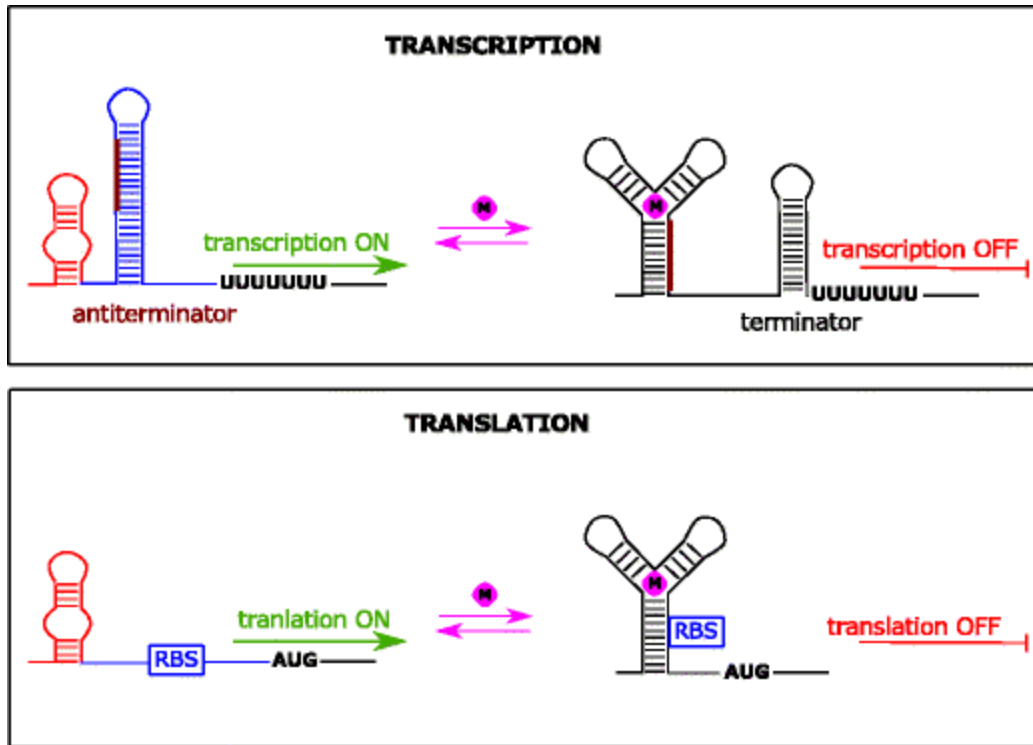


Figure 8 - Mécanismes de régulation des *riboswitchs*.

(Tirée de (Machtel *et al.*, 2016))

« M » représente la molécule(ligand)

« RBS » représente le site de liaison du Ribosome

La séquence en « brun » représente une séquence commune au module aptamère et au module anti-terminateur.

La partie en bleu à gauche représente la plateforme d'expression.

Cela dit on distingue différents types de *riboswitchs* répartis en fonction de la nature de leur ligand. D'une part les *riboswitchs* typiques (ex. TPP - Thiamine Pyrophosphate), connus pour entrer dans l'une des deux voies de repliement mutuellement exclusives pour conférer la régulation. Ces replis d'aptamères déclenchent des signaux structuraux dans la plateforme d'expression adjacente, qui, à leur tour, transduisent un signal pour l'expression des gènes « on » ou « off » (Haller *et al.*, 2013). Cependant, on retrouve également des cas où plusieurs classes de *riboswitchs* reconnaissent le même ligand, notamment les *riboswitchs* SAM (*S-Adenosyl Methionine*). Ils existent au moins cinq classes connues de *riboswitchs* SAM qui se distinguent les unes des autres par leurs caractéristiques architecturales. Par exemple, la classe SAM-I forme une jonction hélicoïdale à quatre voies, SAM-II forme un pseudo-nœud classique (de type H) et SAM-III est défini par une jonction à trois voies (Batey, 2011). La famille d'ARN SAM s'associe

au même ligand pour réguler les mêmes gènes malgré la diversité des classes. Il existe également plusieurs classes de *riboswitchs* associés au second messager di-GMP-cyclique qui régulent un grand nombre de gènes (Sudarsan *et al.*, 2008).

1.2.4.1 Le riboswitch « ykkC-yxkD » connu comme *riboswitch* associé à ppGpp

À l'inverse des différentes classes de *riboswitchs* qui interagissent avec la molécule SAM, il existe quelques cas de classes de *riboswitchs* dont différentes sous-classes interagissent avec différents métabolites. Le *riboswitch* ykkC-yxkD, a été découvert par une étude de génomique comparative en 2004 (Barrick *et al.*, 2004) mais est longtemps resté « orphelin », c'est-à-dire sans ligand connu (Meyer *et al.*, 2011). Ceci était en partie dû au fait que les structures d'ARN ykkC-yxkD étaient trouvées en amont d'une grande diversité de gènes, dont les fonctions étaient assez disparates et ce n'est qu'en 2017 qu'une sous-classe de ses *riboswitchs* a été établie comme interagissant spécifiquement avec la guanidine (Nelson *et al.*, 2017). Ceci laissait la question ouverte quant à la fonction des autres exemplaires de ces *riboswitchs*, question qui a trouvé une réponse en 2018 lorsqu'une autre sous-classe a été définie comme étant le *riboswitch* associé à l'alarmone ppGpp par des études de génomique comparative avec différents outils bio-informatiques, en autres: Infernal v1.1 (Sherlock *et al.*, 2018). Cette sous-classe de *riboswitch* est un élément conservé de l'ARN qui existe en amont de séquences codant pour des protéines qui permettent la résistance à la carence d'acides aminés (figure 9), ce qu'on peut appeler la réponse stringente (Reiss *et al.*, 2017). Ce *riboswitch* constitue ainsi un des rares exemples dont le ligand n'est pas impliqué dans le métabolisme primaire, mais agissant plutôt comme second messager, lui donnant un rôle très important pour la régulation de processus importants chez les bactéries.

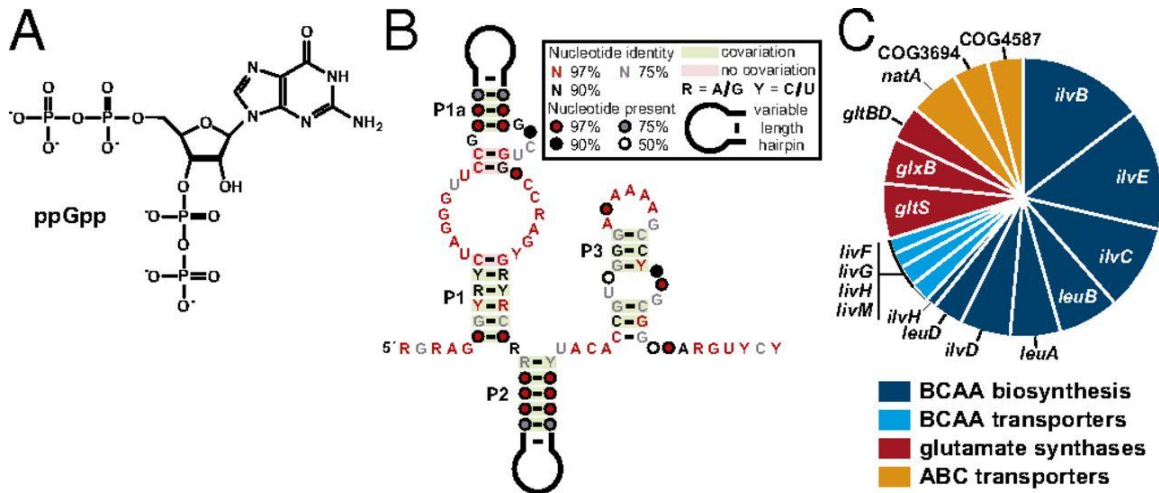


Figure 9 - Le *riboswitch* *ykkC-ykkD*.

(A) La structure chimique de ppGpp. (B) La structure secondaire du *riboswitch* *ykkC-ykkD*, plus précisément de la sous-classe qui interagit avec ppGpp. (C) les gènes prédits associés à cette sous-classe du *riboswitch* *ykkC-ykkD*. (Copyright © 2021 National Academy of Sciences - CC BY-NC-ND license).

Puisque nous porterons un intérêt particulier à la régulation médiée par ppGpp, quelques informations de plus s'imposent. Les cellules procaryotes subissent des conditions environnementales qui limitent leur croissance, où il arrive qu'il y ait un manque d'acides aminés et d'autres nutriments causant une surreprésentation d'ARNt non chargés (ne portant aucun acide aminé), ce qui engendre l'arrêt du ribosome. La cellule réagit par la conversion de GTP en ppGpp avec des enzymes codées par les gènes RelA et SpoT et ce en utilisant de l'ATP. Par conséquent, le ppGpp intervient comme une alarmone qui déclenche la régulation et l'activation de la transcription des gènes impliqués dans la résistance à la carence en nutriment (Dalebroux & Swanson, 2012). La figure 10 résume le processus de la réponse stringente.

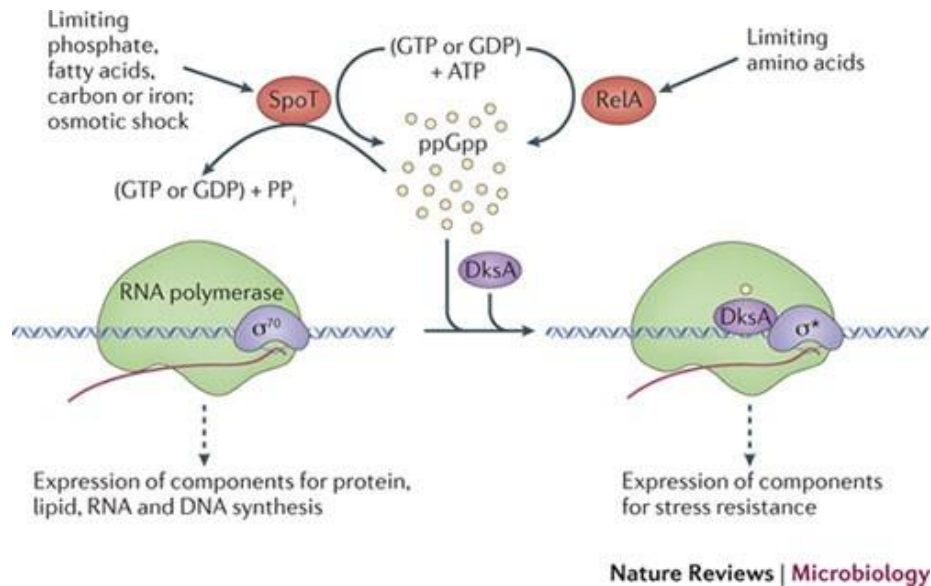


Figure 10 : Mécanisme d'action de ppGpp.

(Dalebroux & Swanson, 2012)

En réponse à des conditions particulière, SpoT et RelA catalyse l'ATP pour synthétiser ppGpp. Avec le suppresseur DksA, ppGpp dirige l'initiation de la transcription au niveau de promoteurs de gènes particuliers via une interaction directe avec l'ARN polymérase, en favorisant l'interaction de l'ARN polymérase avec d'autres facteurs σ (σ^*). Lorsque les précurseurs métaboliques sont abondants, SpoT ppGpp.

1.2.4.2 Recherche de *riboswitchs*

Lorsque des *riboswitchs* régulent plusieurs gènes différents, la fonction de ces gènes est associée d'une façon ou d'une autre au ligand. Pour découvrir ainsi de nouveaux *riboswitchs* on peut se concentrer sur les régions intergéniques (IGR) se trouvant en amont des gènes associés à un ligand donné. Comme cela a été fait dans plusieurs études de génomiques comparatives, avec des IGR associés au même gène (Weinberg et al. 2007 ; Weinberg et al. 2017).

1.3 La base de données « RiboGap »

L'abondance et la complexité de données génomiques imposent un délicat travail d'organisation, de synthèse et de hiérarchisation. Ceci a motivé la réalisation d'une base de données, « RiboGap », pouvant classifier diverses données de façon à être accessibles et utilisables par tout biologiste n'ayant pas de compétences en informatique.

RiboGap est une base de données relationnelle basée sur le langage MySQL qui permet de trouver les séquences intergéniques (IGR), les séquences codantes (cds), les ARNnc et les terminateurs dans l'ensemble du génome de différents organismes procaryotes (Naghdi *et al.*, 2017). Elle avait d'abord été développée, entre autres, pour des études de génomiques comparatives servant notamment à trouver de nouveaux ARNnc. La figure 11 montre le schéma simplifié du diagramme complet qui se trouve dans l'annexe 1.

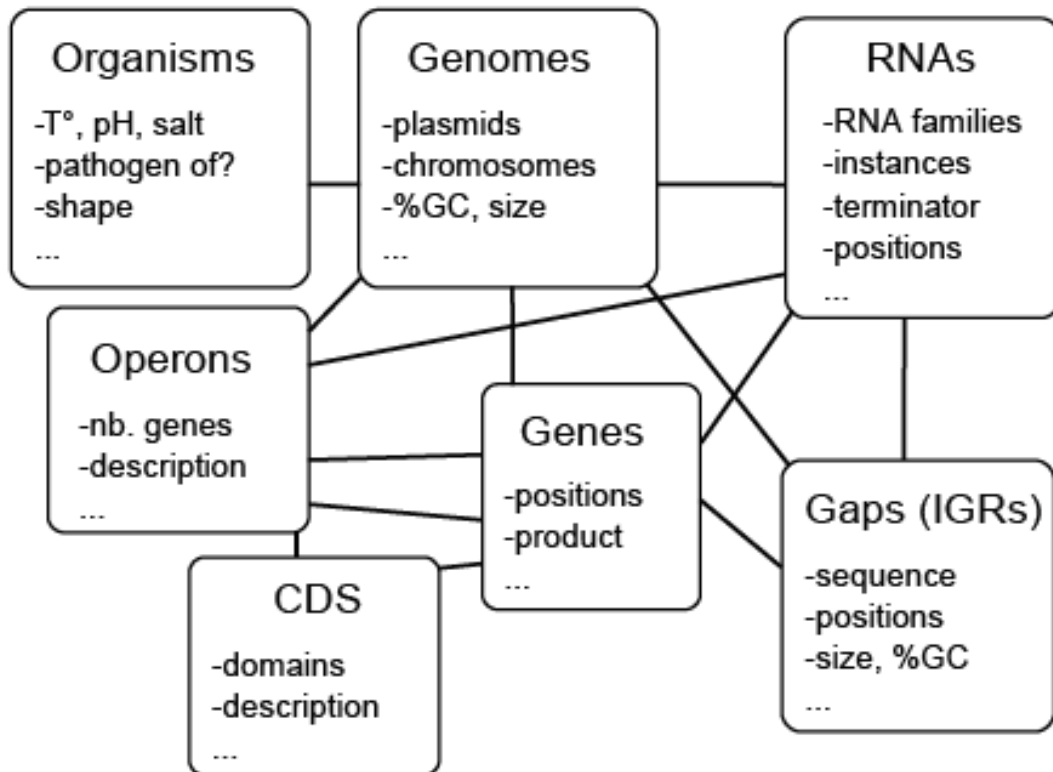


Figure 11 : Schéma simplifié de la base de données de RiboGap-version1.

Tirée de (Naghdi *et al.*, 2017)

La modélisation de RiboGap est réalisée par UML (Unified Modeling language) et les associations entre les tables sont conçues de telle sorte que chaque organisme contient un génome (chromosome et parfois plasmides). Dans ce génome, on retrouve des gènes (séquences codantes - CDS) codants pour des protéines qui sont constituées de domaines conservés. En même temps, ces gènes sont liés à des régions intergéniques du côté 5' et 3' respectivement appelées *gap5* et *gap3*. Finalement, ces régions intergéniques sont souvent transcrites en ARNnc qui peuvent être des petits ARN, des *riboswitchs*, etc. La description de chaque table constituant la base de données est indiquée dans le tableau 1.

Tableau 1 - Description des différentes tables de RiboGapv1

Nom de la table	Description	Source des données
Séquences codantes-<i>CDS</i>	Gènes codants pour des protéines	NCBI (Genbank)
Domaines conservés- <i>CDD</i>	Domaines conservés des protéines	NCBI (base de données CDD)
Région intergéniques - <i>Gap</i>	Régions intergéniques en 3' et en 5' des gènes	NCBI (Genbank)
Génomes-<i>fragment</i>	Les chromosomes et les plasmides des procaryotes	NCBI (liste des chromosomes et des plasmides procaryotes)
Familles d'ARN-<i>RNA_family</i>	Les familles d'ARNnc qui sont trouvés dans les gap5 et les gap3	Rfam
ARN connus-<i>RNA_known</i>	ARNnc connus et découverts jusqu'à maintenant	Rfam
Organismes - <i>Organisms</i>	Organismes bactériens	NCBI
Opérons - <i>Operons</i>	Liste des gènes qui opèrent sous le même opérons	Operon DataBase

1.3.1 Sources des données

1.3.1.1 Centre National pour les informations biotechnologiques (NCBI)

Cet institut est le centre de l'information de biologie moléculaire des États-Unis, qui développe des bases de données et des logiciels pour traiter des données en liens avec les génomes

eucaryotes et procaryotes et les rendre ensuite accessibles pour tout le public et les scientifiques (Sayers *et al.*, 2010). Il existe plusieurs bases de données dans NCBI, telles que: Refseq (Pruitt *et al.*, 2007), Genbank (Benson *et al.*, 2012), Entrez (Schuler *et al.*, 1996), PubMed (Canese & Weis, 2013), CDD (Marchler-Bauer *et al.*, 2002), etc. Ces différentes bases de données citées sont utilisées comme référence de départ pour extraire les données qui existent dans RiboGap v1.

1.3.1.2 Rfam

Rfam est une base de données qui englobe les informations sur les familles d'ARNnc (Gardner *et al.*, 2009). On y trouve également le package Infernal qui est utilisé pour chercher les ARNnc dans les séquences d'ADN en utilisant un modèle de covariance, considéré comme le profil des structures secondaires des ARNnc (Nawrocki *et al.*, 2009).

1.3.2 Interface web de RiboGap

Les tables qui composent la base de données sont accessibles via l'interface web hébergée dans le serveur du laboratoire du professeur Jonathan Perreault dont l'adresse web est : « <http://ribogap.iaf.inrs.ca/> ». Elle a été développée avec le langage perl. Ce qui lui permet d'être accessible et utilisable par tout biologiste n'ayant pas de compétences en informatique et ce pour exécuter des requêtes de sélection de données sans avoir de connaissances du langage SQL, même si la base de données est implémentée en MySQL (système de gestion de bases de données). L'utilisateur a la possibilité de sélectionner les données qu'il désire afficher en cochant simplement les cases (Figure 12).

cdd	Conserved domains
<input type="checkbox"/> cdd_id <input type="checkbox"/> cdd_accession <input type="checkbox"/> cdd_name <input type="checkbox"/> description	ex: 116891 example: pfam05377 or cd06578 or TIGR01188... name of conserved domains like HemD nickel ABC transporter, nickel/metallophore periplasmic binding protein
cds	Coding sequence
<input type="checkbox"/> accession <input type="checkbox"/> gene <input type="checkbox"/> DNA_seq <input type="checkbox"/> locus_tag <input type="checkbox"/> product	protein accession like NP_038276.1 gene name like rhlA DNA sequence of gene (No information available for now!!!) ex:Marme_0002 ex:Mg transporter
gap3	sequence information for 3-prime-UTR
<input type="checkbox"/> start <input type="checkbox"/> end <input type="checkbox"/> strand <input type="checkbox"/> sequence <input type="checkbox"/> size	start position of 3 prime-UTR end position of 3 prime-UTR strand direction of the corresponding gene sequence of 3 prime-UTR in the same strand as the gene size of 3 prime-UTR
rna_family	Family of RNA according to Rfam
<input type="checkbox"/> fam_id <input type="checkbox"/> fam_name <input type="checkbox"/> description <input type="checkbox"/> type <input type="checkbox"/> note	Rfam accession: RF00001 5S_rRNA 5S ribosomal RNA gene; rRNA some description
rna_known	Known RNA according to Rfam
<input type="checkbox"/> start <input type="checkbox"/> end <input type="checkbox"/> strand	start position of RNA end position of RNA strand of RNA

Figure 12 - Tables de l'interface web de RiboGap v1

L'utilisateur peut également restreindre les résultats à afficher, en choisissant les conditions qu'il souhaite (Figure 13).

Condition:

-	▼	-	▼		-	▼
-	▼	-	▼		-	▼
-	▼	-	▼		-	▼
-	▼	-	▼		-	▼
-	▼	-	▼		-	▼
-	▼	-	▼		-	▼
-	▼	-	▼		-	▼

Figure 13: Section des conditions émises par l'utilisateur

1.3.3 Utilisation de RiboGap

L'une des utilisations de RiboGap est d'obtenir les séquences des régions intergéniques qui se trouvent devant un ou des gènes choisis pour la découverte de nouvelles structures d'ARNnc qui réguleraient un ensemble de gènes ayant des fonctions apparentées. À cet effet, deux colonnes sont sélectionnées de la table CDS : *product* et *gene*, ce qui permet l'usage de mots-clés en lien avec la fonction de la protéine encodée par ce gène. De plus, on sélectionne les champs : *fragment* d'ADN et de *description* pour obtenir les numéros d'accèsion des plasmides/chromosomes et les noms de souches bactériennes. Enfin, on sélectionne les séquences IGR dans la table *gap5*, avec tous les champs de *gap5*. La recherche doit être ensuite affinée dans la section des conditions en utilisant soit « *REGEXP* » soit « *find some pattern* » de n'importe quel mot-clé pour les champs *product* ou *gene* de la table *cds*. Dans ce cas-ci les mots clés « *methyl* » et « *RNA* » sont utilisés pour trouver des gènes ayant des fonctions liées à la méthylation de l'ARN, tels que les gènes codant pour les « ARNt méthylases » ou « 16S rARN méthyl transférase » comme le montre la figure ci-dessous.

cds	Coding sequence
<input type="checkbox"/> accession	protein accession like NP_038276.1
<input checked="" type="checkbox"/> gene	gene name like rhlA
<input type="checkbox"/> locus_tag	ex:Marme_0002
<input checked="" type="checkbox"/> product	ex:Mg transporter
fragment	Chromosome information
<input checked="" type="checkbox"/> DNA fragment	Refseq accession number like NC_000913
<input checked="" type="checkbox"/> description	Staphylococcus aureus subsp. aureus str. Newman
gap5	Sequence information for 5-prime-UTR
<input checked="" type="checkbox"/> start	start position of 5 prime-UTR
<input checked="" type="checkbox"/> end	end position of 5 prime-UTR
<input checked="" type="checkbox"/> strand	strand direction of the corresponding gene
<input checked="" type="checkbox"/> sequence	sequence of 5 prime-UTR in same strand as the gene
<input checked="" type="checkbox"/> size	size of 5 prime-UTR

product	▼ find some pattern ▼	methyl	AND ▼
product	▼ find some pattern ▼	RNA	AND ▼
size	▼ >= ▼	25	- ▼
-	▼ - ▼		- ▼
-	▼ - ▼		- ▼
-	▼ - ▼		- ▼
-	▼ - ▼		- ▼

Pour rappel, un *riboswitch* donné régule des gènes dont la fonction est associée à son ligand, par exemple, les gènes impliqués dans la voie de synthèse du ligand en question. Donc avec des mots-clés de noms des gènes ou de protéines (*product*) faisant référence à ce ligand, on peut regrouper les régions intergéniques associées au ligand et qui pourraient éventuellement contenir un *riboswitch* qui régulerait ce gène.

2 PROBLEMATIQUE

La découverte de nouvelles structures d'ARNnc en bio-informatique repose sur des calculs de comparaison de grande ampleur, donc une approche ciblée (justifiable par le fait que les éléments cis-régulateurs tels les *riboswitchs* seront évidemment en amont de gènes qui sont fonctionnellement reliés) pourrait aider à résoudre cela. D'où l'importance d'utiliser RiboGap comme étape primaire pour recenser et extraire toutes les régions intergéniques se trouvant en amont d'un gène précis. L'inconvénient est que RiboGap ne comporte que les génomes complets, sachant que la diversité de séquences est clé pour trouver de nouveaux ARNnc et que l'on peut trouver d'intéressant candidats dans les génomes incomplets. De plus, depuis la réalisation de RiboGap, des mises à jour ont été effectuées sur NCBI et Rfam. Ce qui fait que RiboGap ne comporte pas toutes les données pour toutes les bactéries. Nous avons également remarqué que certains éléments régulateurs, tels que des promoteurs, pouvaient causer des problèmes lors des prédictions car leur conservation augmente la probabilité d'avoir de faux positifs pour la recherche de structures d'ARN conservées. Ainsi, de multiples ajouts, de même qu'une mise-à-jour majeure, semblent essentiels pour un usage optimal de RiboGap pour de telles études de génomique comparative.

3 HYPOTHESE

Dans notre cas, le ligand sur lequel on a concentré les recherches est ppGpp, appelé aussi guanosine tetraphosphate. Son importance et son implication complexe dans le monde de la régulation des expressions des gènes nous laisse croire qu'il doit sûrement y avoir des *riboswitchs*, ou autres types d'ARN régulateurs, associés à ppGpp. Mais pour augmenter nos chances, il faut créer deux nouvelles versions de RiboGap v2 et v2.1 qui correspondront respectivement à une mise à jour des génomes complets et aux génomes incomplets. Nous émettons l'hypothèse que ceci augmentera le nombre de candidats pour la découverte de nouvelles structures d'ARNnc qui pourront être de nouveaux *riboswitchs*. De plus, l'ajout de la table promoteur à RiboGap pourrait aider à assigner les conservations de séquence trouvées à leur rôle approprié, ARN régulateur vs promoteur.

4 OBJECTIFS

1. Créer deux bases de données RiboGap v2 et v2.1
 - 1.1. Extraction des séquences codantes et des régions intergéniques des génomes complets et incomplets
 - 1.2. Annotation des gènes ARN et des éléments régulateurs
 - 1.3. Mise à jour des données
2. Recherche de nouveaux ARNnc
 - 2.1. Extraction des régions intergéniques associés à ppGpp de RiboGap v2 et v2.1
 - 2.2. Recherche de séquences et structures conservées parmi ces séquences

5 RIBOGAP: A RELATIONAL DATABASE FOR PROKARYOTE GENOMICS

RiboGap : une base de données relationnelle pour les procaryotes génomiques

Afin de vérifier notre hypothèse, la première étape du projet consiste à réaliser deux nouvelles versions de la base de données RiboGap. Vue les importants changements de tailles de données et fonctionnalités existants dans ces versions. Nous avons jugé intéressant d'en faire un article qui est en cours de soumission.

Samia Djerroud¹, Abdellatif Elghizi¹, Jonathan Perreault¹

¹ INRS – Centre Armand-Frappier Santé Biotechnologie, 531 boulevard des Prairies, Laval (Québec), H7V 1B7, Canada

5.1 Résumé

Introduction : Les ARN noncodants (ARNnc) sont importants pour la régulation de l'expression des gènes et leur association avec des régions intergéniques a conduit au développement de la base de données RiboGap pour faciliter leur étude.

Description : RiboGap v2 est une base de données relationnelle qui donne accès à des informations concernant les phénotypes d'organismes, les génomes, les gènes, les régions intergéniques, les ARNnc, les opérons et les promoteurs prédits de tous les procaryotes séquencés dans les bases de données publiques. La ressource offre des outils de recherche et d'exploration de données utiles, y compris la possibilité de rechercher des gènes et des ARNnc par nom ou à l'aide de mots-clés de fonction. Les données sources ont été extraites de différentes bases de données et générées à partir de séquences procaryotes avec différents outils bio-informatiques.

Conclusion : De nouvelles bases de données (RiboGap v2 et v2.1) ont été développées pour inclure respectivement tous les génomes complets et incomplets. RiboGap v2 et RiboGap v2.1 sont accessibles à l'adresse <http://ribogap.iaf.inrs.ca>.

Mot clés : Région intergénique, ARN non codant (ARNnc), base de données, génome procaryote, gène, promoteur, terminateur Rho-dépendant, terminateur Rho-indépendant, riboswitch, thermorégulateur, séquence leader, t-box, *cis*-régulateur, virulence, ARNt, ARNr

5.2 Abstract

Background: Noncoding RNAs (ncRNAs) are important for the regulation of gene expression and their association with intergenic regions led to the development of the RiboGap database to facilitate their study.

Description: RiboGap v2 is a relational database that provides access to information with regards to organism phenotypes, genomes, genes, intergenic regions, ncRNAs, operons and predicted promoters of all sequenced prokaryotes in public databases. The resource offers useful searching and data mining tools, including the ability to search for genes and ncRNAs by name or using function keywords. Source data was extracted from different databases and generated from prokaryotic sequences with different bioinformatics tools.

Conclusion: New databases (RiboGap v2 and v2.1) have been developed to include respectively all the complete and incomplete genomes. RiboGap v2 and RiboGap v2.1 can be accessed at <http://ribogap.iaf.inrs.ca>.

Keywords: Intergenic region, noncoding RNA (ncRNA), database, prokaryotic genome, gene, promoter, Rho-dependent terminator, Rho-independent terminator, riboswitch, thermoregulator, leader sequence, t-box, *cis*-regulatory, virulence, tRNA, rRNA

5.3 Introduction

The quantity of data represented by genome sequences from different organisms, structural and functional annotations of genes, noncoding RNAs (ncRNAs), and the need of storing efficiently this increasing biological data to interrogate it efficiently require well designed biological databases (Herbert *et al.*, 2007). For example: articles are provided by PubMed (Canese & Weis, 2013), sequencing and annotation of genomic data is available on GenBank (Benson *et al.*, 2012) or conserved operons are hosted in Operon database (ODB) (Okuda & Yoshizawa, 2010).

Most biological databases are accessible on websites where users can browse information. In general, it is also possible to download data in various formats: text, CSV (comma-separated values) or fasta files. Each database is managed by a management system (ex: MySQL, SQL server, Oracle, etc.) and schematized by a data model which describes the organization and constitution of the tables and which provides information about the relationships between the different data (Coronel & Morris, 2016). In this case, a database RiboGap (Naghdi *et al.*, 2017) was created by using MySQL (DuBois, 2008) as a relational database management system and Perl (Gundavaram *et al.*, 2000) for conception and the development of the web interface that is available at <http://ribogap.iaf.inrs.ca/> (see the supplementary material section 1 for more information about the first version of RiboGap). The goal of the platform was to give scientists access to different genomic data by combining the intergenic regions and genes downloaded from NCBI (Sayers *et al.*, 2019), ncRNAs from Rfam (Gardner *et al.*, 2009) and operons from the Operon Database.

Genomes are classified according to the assembly level into complete genomes (fully assembled) and Scaffold/Contigs (incomplete genomes that are partially assembled). Each genome is made up of coding sequences (genes encoding proteins) which are interspaced by intergenic regions (Tatusova *et al.*, 2016) (for more details on completely and incompletely assembled genomes see supplementary material section 2.1). Intergenic regions (IGRs) are sequences located between two genes, while untranslated regions (UTRs) can generally be regarded as a subset of the former and are transcribed into a ncRNA (or at least a noncoding region of a mRNA) which is not translated into a protein (Tropp, 2012). NcRNAs are classified into families in Rfam, such as: tRNAs, rRNAs, riboswitches, CRISPR RNAs or small RNAs, to give only a few examples. The first version of RiboGap included only complete genomes (2,757 genomes) and 2,500 RNA families taken from Rfam. Given that there are now 9,857 complete genomes, and much more incomplete genomes in NCBI, combined with numerous new RNA families (3,019 families of RNA in Rfam), an update was required. In this work, two different databases were created: A second version of RiboGap to update the genomic data of complete genome, in addition, of a version 2.1 of RiboGap which contains incomplete genomes.

Different types of transcription's direction: unidirectional, convergent and divergent (Rogozin *et al.*, 2002). Some directions may differ in terms of the types of regulatory sites they contain:

- Sequences between unidirectional genes may contain a terminator for the upstream gene, a promoter and an operator for the downstream gene; or neither if both genes are co-transcribed as a polycistronic mRNA within an operon.
- Sequences between convergent genes would typically only have terminators.
- Sequences between divergent genes are expected to have at least two promoters with binding sites for transcriptional factors.

In order to provide as much information as possible on the intergenic regions, several features were added to the new versions of RiboGap, mainly predictions for: promoters; and Rho-dependent transcription terminators in addition to the Rho-independent terminators.

5.4 Materials and methods

5.4.1 Genomic data download

The National Center for Biotechnology Information (NCBI) is considered a reference tool because the majority of RiboGap's data comes from there: the list of prokaryotes, the Fasta files and the GenBank files of the complete and incomplete genomes. It also provides different databases for biological information, such as: CDD for proteins database (Marchler-Bauer *et al.*, 2002) and Entrez for prokaryotic organism (Schuler *et al.*, 1996). In this work, GenBank files of complete and incomplete assembled genomes were downloaded from NCBI (details in supplementary material 2.1). Perl programs were then used to extract coding sequences (CDS) and IGRs of upstream and downstream of genes for all complete and incomplete genomes (supplementary material section 2.2 and 2.3).

5.4.2 Noncoding RNA detection

NcRNAs can perform various functions usually determined by the secondary structure that RNA adopts. Those that do not fold in a specific way most often combine with other molecules, usually proteins, to form complexes (Mattick, 2005). Rfam is the most used platform to get all information about the known ncRNAs and their families. The classification of RNAs into families is performed based on the multiple sequence alignments (MSA) that have evolved from a common ancestor. This can also provide insight into their consensus secondary structure, function and covariance models (Childs *et al.*, 2009). The Rfam database is available on the web (Downard, 2004) and it can be used to fetch information for each ncRNA family. Most of the database is populated by

structured RNA instances found by homology searches using the Infernal software suite (Nawrocki & Eddy, 2013). We also used this package to annotate genomic sequences based on the covariance models of ncRNA families available in Rfam. In RiboGap v1, the prediction of ncRNAs was based on Rfam sequences and carried out by BLAST (Johnson *et al.*, 2008). However, this method generated few hits with aberrant similarity due to an insertion. Even if the e-value of the blast was very good, the e-value that corresponds to the same hit in infernal was very bad. (see Fig.S4 in the supplementary data). So, to avoid these false positives in the new version of RiboGap (v2), Infernal has been applied on the complete and incomplete genomes with covariance models of known ncRNA families to look for homologs. Besides, a new attribute "e-value" was added to the "Known RNA" table of the database to measure the level of correspondence between the intergenic regions and the ncRNAs. The detailed pipeline of the execution of Infernal on prokaryotic genomes is shown in the supplementary material section 2.3.1.1.

5.4.3 tRNA and rRNA determination

To get transfer RNA (tRNA) from genomic sequences, tRNAscan-SE is the adopted tool. The software uses essentially Infernal v1.1 and covariance models that give the secondary structure of tRNA. The program gives an exhaustive list of tRNAs and more details about their types (including pseudo-tRNAs) instead of extracting them directly from GenBank files (tRNA annotated by NCBI) or from Rfam (Chan & Lowe, 2019). The full version of tRNAscan-SE is available as a web server (Lowe TM). In this work, the program has been downloaded and executed by command lines of Linux (Sobell, 2013) in order to use it on all the complete and incomplete genomes. The different steps are presented in 2.3.1.2 of the supplementary material.

In RiboGap v2, the rRNAs were extracted from the GenBank files of the genomes. The annotation is already done by NCBI (as shown in the Fig.S6). This new rRNAs list is more exhaustive than in RiboGap v1.

5.4.4 Promoter predictions

Different bioinformatic tools for prediction of promoters in prokaryotes have been tested in this work, such as: BPROM (Salamov & Solovyevand, 2011), bTSSfinder (Shahmuradov *et al.*, 2017) and BacPP (e Silva *et al.*, 2011). However, according to the work of Shahmuradov and co-workers, bTSSfinder achieved higher accuracy compared to other tools and it was the only one that predicts

putative promoters for five classes of σ factors (σ_{70} , σ_{38} , σ_{32} , σ_{28} and σ_{24}) while the other programs target σ_{70} promoters only. More importantly, bTSSfinder demonstrated better usability for large scale predictions on all IGRs of complete and incomplete genomes. It is available as a standalone program and online (Marchler-Bauer *et al.*, 2002) (its use is explained in supplementary data at 2.3.2). The Fig.S7 shows how bTSSfinder predicts promoters by looking for the motifs of boxes -35 and -10. To predict promoters for all intergenic sequences for both RiboGap v2 and RiboGap v2.1, we first extracted all the intergenic regions that we put into individual files. Then, we performed bTSSfinder analyses in parallel on the different files.

5.4.5 Terminator annotations

As in RiboGap v1, to predict Rho-independent terminators (RITs), a probabilistic method is used: RNIE (Gardner *et al.*, 2011). The program is based on covariance models of RITs, since the secondary structure of this type of terminators is already known: a conserved simple hairpin structure followed by several uridines (U). RNIE is available and can be downloaded (Nawrocki *et al.*, 2009). The program has been executed on both complete and incomplete genomes (see the section 2.3.3.1 in the supplementary material to see how RNIE was executed).

To annotate terminator sequences more thoroughly, we also looked for Rho dependent terminators (RDTs). For this, we used RhoTermPredict to predict RDTs in bacterial complete and incomplete genomes (supplementary material 2.3.3.2 for details). The program RhoTermPredict was the first choice because it is a powerful tool usable for whole genome analysis. It is based on the search of a 78 nt long RUT (Rho utilization site) followed by a putative pause site for RNA polymerase in sequences which exhibit a secondary structure of hairpin. Numerous RUT sites and RNA polymerase pause sites are already known for many genomes (Di Salvo *et al.*, 2019). The RhoTermPredict algorithm, available on line (Sherlock *et al.*, 2018), performs predictions based on a model derived from these experimental data. RDT annotation is an important new feature for the RiboGap database, since transcription termination triggered by the Rho protein factor is not only involved in mRNA termination, but also in the regulation of gene expression (Nadiras *et al.*, 2018).

5.4.6 Functional gene annotation

A functional annotation of genes, with protein domains and families, already exists in the NCBI GenBank files. However, often the gene homology threshold is insufficient to assign it a protein and is thus assigned as a “hypothetical protein”. To have a more thorough annotation, "InterProScan" (Rivas *et al.*, 2017) was used on all predicted coding sequences to annotate genes. The program uses a set of databases Pfam (Bateman *et al.*, 2004), PROSITE (Hulo *et al.*, 2006), PRINTS (Attwood *et al.*, 2012), ProDom (Bru *et al.*, 2005), SMART (Letunic *et al.*, 2002), TIGRFAMs (Haft *et al.*, 2012), PIRSF (Nikolskaya *et al.*, 2006), SUPERFAMILY (Pandurangan *et al.*, 2019), Gene3D (Lewis *et al.*, 2018), and PANTHER (Mi *et al.*, 2007) to annotate protein sequences given as input and to find additional information on protein families and domains (Jones *et al.*, 2014).

A downloadable version exists in order to be able to run the program easily in command line on all complete and incomplete genomes (section 2.4 of the supplementary material).

5.5 Results and discussion

5.5.1 Data in RiboGap v2

The database’s diagram now includes a table for predicted promoters, as well as several new attributes for some tables (addition of the “virulence” attribute in the coding sequence table; modification of the attributes of the CDD table; and addition of e-values). In Fig.1, we can see the organization and the associations of tables in a simplified diagram (detailed diagram in Fig. S10 of the supplementary data). By comparing the data between RiboGap v1, RiboGap v2 and v2.1, we notice a considerable increase in the volume of genomic data (Table 1).

Furthermore, in addition to numerous new RNA families derived from the updated version of Rfam, the RNA tables now include RDTs; and not only RITs like in the previous instalment of RiboGap. This allows RiboGap to cover most known RNA elements. Among the existing RNA families, the modification of tRNA annotations, by using directly tRNAscan-SE instead of NCBI’s annotations, allow us to provide additional information: the amino acid corresponding to that tRNA, as well as identification of “pseudo tRNAs”. Finally, one of the major upgrades to RiboGap v2 is the addition of a promoter table. With predictions of putative promoters for all existing IGRs

in the database. However, Promoter prediction should be used with caution as there are several limitations.

5.5.2 Use of RiboGap v2

The database can be used by many scientists to answer various biological questions. We think the database can be especially useful to quickly get preliminary insight for various hypotheses and determine a course of action with regards to more time-consuming analysis or experiments. Here is the description of some examples of such questions and the corresponding findings.

5.5.2.1 Identification of multiple regulatory elements.

Riboswitches belong to ncRNA families and over 30 different riboswitch families exist in RiboGap v2. Besides well known transcription factors, promoters and terminators, the conserved RNA structures of riboswitches regulate gene expression after binding to a given ligand (Garst *et al.*, 2011). Using a MySQL query in RiboGap v2, we can get the intergenic sequences that may have simultaneously a predicted putative promoter, a riboswitch and a terminator (query details in supplementary material, section 4.1) as schematized in Fig.2 (with complete results in supplementary Table S9).

5.5.2.2 Search for expression platforms of all riboswitches

With a few exceptions, the covariance models of riboswitches only describe the aptamer domain, responsible for metabolite binding, and as such the so-called “riboswitches” annotated in Rfam and in RiboGap (including v2) exclude the expression platform. However, typical expression platforms (transcription terminators/anti-terminators and Shine-Dalgarno/sequestering sequences) can be at least partially deduced with additional information from RiboGap.

To search for these expression platforms, multiple MySQL queries were executed using RiboGap to select intergenic sequences that contain riboswitches and other upstream elements (RITs/RDTs or ncRNAs) to get the distance between them. According to IGR's content we can distinguish three cases: riboswitch and terminators (RITs or RDTs), riboswitch alone or riboswitch and ncRNA (illustrated respectively in Fig. 3A, B and C). The aim of the queries was to obtain for each case the position of the riboswitches and the element which follows it which can be either a terminator,

a ncRNA or none of them. This data was then used individually as if to calculate several combinations of distances by running Perl programs (the complete MySQL queries and programs can be found in section 4.2).

At this step, we calculated for each set of distances the 1st quartile, the minimum, the median, the maximum and the 3rd quartile which allowed us to draw the boxplot for each case (Fig. 3D).

5.5.2.3 Search for putative small RNA signature

Small RNAs are short (generally <200 bases) ncRNA molecules that can regulate the expression of target genes by binding to mRNAs (Shimoni *et al.*, 2007). In many bioinformatic approaches, the prediction of new sRNAs is based on the existence of DNA sequences containing promoters within a short distance of rho independent terminators (Chen *et al.*, 2002). We looked thus for IGRs with high likelihood of small RNA occurrences based on that approach. However, for reduced false positives, we looked for RITs included between two promoters (Fig. 4). These were searched for in RiboGap by performing a MySQL query (supplementary material 4.3). We can find 227,810 such putative sRNA signatures (Table S11), which were compared to the 319,288 known sRNAs from RiboGap v2 - Complete genomes. We found that close to 1.47% of hits from this simple query correspond to known sRNAs. The 98.53% left (~224,000 from the 227,810 hits) could either correspond to not yet discovered sRNAs, to genes with multiple promoters or mis annotations (either missing CDS or false promoter predictions).

5.5.2.4 Motif finder

Regex is based on pattern matching and is an easy way to search for simple (as well as not so simple) sequence motifs. In RiboGap, it can conveniently be used to find instances of motifs in large numbers of sequences, whether IGRs, coding sequences, or amino acid sequences. This can even be done in queries combining different motifs and other search criteria. In this example, we have searched for potential G-quadruplexes.

G-quadruplex structures (G4) are formed in sequences rich in guanine. They have an architecture containing guanine tetrads with four parallel (and/or anti-parallel) strands. They often occur naturally near the ends of the chromosomes of eukaryotes and in transcriptional regulatory regions of multiple genes (Lipps & Rhodes, 2009). Some of these structures are involved in mechanisms

regulating different biological pathways such as replication, transcription and translation (Rhodes & Lipps, 2015), as well as pri-miRNA processing (Rouleau *et al.*, 2018), miRNA binding (Rouleau *et al.*, 2017), poly-adenylation (Beaudoin & Perreault, 2013) and numerous instances in the human transcriptome (Vannutelli *et al.*, 2020). Several bioinformatic tools already exist to predict G4s in nucleic sequences such as G4RNA (Garant *et al.*, 2018).

As an example of a query using Regex in RiboGap, a search for G4motifs in all coding sequences was performed with this pattern $G_4N_{1-7}G_4N_{1-7}G_4N_{1-7}G_{4+}$, where N is any nucleotides (supplementary material section 4.4). Even if additional criteria can be checked with specialized tools to have more accurate results, RiboGap is a good starting point for those who want to go further in their research whether in G-quadruplex or in any other pattern/box because such pattern matching allows a rapid screening of potentially interesting candidates. The complete list of 6,626 putative G4 motif scan be found in Table S13, and some examples are given in Table S14.

5.5.2.5 Cis-regulatory RNAs may require more often Shine-Dalgarno (SD) sequences that fit the consensus

RiboGap is focused on genes and their surrounding sequences, as well as most types of known regulatory sequences, allowing queries directly inquiring numerous hypotheses related to gene control. For instance, to evaluate whether there might be a general pattern with regards to translation initiation and known cis-acting RNA families, queries can be formulated with the following concepts: i) calculate the number of hits of gap5 regions with cis-regulatory RNAs (such as riboswitches, thermoregulators, T-boxes or leader types) and ii) look for association with a perfect SD consensus. This was addressed by evaluating the presence or absence of the sequence AGGAGG in the range expected for the SD, between 5 and 12 bases (supplementary material section 4.5 for more details), iii) while also taking in consideration different start codons.

Results indicate that cis-regulatory elements (at least these four major examples) more often co-occur with a perfect SD consensus (AGGAGG) at the expected position for the SD (between 5 and 12 bases, inclusively): from 0.076 to 0.323 compared to 0.036 for IGRs overall (Table 3). However, among cis-regulatory RNAs thermoregulators less frequently have a "perfect" Shine-Dalgarno, 0.076 compared with 0.111, 0.193 and 0.323 for leaders, riboswitches and T-boxes, respectively. This is even more true when the annotated start codon is a GTG (in this case the ratio

is 0.011, vs 0.039 for overall IGRs and >0.1 for other IGRs with cis-regulatory elements). Perhaps this is related to the subtle balance between the ON and OFF structures which depends essentially on the T_m of the structures and which the ribosome could easily melt if it could hybridize too easily. Obviously, much more work is required to decipher these mechanisms. Nevertheless, a few relatively simple RiboGap v2 queries allowed to rapidly survey most of the available genomic data to get a global picture and provide new hypotheses to examine.

5.6 Conclusion

The RiboGap database has been developed to facilitate biologists' research, with a focus on regulatory regions. By providing a powerful, yet easy to use, relational database framework to permit simple and complex queries combining diverse types of genomic information based on intergenic regions, putative promoter predictions, riboswitches, ncRNAs, terminator predictions or gene functions, to name only a few. Scientists may use RiboGap without having any knowledge in computer science or SQL language, even if we do allow users knowledgeable in SQL to formulate their own queries to further increase the power of RiboGap. The results mentioned are given just as examples to demonstrate the usefulness of the database, and perhaps as starting points for projects relevant to researchers in their respective fields.

In the future, we envision RiboGap metagenomic versions to include separate databases for the various metagenomes being published and which hide an immense wealth of data for comparative genomics. In this case, we will have to include a first step which is the structural annotation of the metagenomes to generate the GenBank files and thus find the coding sequences and the intergenic regions from the fasta files. Even if the tables related to organisms and genomes cannot be generated from these metagenomes, because we do not know the origin of metagenomes, the benefit of making queries on metagenomes in a manner similar to what is described in this manuscript would be very useful.

5.7 Abbreviation

Abbreviation	Explanations
RiboGap v2	RiboGap version 2 (complete genomes)
RiboGap v2.1	RiboGap version 2 (incomplete genomes)
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
tRNA	Transfer RNA
sRNA	Small RNA
rRNA	Ribosomal RNA
ncRNA	Noncoding RNA
CDS	Coding sequence
CDD	Conserved Domains
IGR/GAP	Intergenic sequences
NCBI	National Center for Biotechnology Information
BLAST	Basic Local Alignment Search Tool
ODB	Operon Database
SQL	Structured Query Language
CSV	Comma-separated values
SD	Shine- Dalgarno
RDT	Rho-Dependent Terminator
RIT	Rho-Independent Terminator
G4	G-quadruplex

5.8 Acknowledgements

We thank Calcul Quebec and Calcul Canada for access to *clusters* essential to perform several tasks with parallel file systems and persistent storage; in particular, we thank Hui Zhong Lu and the entire Quebec calculation team for their support. We also wish to thank E Boutet and MR Naghdi for critical reading of the manuscript.

5.9 Funding

This study and open access charge were funded by NSERC [RGPIN-2019-06403]. J.P. is a junior 2 FRQS research scholar.

5.10 Contributions

JP and SD had the ideas for the database and manuscript. SD conceived the methods to fetch and format the data. SD designed and implemented the database. AE prepared some preliminary results and did test queries and verified data. SD wrote the manuscript. All authors revised the manuscript.

5.11 References

1. Herbert KG, Spirollari J, Wang JT, Wang JT, Piel WH, Westbrook J, Barker WC, Hu ZZ, Wu CH: Bioinformatic databases. Wiley Encyclopedia of Computer Science and Engineering 2007.
2. Canese K, Weis S: PubMed: the bibliographic database. In: The NCBI Handbook [Internet] 2nd edition. National Center for Biotechnology Information (US); 2013.
3. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: GenBank. Nucleic acids research 2012, 41(D1):D36-D42.
4. Okuda S, Yoshizawa AC: ODB: a database for operon organizations, 2011 update. Nucleic acids research 2010, 39(suppl_1):D552-D555.
5. Coronel C, Morris S: Database systems: design, implementation, & management: Cengage Learning; 2016.
6. Naghdi MR, Smail K, Wang JX, Wade F, Breaker RR, Perreault J: Search for 5'-leader regulatory RNA structures based on gene annotation aided by the RiboGap database. Methods 2017, 117:3-13.
7. DuBois P: MySQL: Pearson Education; 2008.
8. Gundavaram S, Birznieks G, Guelich S: CGI Programming with Perl: O'reilly; 2000.
9. Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, Connor R, Fiorini N, Funk K, Hefferon T: Database resources of the national center for biotechnology information. Nucleic acids research 2019, 47(Database issue):D23.
10. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR: Rfam: updates to the RNA families database. Nucleic acids research 2009, 37(suppl_1):D136-D140.
11. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J: NCBI prokaryotic genome annotation pipeline. Nucleic acids research 2016, 44(14):6614-6624.
12. Tropp BE: Principles of molecular biology: Jones & Bartlett Publishers; 2012.

13. Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL, Wolf YI, Yin J, Koonin EV: Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic acids research* 2002, 30(19):4264-4271.
14. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH: CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic acids research* 2002, 30(1):281-283.
15. Schuler GD, Epstein JA, Ohkawa H, Kans JA: [10] Entrez: Molecular biology database and retrieval system. *Methods in enzymology* 1996, 266:141-162.
16. Mattick JS: The functional genomics of noncoding RNA. *Science* 2005, 309(5740):1527-1528.
17. Childs L, Nikoloski Z, May P, Walther D: Identification and classification of ncRNA molecules using graph properties. *Nucleic acids research* 2009, 37(9):e66-e66.
18. Downward J: RNA interference. *Bmj* 2004, 328(7450):1245-1248.
19. Nawrocki EP, Eddy SR: Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 2013, 29(22):2933-2935.
20. Johnson M, Zaretskaya I, Raytselis Y, Merezuk Y, McGinnis S, Madden TL: NCBI BLAST: a better web interface. *Nucleic acids research* 2008, 36(suppl_2):W5-W9.
21. Chan PP, Lowe TM: tRNAscan-SE: searching for tRNA genes in genomic sequences. In: *Gene Prediction*. Springer; 2019: 1-14.
22. Lowe TM PP, Chan P. : TRNAscan-SE Search Server.
23. Sobell MG: *A practical guide to Linux commands, editors, and shell programming*: Prentice Hall; 2013.
24. Salamov VSA, Solovyevand A: Automatic annotation of microbial genomes and metagenomic sequences. *Metagenomics and its applications in agriculture, biomedicine and environmental studies* Hauppauge: Nova Science Publishers 2011:61-78.
25. Shahmuradov IA, Mohamad Razali R, Bougouffa S, Radovanovic A, Bajic VB: bTSSfinder: a novel tool for the prediction of promoters in cyanobacteria and Escherichia coli. *Bioinformatics* 2017, 33(3):334-340.
26. e Silva SdA, Echeverrigaray S, Gerhardt GJ: BacPP: bacterial promoter prediction—a tool for accurate sigma-factor specific assignment in enterobacteria. *Journal of theoretical biology* 2011, 287:92-99.
27. Gardner PP, Barquist L, Bateman A, Nawrocki EP, Weinberg Z: RNIE: genome-wide prediction of bacterial intrinsic terminators. *Nucleic acids research* 2011, 39(14):5845-5852.
28. Nawrocki EP, Kolbe DL, Eddy SR: Infernal 1.0: inference of RNA alignments. *Bioinformatics* 2009, 25(10):1335-1337.
29. Di Salvo M, Puccio S, Peano C, Lacour S, Alifano P: RhoTermPredict: an algorithm for predicting Rho-dependent transcription terminators based on Escherichia coli, Bacillus subtilis and Salmonella enterica databases. *BMC bioinformatics* 2019, 20(1):1-11.
30. Sherlock ME, Sudarsan N, Breaker RR: Riboswitches for the alarmone ppGpp expand the collection of RNA-based signaling systems. *Proceedings of the National Academy of Sciences* 2018, 115(23):6052-6057.

31. Nadiras C, Eveno E, Schwartz A, Figueroa-Bossi N, Boudvillain M: A multivariate prediction model for Rho-dependent termination of transcription. *Nucleic acids research* 2018, 46(16):8245-8260.
32. InterPro. <https://www.ebi.ac.uk/interpro/download/>.
33. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL: The Pfam protein families database. *Nucleic acids research* 2004, 32(suppl_1):D138-D141.
34. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ: The PROSITE database. *Nucleic acids research* 2006, 34(suppl_1):D227-D230.
35. Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB, Popov I, Roma-Mateo C, Theodosiou A, Mitchell AL: The PRINTS database: a fine-grained protein sequence annotation and analysis resource—its status in 2012. *Database* 2012, 2012.
36. Bru C, Courcelle E, Carrère S, Beausse Y, Dalmar S, Kahn D: The ProDom database of protein domain families: more emphasis on 3D. *Nucleic acids research* 2005, 33(suppl_1):D212-D215.
37. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P: Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic acids research* 2002, 30(1):242-244.
38. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E: TIGRFAMs and genome properties in 2013. *Nucleic acids research* 2012, 41(D1):D387-D395.
39. Nikolskaya AN, Arighi CN, Huang H, Barker WC, Wu CH: PIRSF family classification system for protein functional and evolutionary analysis. *Evolutionary Bioinformatics* 2006, 2:117693430600200033.
40. Pandurangan AP, Stahlhacke J, Oates ME, Smithers B, Gough J: The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. *Nucleic acids research* 2019, 47(D1):D490-D494.
41. Lewis TE, Sillitoe I, Dawson N, Lam SD, Clarke T, Lee D, Orengo C, Lees J: Gene3D: extensive prediction of globular domains in proteins. *Nucleic acids research* 2018, 46(D1):D435-D439.
42. Mi H, Guo N, Kejariwal A, Thomas PD: PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic acids research* 2007, 35(suppl_1):D247-D252.
43. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G: InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014, 30(9):1236-1240.
44. Smale ST, Kadonaga JT: The RNA polymerase II core promoter. *Annual review of biochemistry* 2003, 72(1):449-479.
45. Garrity GM, Bell JA, Lilburn T: *Proteobacteria phyl. nov. Bergey's Manual of Systematics of Archaea and Bacteria* 2015:1-1.
46. Superson AA, Phelan D, Dekovich A, Battistuzzi FU: Using taxon resampling to identify species with contrasting phylogenetic signals: an empirical example in Terrabacteria. *BioRxiv* 2018:369264.

47. Garst AD, Edwards AL, Batey RT: Riboswitches: structures and mechanisms. *Cold Spring Harbor perspectives in biology* 2011, 3(6):a003533.
48. Shimoni Y, Friedlander G, Hetzroni G, Niv G, Altuvia S, Biham O, Margalit H: Regulation of gene expression by small non-coding RNAs: a quantitative view. *Molecular systems biology* 2007, 3(1):138.
49. Chen S, Lesnik EA, Hall TA, Sampath R, Griffey RH, Ecker DJ, Blyn LB: A bioinformatics based approach to discover small RNA genes in the Escherichia coli genome. *Biosystems* 2002, 65(2-3):157-177.
50. Lipps HJ, Rhodes D: G-quadruplex structures: in vivo evidence and function. *Trends in cell biology* 2009, 19(8):414-422.
51. Rhodes D, Lipps HJ: G-quadruplexes and their regulatory roles in biology. *Nucleic acids research* 2015, 43(18):8627-8637.
52. Rouleau SG, Garant J-M, Bolduc F, Bisailon M, Perreault J-P: G-Quadruplexes influence pri-microRNA processing. *RNA biology* 2018, 15(2):198-206.
53. Rouleau S, Glouzon J-PS, Brumwell A, Bisailon M, Perreault J-P: 3' UTR G-quadruplexes regulate miRNA binding. *Rna* 2017, 23(8):1172-1179.
54. Beaudoin J-D, Perreault J-P: Exploring mRNA 3'-UTR G-quadruplexes: evidence of roles in both alternative polyadenylation and mRNA shortening. *Nucleic acids research* 2013, 41(11):5898-5911.
55. Vannutelli A, Belhamiti S, Garant J-M, Ouangraoua A, Perreault J-P: Where are G-quadruplexes located in the human transcriptome? *NAR Genomics and Bioinformatics* 2020, 2(2):lqaa035.
56. Garant J-M, Perreault J-P & Scott MS (2018) G4RNA screener web server: user focused interface for RNA G-quadruplex prediction. *Biochimie* 151:115-118

5.12 Figures

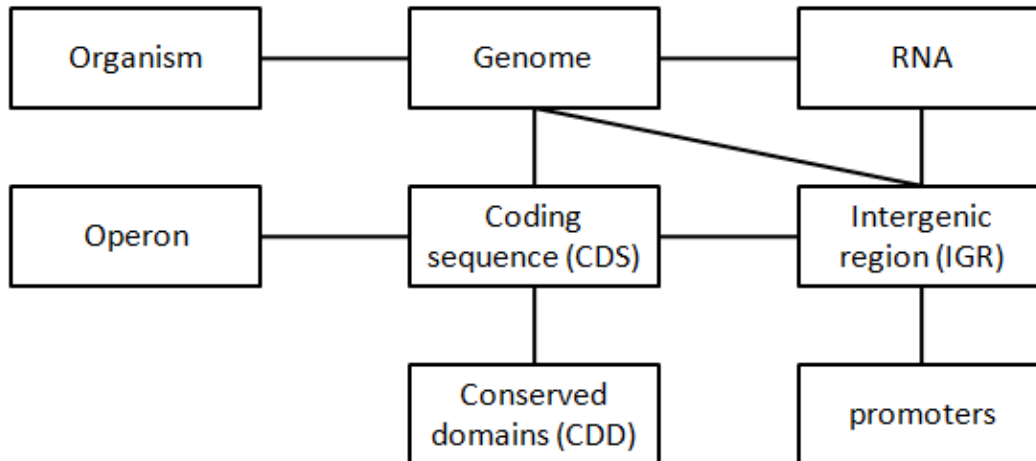


Figure 14 - RiboGap's v2 and v2.1 simplified diagram (Fig 1 in the article)

This diagram illustrates the general structure of the database with associations between tables and data.

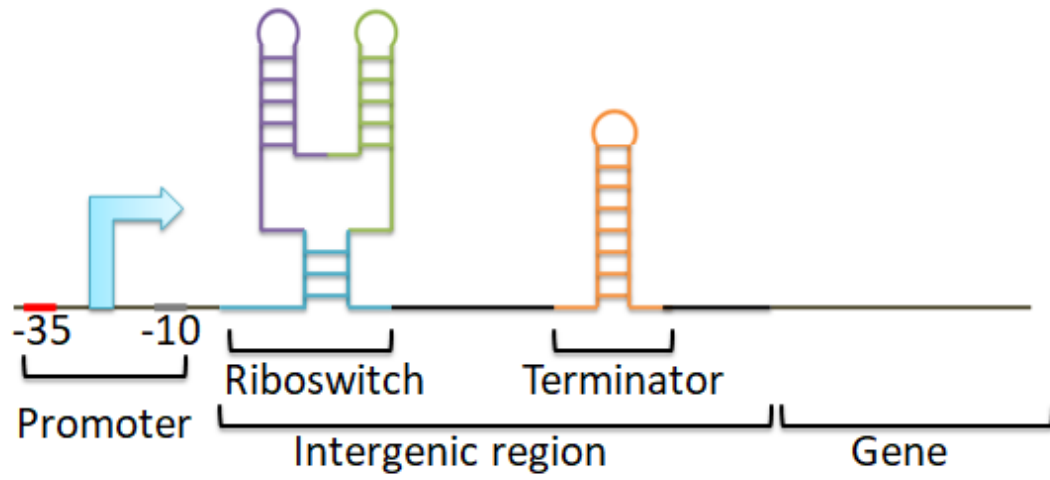
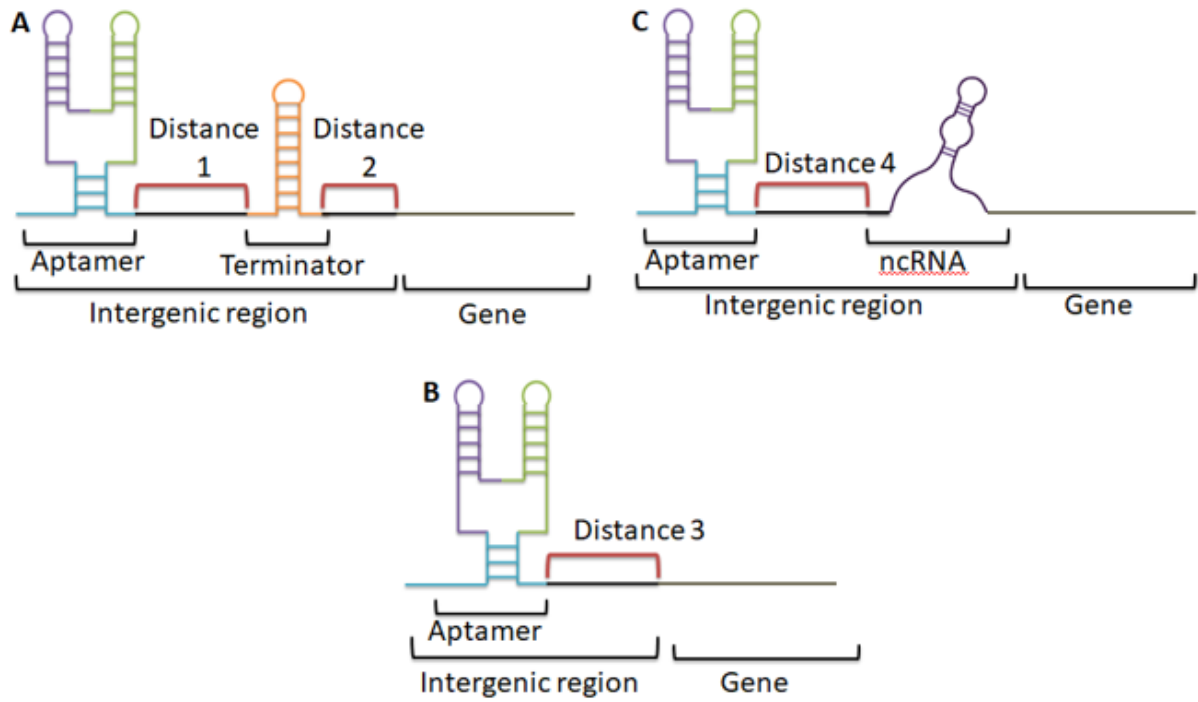


Figure 15- Illustration of different regulatory elements (Fig 2 in the article)



D

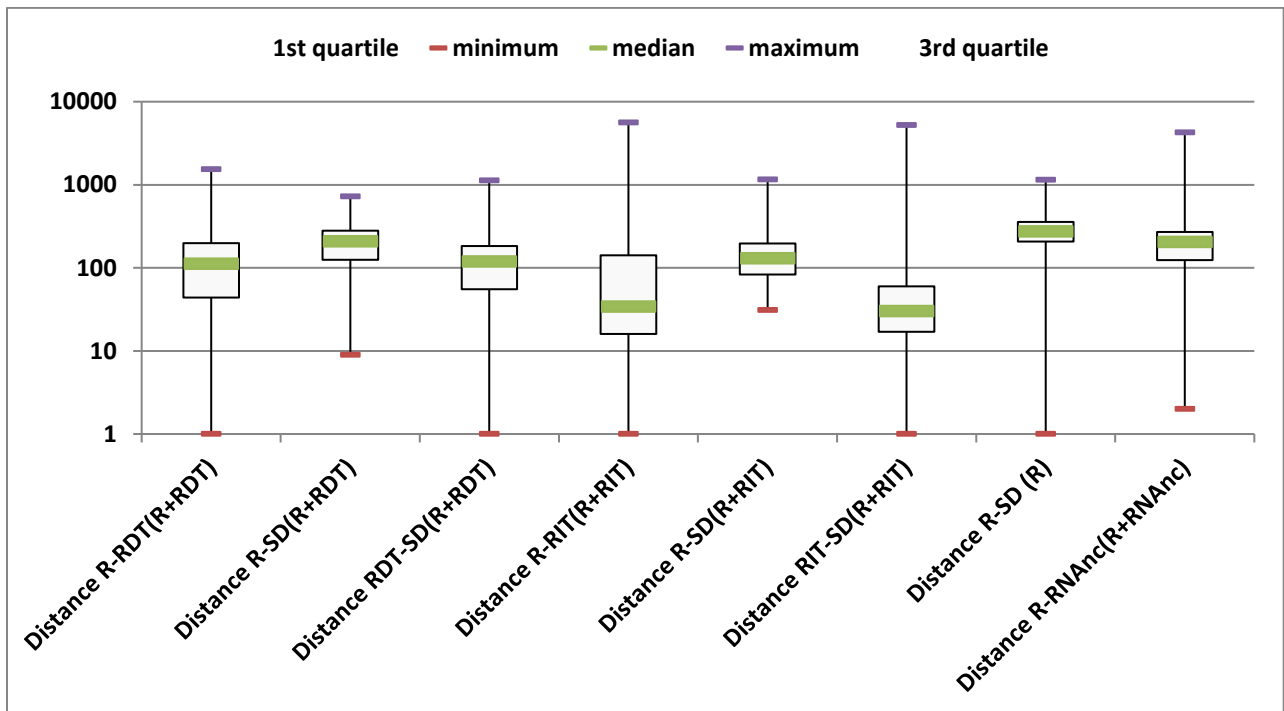


Figure 16 - Distances between riboswitches and expression platforms (Fig 3 in the article)

Different riboswitch expression platforms are illustrated with the distances calculated between these different elements. **a** which corresponds to a riboswitch with a terminator as an expression platform, we calculated the riboswitch-terminator, terminator-start codon and riboswitch-start codon distances. **b** where there is only the riboswitch, we calculated the distance riboswitch - codon start. **c** we calculated the riboswitch - ncRNA distance. **d** Box plot of distances between riboswitches and their putative expression platforms. Outliers (defined as the top 0.5% largest distances) were excluded. The position of the end of the intergenic region (i.e., immediately before the start codon) was used as a surrogate for SD-based expression platforms. R: Riboswitch, RDT: Rho Dependent Terminator, RIT: Rho Independent Terminator, AUG: start codon (including the less common non-AUG start codons).

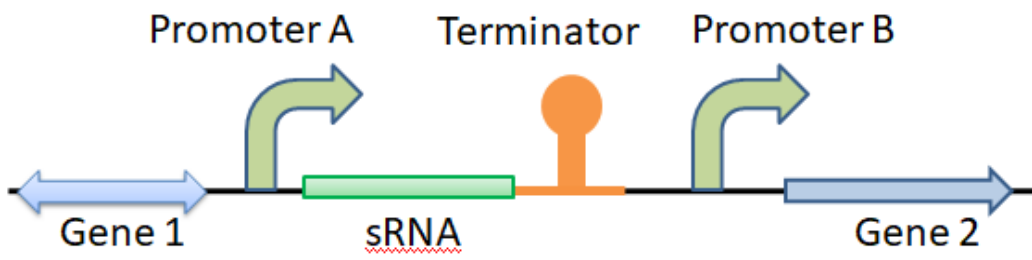


Figure 17 - Regulatory element context of a small RNA (Fig 4 in the article)

The presence of a terminator between two promoters can be a sign of the presence of a small RNA.

5.13 Tables

Tableau 2 - Comparison of available data in RiboGap v1, v2 and v2.1 (table 1 in the article)

	RiboGap v1	RiboGap v2	RiboGap v2.1
Genomes	2,757	9,857	45,122
CDS	8,708,354	67,433,293	680,476,616
Promoters	0	97,656,322	940,568,520
ncRNA families	2,452	3,019	3,019
RDTs	0	11,510,176	104,733,418
RITs	873,625	7,244,883	75,694,296
tRNA	153,228	1,164,179	10,496,574
rRNA	39,813	268,974	1,590,529
Other known ncRNAs	278,049	3,826,019	23,082,460

RiboGap v2 correspond to the RiboGap complete genomes and RiboGap v2.1 correspond to the RiboGap incomplete genomes. A large proportion of the numerous putative promoter predictions are presumed to be false positives (as described more thoroughly in the text).

Tableau 3 - Distances between riboswitch aptamers and putative expression platforms (table 2 in the article)

Content of IGR	Distances	Mean distance (Standard deviation)
Riboswitch + RDTs	Riboswitch –RDTs	85 (+/-130)
	RDTs – start codon	95 (+/-150)
	Riboswitch –start codon	260 (+/- 130)
Riboswitch + RITs	Riboswitch –RITs	120 (+/- 920)
	RITs – start codon	100 (+/- 1000)
	Riboswitch –start codon	260 (+/- 800)
Riboswitch (no terminator)	Riboswitch –start codon	210 (+/- 400)
Riboswitch + ncRNA	Riboswitch – ncRNA	960 +/- (1600)

Tableau 4 - Intergenic sequences with cis-regulatory RNAs which have a perfect Shine-Dalgarno close to SD (table 3 in the article)

	Start Codon	leader	riboswitch	thermoregulator	T-box	All
No AGGAGG close to start codon	ATG	94,195	116,889	26,812	41,554	56,821,227
	CTG	138	469	49	31	326,787
	TTG	5,953	6,291	1,223	1,859	2,772,756
	GTG	5,834	13,506	1,956	2,603	6,325,439
	Total	106,120	137,155	30,040	46,047	66,246,209
AGGAGG close to start codon (<= 12 bases)	ATG	18,668	12,875	2,138	13,549	1,952,251
	CTG	1	2	1	1	2,922
	TTG	949	886	119	787	179,880
	GTG	847	1,506	21	532	245,518
	Total	20,465	15,269	2,279	14,869	2,380,571
Ratio: no AGGAGG	ATG	0,198	0,110	0,080	0,326	0,034
	CTG	0,007	0,004	0,020	0,032	0,009
	TTG	0,159	0,141	0,097	0,423	0,006
AGGAGG	GTG	0,145	0,112	0,011	0,204	0,039
	Total	0.193	0.111	0.076	0.323	0.036

Start codon, as annotated, leader, riboswitch, thermoregulatory, T-box: all Rfam “types” corresponding to cis-regulatory RNAs,

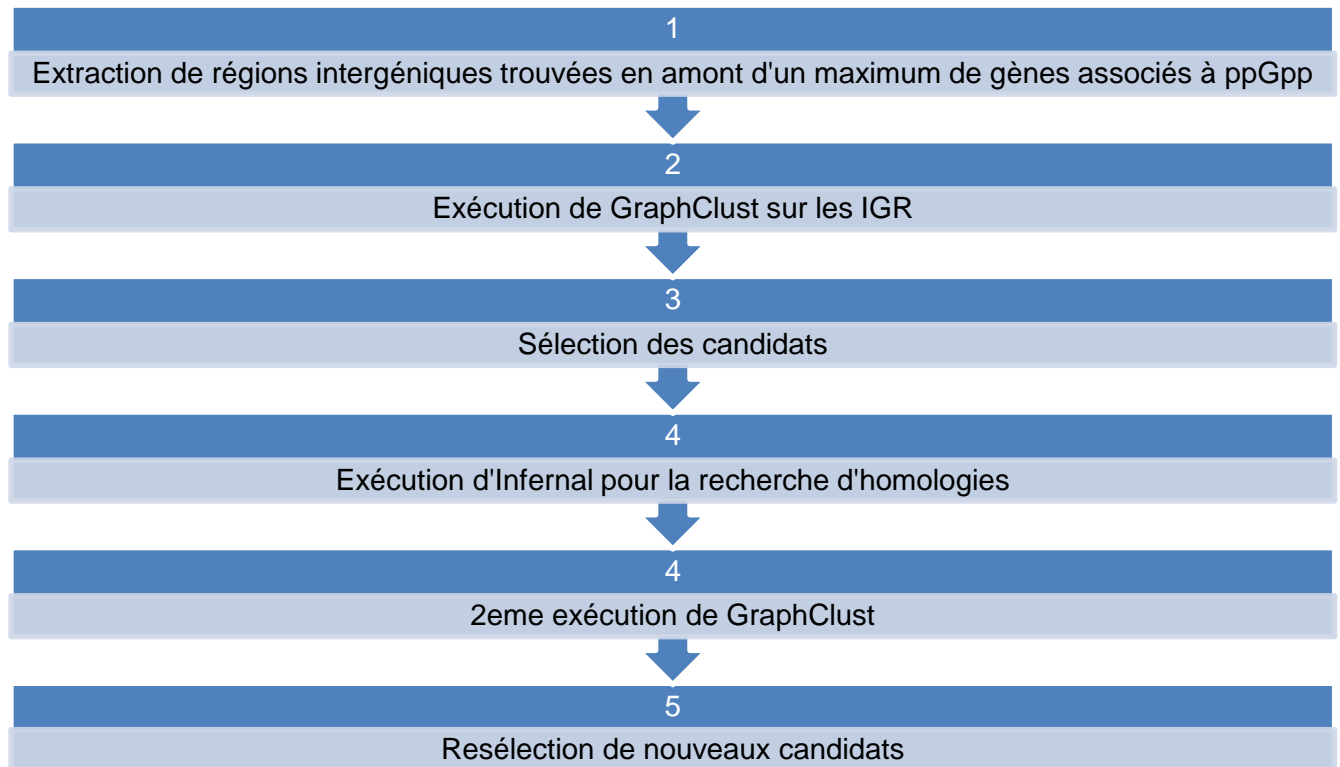
Ratio: ratio of the number of hits with a good Shine-Dalgarno (i.e., “AGGAGG” at less than 13 bases and more than 4 bases from the start codon) on the total number of hits for the corresponding cis-regulatory RNA.

6 DECOUVERTE DE NOUVEAUX ARN NONCODANTS ASSOCIES A PPGPP

Dans cette partie du projet, le but est d'utiliser les bases de données pour retrouver de nouvelles structures d'ARNnc, voire de nouveaux *riboswitchs*. De plus, à la lumière des *riboswitchs* pour lesquels de multiples classes existent pour un même ligand, comme c'est le cas pour SAM et le second messenger diGMP cyclique, il est imaginable que cela soit le cas pour plusieurs autres *riboswitchs*, tel que pour l'alarmone ppGpp. En effet, comme cette molécule joue un rôle majeur dans la régulation de nombreuses bactéries, il apparaît plausible que d'autres *riboswitchs* spécifique à ppGpp puissent exister. Cependant, certaines approches passées basées simplement sur les comparaisons des régions en amont d'un même gène seraient vraisemblablement limitantes, étant donné la diversité des gènes pouvant être régulés par ppGpp. À cet effet, la flexibilité de RiboGap pourrait contourner ce problème en permettant assez facilement de recueillir tous les IGR de tous les gènes pertinents.

6.1 Matériels et méthodes

Les étapes de la recherche de nouveaux ARNnc sont résumées dans le pipeline suivant :



Le *riboswitch* *ykkC-yxkD*, dont une sous-classe est connue pour lier ppGpp, sera utilisé comme test positif pour la méthode bio-informatique réalisée dans ce travail et pour aider à dresser une liste de gènes régulés par la molécule ppGpp.

6.1.1 RiboGap

Dans cette étape, deux types de requêtes SQL ont été exécutées pour extraire les régions intergéniques pour chacune des bases de données RiboGap v1, RiboGap v2 puis RiboGap v2.1. Comme nous recherchons un élément régulateur (pas nécessairement un *riboswitch*) relié à l'alarmone ppGpp, les IGR d'intérêt seront en amont de gènes associés d'une façon ou d'une autre à cette molécule.

6.1.1.1 Requête simple :

Dans cette requête, la plus simple, nous avons toutes les régions intergéniques de toutes les bactéries qui se trouvaient devant les gènes ou dont les produits des gènes étaient « ppGpp », « pentaphosphate », « tetraphosphate », incluant les gènes bien connus *relA*, *spoT* et *dksA*. Comme suivant :

```
Select      distinct cds.num_cle, cds.gene, cds.product, fragment.fragment,
gap5.num_cle, gap5.sequence, gap5.size
From        fragment inner join cds on fragment.fragment = cds.fragment
            inner join gap5 on cds.num_cle = gap5.num_cle
where       (cds.product like '%ppgpp%'           OR
cds.product like '%pentaphosphate%'             OR
cds.product like '%tetraphosphate%'             OR
cds.gene like '%ppgpp%'                         OR
cds.gene like '%pentaphosphate%'               OR
cds.gene like '%tetraphosphate%'               OR
cds.gene like '%relA%'                         OR
cds.product like '%relA%'                      OR
cds.gene like '%spoT%'                         OR
cds.product like '%spoT%'                     OR
cds.product like '% dksA %'                   OR
cds.gene like '%dksA%')                       AND
gap5.size > 50
```

```
into outfile '/var/lib/mysql-files/ppgpp_partie_query.txt';
```

6.1.1.2 Requête complexe :

Dans ce cas, nous voulions aller plus loin et inclure tous les gènes qui sont régulés par le *riboswitch* ykkC-yxkD, les gènes sont cités dans la figure 9-C.

```
Select      distinct cds.num_cle, cds.gene, cds.product, fragment.fragment, gap5.num_cle,
            gap5.sequence
from        fragment inner join cds on fragment.fragment = cds.fragment
inner join  gap5 on cds.num_cle = gap5.num_cle
where       cds.product like '%ppgpp%'           OR
            cds.product like '%pentaphosphate%' OR
            cds.product like '%tetraphosphate%' OR
            cds.product like '%COG3694%'        OR
            cds.product like '%COG4785%'        OR
            cds.product like '%BCAA biosynthesis%' OR
            cds.product like '%BCAA transporters%' OR
            cds.product like '%glutamate synthases%' OR
            cds.product like '%ABC transporters%' OR
            cds.product like '%relA%'           OR
            cds.product like '%spoT%'          OR
            cds.product like '%dksA%'          OR
            cds.product like '%ilvH%'          OR
            cds.product like '%ilvD%'          OR
            cds.product like '%ilvB%'          OR
            cds.product like '%ilvE%'          OR
            cds.product like '%ilvC%'          OR
            cds.product like '%leuB%'          OR
            cds.product like '%leuA%'          OR
            cds.product like '%leuD%'          OR
            cds.product like '%livM%'          OR
            cds.product like '%livH%'          OR
            cds.product like '%livG%'          OR
            cds.product like '%livF%'          OR
```

```

cds.product like '%gltS%' OR
cds.product like '%glexB%' OR
cds.product like '%gltBD%' OR
cds.product like '%natA%' OR
cds.gene like '%relA%' OR
cds.gene like '%spoT%' OR
cds.gene like '%dksA%' OR
cds.gene like '%ykkC%' OR
cds.gene like '%yxD%' OR
cds.gene like '%ilvH%' OR
cds.gene like '%ilvD%' OR
cds.gene like '%ilvB%' OR
cds.gene like '%ilvE%' OR
cds.gene like '%ilvC%' OR
cds.gene like '%leuB%' OR
cds.gene like '%leuA%' OR
cds.gene like '%leuD%' OR
cds.gene like '%livM%' OR
cds.gene like '%livH%' OR
cds.gene like '%livG%' OR
cds.gene like '%livF%' OR
cds.gene like '%gltS%' OR
cds.gene like '%glexB%' OR
cds.gene like '%gltBD%' OR
cds.gene like '%natA%'

```

```

into outfile '/var/lib/mysql-files/ppgpp_complet_query.txt';

```

6.1.2 GraphClust

GraphClust est un outil bio-informatique de regroupement (*clustering*) des ARN basé sur l'alignement des structures secondaires des séquences (Heyne *et al.*, 2012). Il existe une version implémentée de GraphClust dans Galaxy (Miladi *et al.*, 2019) sous forme de workflow et se compose d'un ensemble d'étapes. Chaque étape est réalisée par un outil spécifique intégré dans GraphClust (Figure 18).

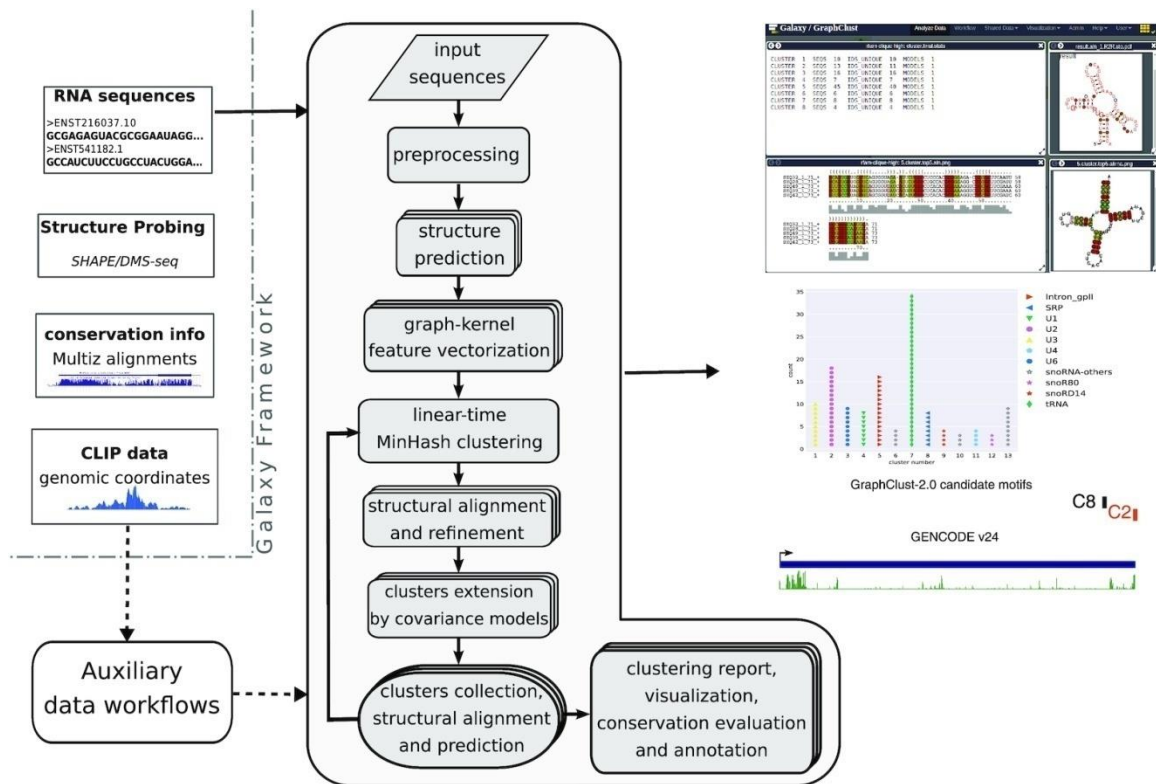


Figure 18 : Diagramme du pipeline de *clustering* complet. Les phases exécutées en parallèle sont représentées dans des boîtes empilées

(Copyright (C) 2007 Free Software Foundation, Inc. <http://fsf.org/>)

GraphClust fait intervenir plusieurs outils durant les étapes, à savoir : BlastClust (regroupement basé sur le score d'identité), RNAShape (Recherche de motifs structuraux dans l'ARN), LocARNA (alignement de séquences d'ARN de structure inconnue) et Infernal (Recherche d'homologie).

Etape1 : Prétraitement séquentiels

Le pipeline permet de regrouper plusieurs séquences qui proviennent de différentes sources. D'où la nécessité de faire un prétraitement des séquences d'entrée. Dans notre cas, le fichier d'entrée est l'ensemble des séquences intergénomiques extraites précédemment de RiboGap v1, RiboGap v2 et v2.1. Le prétraitement des séquences est réalisé par cd-hit (Fu *et al.*, 2012), qui supprime les séquences identiques à plus de 97%, pour éviter d'avoir 100% de conservation dans les alignements, et classe le reste des séquences qui s'alignent dans des *clusters*.

Etape2 : Détermination de la structure

Dans cette phase, la structure de chaque séquence est prédite en utilisant l'outil RNAShape (Janssen & Giegerich, 2015), qui cartographie les structures à un domaine arborescent de formes, en conservant la contiguïté et l'imbrication des caractéristiques structurales, sans tenir compte des longueurs d'hélice.

Etape3 : Alignement structurel

Afin de créer un modèle de haute qualité du cluster, un alignement de séquence-structure de l'ensemble des candidats est effectué avec l'outil LocARNA (Will *et al.*, 2012). Il comprend plusieurs outils permettant de produire des paires et des alignements multiples des séquences. Ces outils s'appuient sur le modèle d'énergie libre de Turner pour replier et aligner simultanément les séquences en fonction de leurs caractéristiques et de leurs structures.

Etape4 : Modèle de cluster

Les candidats les mieux classés sont réalignés avec l'outil LocARNA pour identifier une région d'alignement fiable et estimer les bordures du motif local commun grâce au calcul d'un score de fiabilité d'alignement. Suite à cela, un modèle de covariance (CM) est finalement créé en appliquant l'outil Infernal sur les sous-séquences identifiées.

Etape5 : Numérisation du modèle

Chaque cluster contient un modèle CM qui est utilisé pour rechercher des structures homologues aux séquences du cluster (fonction `cmsearch`) dans l'ensemble des séquences des IGR. Les hits de séquence qui sont considérés comme significatifs (bit-score ≈ 20) sont ajoutés au cluster final.

Etape6 : Itération et suppression

Les membres du cluster trouvés dans la phase précédente sont supprimés de l'ensemble de données et une nouvelle itération commence à partir de la phase 4. La condition de terminaison est donnée soit par un nombre maximum d'itérations prédéterminé dans la configuration du GraphClust, une limite de temps ou lorsque l'ensemble de données restant est épuisé.

Etape7 : Post-traitement

Les *clusters* redondants sont fusionnés et les instances appartenant à plusieurs *clusters* sont attribuées sans ambiguïté. Les membres du cluster sont enfin classés par leur bitscore CM.

6.1.3 Sélection de *clusters* candidats

Comme mentionnés précédemment, plusieurs *clusters* sont obtenus de GraphClust. Donc une analyse de chaque alignement de chaque cluster est requise pour sélectionner les candidats les plus intéressants. Cette sélection est faite par rapport à la structure secondaire qui pourrait être potentiellement une nouvelle structure d'ARNnc. Pour réaliser cette sélection manuelle, trois critères sont vérifiés à savoir : la conservation des séquences, l'existence de co-variation dans plusieurs paires de bases de l'alignement et la non-existence d'ARNnc connu en vérifiant les séquences constituant les *clusters* avec Rfam. La vérification de la conservation et la co-variation est faite suivant un code couleur. La co-variation est indiquée en jaune pour deux paires de bases, vert pour 3 types de paires de base, bleu clair pour 4 types de paires de base, bleu foncé pour 5 types de pair de base et mauve pour 6 types de paires de base.

		Types of Pairs					
		1	2	3	4	5	6
Incompatible Pairs	0	Red	Yellow	Green	Cyan	Blue	Purple
	1	Light Red	Light Yellow	Light Green	Light Cyan	Light Blue	Light Purple
	2	Very Light Red	Very Light Yellow	Very Light Green	Very Light Cyan	Very Light Blue	Very Light Purple

6.1.4 Infernal

L'outil Infernal (*inference of RNA alignment*) permet de chercher dans une base de données des instances homologues à des alignements de structures donnés en entrée au programme. Dans notre cas, la base de données est l'ensemble des séquences des génomes complets/incomplets de procaryotes et les entrées sont les alignements des *clusters* sélectionnés de format modèle de covariance des *clusters* intéressants. Cette étape pourrait donc permettre de trouver des homologues dans des IGR qui n'avaient jusqu'alors aucune association connue avec ppGpp. Des séquences additionnelles correspondant aux modèles de structure trouvés sont ainsi été obtenus et une 2^{ème} utilisation de GraphClust est ainsi réalisée pour avoir d'autres *clusters* qu'on a aussi sélectionnés.

6.2 Résultats

En exécutant les deux types de requêtes (simple et complexe mentionnées) consécutivement sur RiboGap v2 puis sur RiboGap v2.1, quatre ensembles de séquences intergéniques en format fasta ont été générés. Suivant le pipeline bio-informatique, GraphClust a été exécuté sur les quatre fichiers fasta. La sélection des candidats intéressants a été faite sur les différents cas et ce de façon indépendante. La liste complète des *clusters* pour chaque est donné en annexe en montrant pour chaque cluster l'alignement structurel ainsi que quelques structures secondaires donnés par GraphClust.

- i. Résultats de la requête simple exécutée sur RiboGap v2 (Voir l'annexe 11.2.1.1)
- ii. Résultats de la requête complexe exécutée sur RiboGap v2 (Voir l'annexe 11.2.1.2)
- iii. Résultats de la requête simple exécutée sur RiboGap v2.1 (Voir l'annexe 11.3.1.1)
- iv. Résultats de la requête complexe exécutée sur RiboGap v2.1 (Voir l'annexe 11.3.1.2)

Après la recherche d'homologues et la réexécution de GraphClust, d'autres *clusters* ont été générés. La liste des candidats sélectionnés est présentée comme tel :

- i. Résultats de la requête simple exécutée sur RiboGap v2 (Voir l'annexe 11.2.2.1)
- ii. Résultats de la requête complexe exécutée sur RiboGap v2 (Voir l'annexe 11.2.2.2)
- iii. Résultats de la requête simple exécutée sur RiboGap v2.1 (Voir l'annexe 11.3.2.1)
- iv. Résultats de la requête complexe exécutée sur RiboGap v2.1 (Voir l'annexe 11.3.2.2)

Afin de vérifier notre hypothèse de départ, qui était qu'une mise à jour de la base de données donnerait plus de candidats, nous avons joint les résultats de GraphClust obtenus durant l'exécution de la requête simple sur RiboGap v1. La liste complète est également disponible en Annexe 11.1.

A titre d'exemple, dans ce qui suit nous analysons de façon détaillée un exemple de *cluster* obtenu choisi au hasard parmi les clusters intéressants. Les séquences et la structure du *cluster* ne sont pas connus de Rfam ce qui rend le *cluster* un bon candidat pour être un nouvel ARN noncodant. Pour commencer, nous présentons l'alignement (Figure 20). Il est à noter que certaines prédictions peuvent parfois avoir des portions où les structures prédites sont moins fiables. Mais plusieurs autres régions où les tiges prédites semblent avoir une haute probabilité de se former et où plusieurs pb montrent des variations compatibles (incluant de la covariation) supportant la structure prédite.

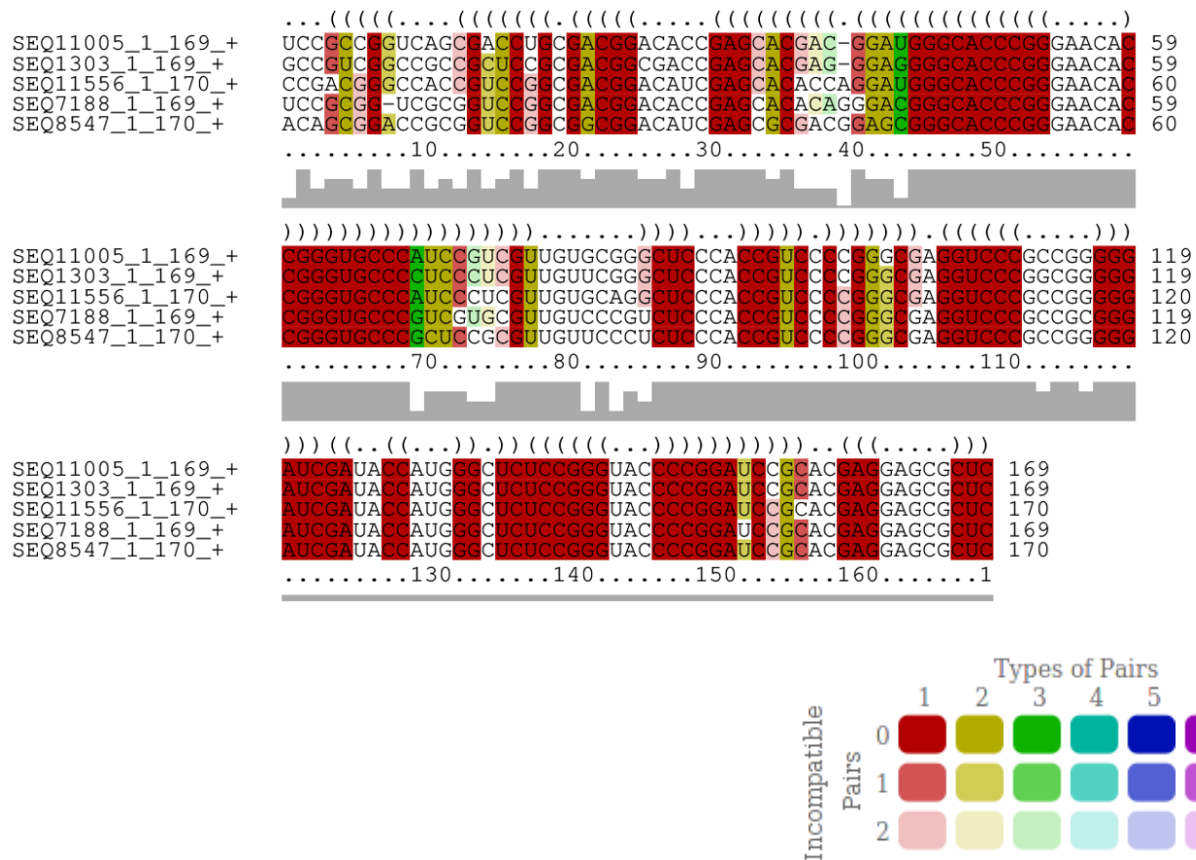
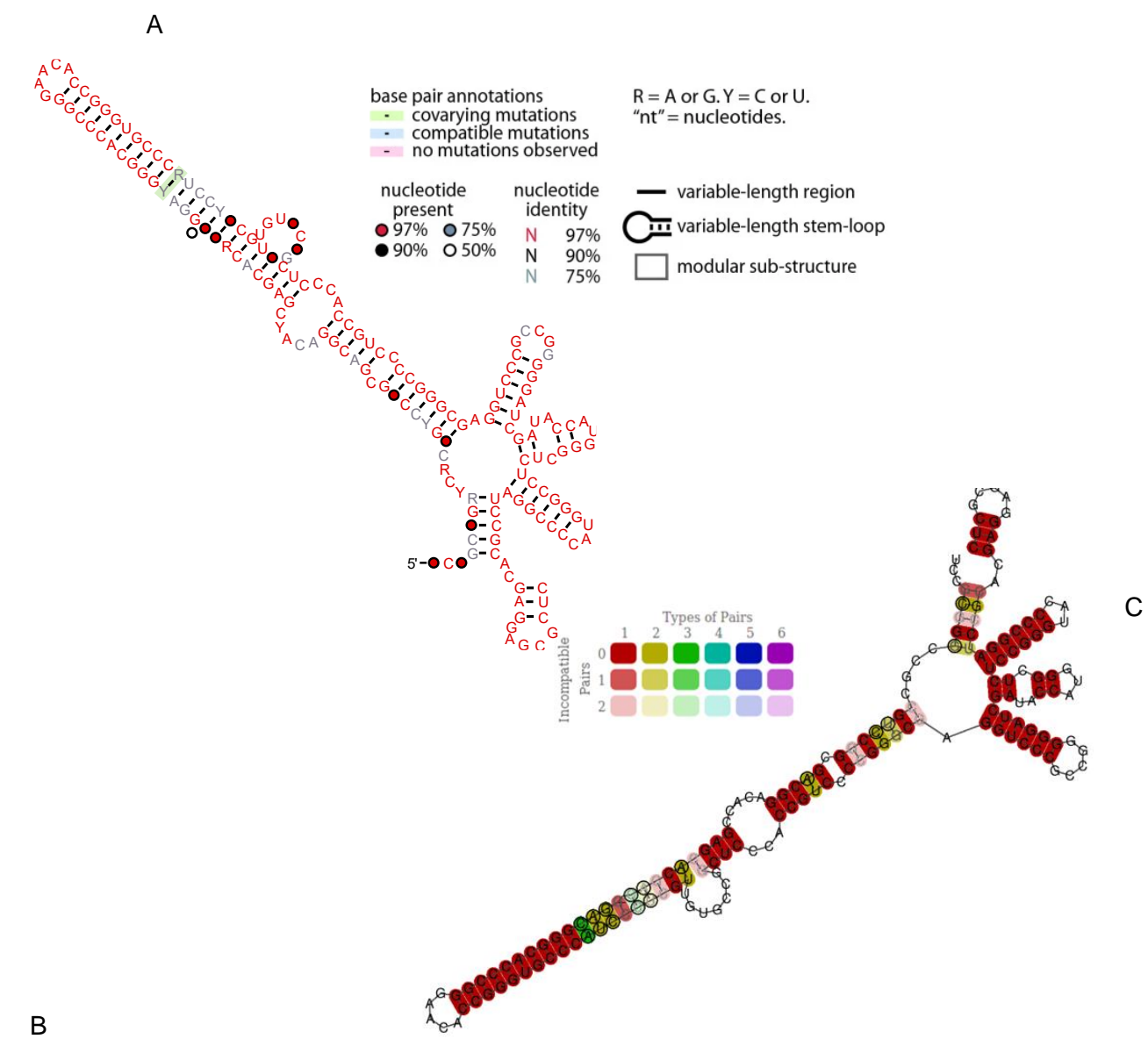


Figure 19 - Alignement structurel du Cluster 45

Les parenthèses = paires de base

Ombre grise = conservation

Pour vérifier si la co-variation notée dans l'alignement est statistiquement significative, nous avons exécuté R-scape (Rivas *et al.*, 2017) qui évalue les covariations par paires observées dans un alignement multiple de séquences pour déterminer s'il est plus probable qu'elle soit survenue par hasard ou dû à une conservation de structure.



List of covarying basepairs

in given structure	Left base	Right base	Covariation Score	E-value	Substitutions	Power
*	44	70	5.10481	0.00131526	8	0.05

Figure 20 - Structure secondaire prédite pour le cluster 45
A - Structure secondaire générée par R2R. B- Covariation générée par R-scape Les positions correspondent à la covariation. **C- Structure secondaire prédite par GraphClust du cluster 45 « - »** correspond aux Gaps.

Afin de comprendre et d'analyser le contexte génétique des séquences constituant l'alignement du *cluster* 45. Nous avons utilisé RiboGap pour retrouver le gène situé en aval de ces régions

intergéniques de même que l'existence de potentiels promoteurs ou de terminateurs Rho dépendants/indépendants.

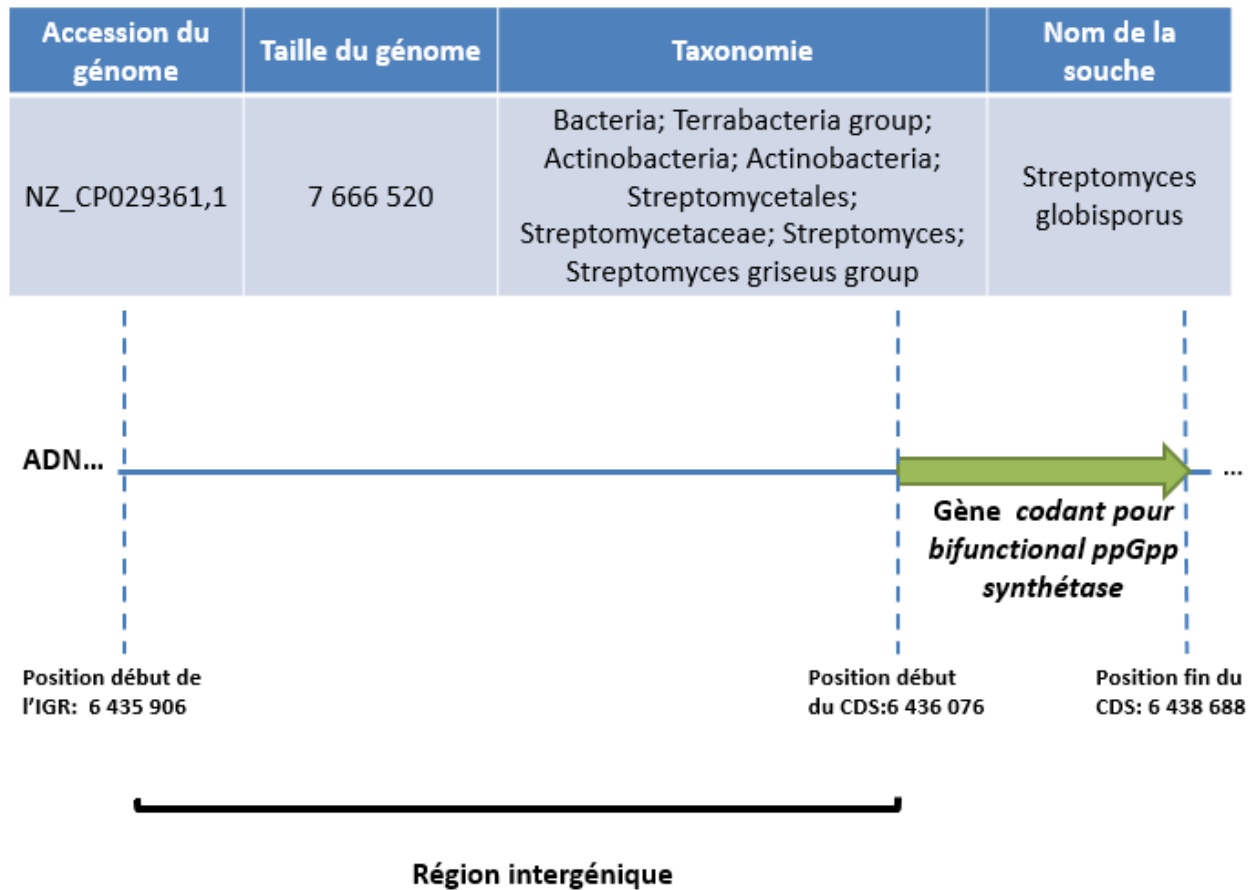


Figure 21 - Contexte génétique de l'une des séquences intergéniques du cluster 45

Ceci est une séquence parmi les séquences alignées de la figure 19. En utilisant RiboGap v2, on a trouvé que les autres séquences du même alignement sont aussi devant le même gène et appartiennent à la même taxonomie.

Le *cluster 45* fait partie des *clusters* les plus intéressants, donc après l'avoir sélectionné nous avons cherché des homologues des séquences constituant cet alignement dans tous les génomes complets, et ce en exécutant Infernal (cmcalibrate et cmsearch) avec comme input le modèle de covariance de ce *cluster*. On a obtenu 29 séquences homologues aux séquences du *cluster 45*. Ensuite nous avons exécuté une autre fois GraphClust sur les séquences homologues trouvées. A noter que ce travail a été fait pour chaque cluster sélectionné des requêtes simples et complexes exécutées dans les différentes versions de RiboGap. On a obtenu le *cluster* suivant :

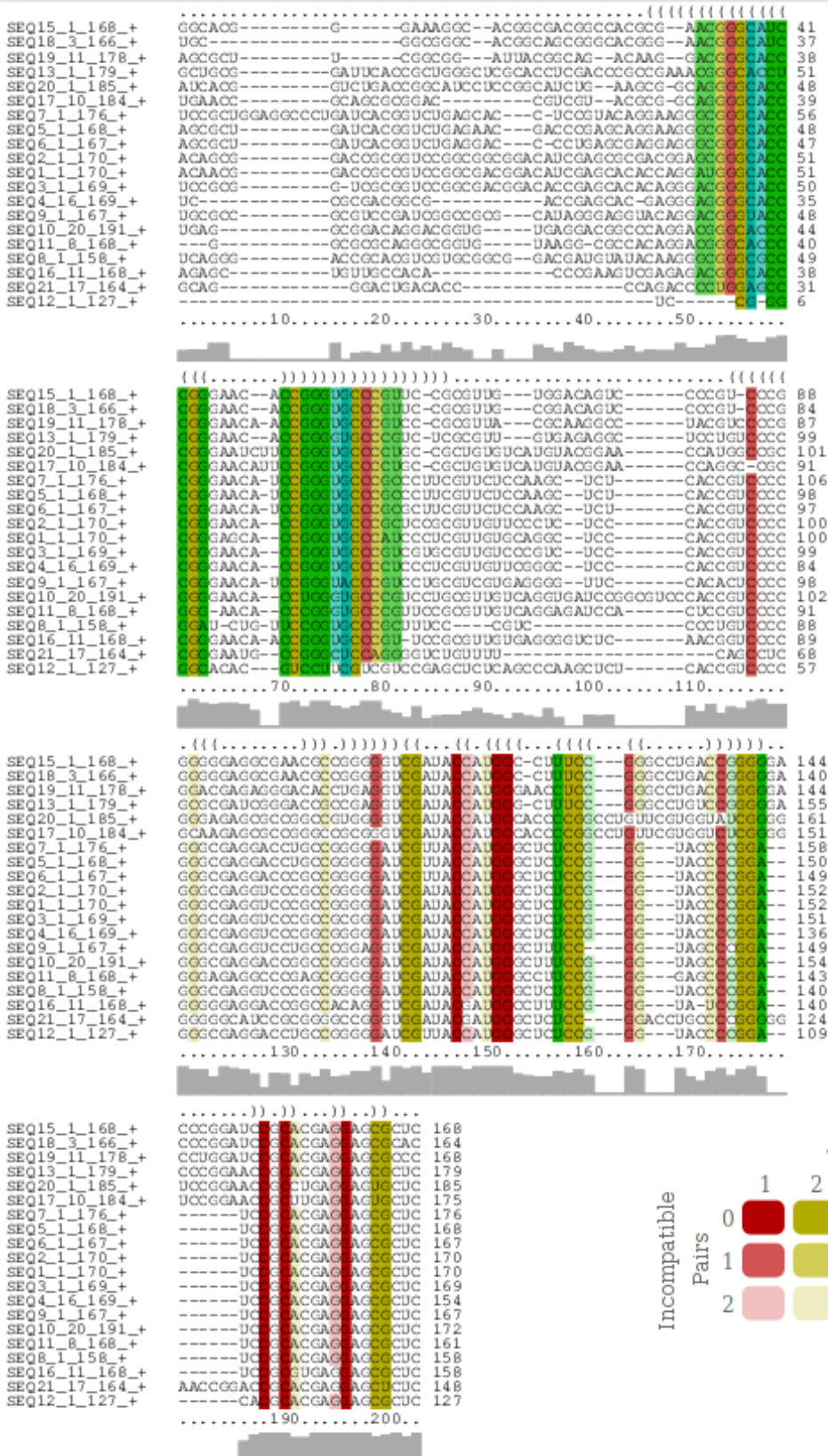


Figure 23 - Cluster résultant des séquences homologues du cluster 45

Le tableau suivant récapitule le nombre de *clusters* obtenus après l'exécution de GraphClust et après la recherche d'homologues pour les différentes requêtes réalisées sur RiboGap v1, RiboGap v2 et RiboGap v2.1.

Bases de données	Requête exécutée	Nombre total de <i>clusters</i> obtenu	<i>Clusters</i> sélectionnés	Nombre total de <i>clusters</i> obtenu après recherche d'homologie	<i>Clusters</i> sélectionnés après recherche d'homologie
RiboGap v1	Requête 1 (Seulement une requête a été exécutée)	14	4	-	-
RiboGap v2	Requête 1	84	26	7	5
	Requête 2	89	18	4	4
RiboGap v2.1	Requête 1	59	5	7	1
	Requête 2	68	4	5	4

Tableau 5 - Comparaison du nombre de candidats trouvés entre les trois versions des bases de données

Nous remarquons dans ce tableau que le nombre de cluster obtenus dans les nouvelles versions de RiboGap est largement plus grand que ceux obtenus avec RiboGap v1. De plus, nous remarquons que RiboGap v2.1 présente moins de clusters que RiboGap v2 ce qui s'explique par la condition rajoutée aux requête du fait que la taille de la séquence intergénique devrait être plus grande que 50 nt.

7 DISCUSSION GÉNÉRALE

7.1 Utilité de RiboGap

Extraire toutes les régions intergéniques se trouvant en amont d'un ou plusieurs gènes dans différentes bactéries peut être une tâche assez complexe pour les scientifiques n'ayant pas de compétences en bio-informatique ou en programmation. De plus, il n'est pas évident de

déterminer la présence des différents éléments (les terminateurs Rho dépendants ou Rho indépendants, les promoteurs ou les parties des IGR transcrites en ARNnc) contenus dans ces régions intergéniques. La raison est que ceci nécessite l'utilisation de plusieurs outils de prédiction pour un seul IGR. Ceci met en évidence l'importance de RiboGap qui donne la possibilité d'avoir les régions intergéniques en quelques clics à partir de l'interface web, tout en récupérant les éléments régulateurs qu'ils peuvent contenir. Comme mentionné dans le manuscrit du chapitre 4, la base de données peut être utilisée pour extraire différents types de données permettant ainsi de répondre à différentes questions biologiques, à titre d'exemples : l'existence de structures G-quadruplex, la recherche de nouveaux petits ARN, la recherche de plusieurs éléments régulateurs (*riboswitchs*, promoteurs, terminateurs). La possibilité de croiser les données de séquences (de nucléotides ou d'acides aminés), gènes, génomes, éléments régulateurs connus/prédits et même des phénotypes, permet de facilement obtenir, à l'échelle génomique, des réponses à des hypothèses qui peuvent être le début de nouveaux projets ou qui peuvent aider à supporter certaines hypothèses émergeant de données expérimentales.

7.2 Outils de prédiction utilisés dans la mise à jour de la base de données

Comme mentionné dans l'article, en plus de la mise à jour de la base de données, nous avons entrepris de prédire plusieurs éléments tels que : les promoteurs en utilisant bTSSfinder, les terminateurs Rho-dépendants avec l'outil RhoTermPredict, les terminateurs Rho-indépendants avec RNIE, les ARN noncodants, les ARNt et les ARNr. Différents outils bio-informatiques ont été exécutés individuellement sur tous les génomes, dont la taille peut atteindre 14 782 100 paires de bases, et la durée d'exécution pour un outil (entre bTSSfinder, RhoTermPredict, tRNAscan, Infernal, ...) a varié entre 5 à 30 minutes pour un seul génome. Sachant que le nombre de génomes complets est de 9 857 et que le nombre des génomes incomplets est de 45 122, la durée d'exécution de chaque outil bio-informatique pour tous les génomes complets et incomplets auraient pu prendre plusieurs mois. En plus du temps, la taille des données nécessitait plusieurs interventions manuelles. Ces deux défis majeurs ont pu être résolus en utilisant les serveurs de Calcul Canada, on a pu ainsi réduire le temps à juste quelques semaines pour chaque programme.

Un autre défi avec les prédictions Bio-informatiques est la faisabilité d'exécuter un outil sur le génome au complet. En effet, si on prend à titre d'exemple la prédiction des promoteurs, nous avons testé plusieurs outils, au final on s'est arrêté sur bTSSfinder car c'était un programme qui nous permettait de traiter toutes les IGR d'un génome en une seule exécution et qui a réussi à

nous donner un maximum d'information sur le promoteur (position du promoteur, polarité, position de la boîte -35, séquence de la boîte -35, position de la boîte -10, séquence de la boîte -10) contrairement aux autres outils qui prenaient une séquence de taille limitée et cela aurait pu nous prendre encore beaucoup plus de temps car il aurait fallu couper la séquence du génome en plusieurs fragments. En résumé, on a dû tester différents outils pour la prédiction de chaque élément pour trouver un juste milieu entre avoir un maximum d'informations, le temps disponible et la compatibilité avec l'infrastructure de Calcul Canada. Enfin, toujours en restant dans l'exemple des promoteurs, une fois que nous avons inséré toutes nos données, nous avons remarqué plusieurs faux positifs, nous espérons résoudre une partie de ce souci en incluant pour chaque élément un score ou un *e-value*, mais cela ne permettait malheureusement pas de discriminer les faux-positifs. Les prédictions de promoteurs demeurent donc peu fiables, mais tout de même potentiellement informatives pour donner des voies à explorer de façon plus approfondie.

Donc de façon générale, la base de données reste utile pour un utilisateur qui voudrait générer un maximum de données en utilisant des requêtes qui permettent de combiner les différents terminateurs, promoteurs, ARNnc, opérons, séquences codantes et régions intergéniques. D'après les différents exemples réalisés sur les différents éléments tels que cités dans l'article, nous avons conclu que dans le cas des promoteurs, il faudra penser à rajouter d'autres étapes pour trier ces données et permettre ainsi de réduire les faux positifs. En revanche, les prédictions d'ARNnc, faites via Infernal, lorsqu'utilisées conjointement avec les *e-value*, sont beaucoup plus fiables. Quant à la prédiction des terminateurs, cela nous a parues relativement fiable.

7.3 Augmentation du nombre de candidats

Dans ce mémoire, nous décrivons l'utilisation de RiboGap pour extraire les régions intergéniques se trouvant en amont des gènes associés à ppGpp et au *riboswitch* connu ykkC-yxkD. La récupération de ces séquences est l'étape primaire du pipeline bio-informatique basé sur la génomique comparative. En exécutant le programme GraphClust sur les données extraites de RiboGap v1, puis de RiboGap v2 et v2.1. Nous avons remarqué une grande différence dans le nombre de *clusters* obtenus avec GraphClust et du nombre de candidats sélectionnés comme le montrait le tableau 5 qui peuvent être des potentiels nouveaux ARNnc, voire de nouveaux *riboswitchs*. Nous avons donc réalisé notre objectif qui était d'augmenter notre capacité à trouver plus de candidats en mettant à jour la base de données et en lui rajoutant de nouvelles fonctionnalités. En perspective, il serait également intéressant de penser à une 3^{ème} version qui

regrouperait des métagénomés. Vu le grand nombre de métagénomés qui existent et la quantité de données représentée par chaque métagénome, ceci augmentera considérablement le nombre de candidats pour la découverte de nouveaux ARNnc. À ce titre, il faudrait retravailler « l'automatisation » de la mise-à-jour de façon à pouvoir plus facilement faire de nouvelles bases de données pour chaque projet de métagénomique.

7.4 Absence du *riboswitch* ykkC-yxkD de la sous-classe ppGpp

En utilisant Rfam, nous avons cherché le *riboswitch* ykkC-yxkD dans toutes les séquences des différents *clusters* obtenus comme test positif pour vérifier si avec notre pipeline bio-informatique nous retrouvons ce *riboswitch*. Cela dit le *riboswitch* n'a pas été trouvé et cela peut s'expliquer soit par la faiblesse du pipeline bio-informatique (GraphClust) ou le fait que les séquences où le *riboswitch* ykkC-yxkD existe sont des métagénomés et que dans nos bases de données, on regroupe uniquement les génomes complets et incomplets.

En effet, une recherche de RiboGap révèle qu'avec les mots-clés utilisés, on ne retrouve qu'un seul exemplaire de l'ARN ykkC-yxkD. Même en tentant d'utiliser des mots-clés plus généralistes associés aux fonctions en lien avec le *riboswitch* ppGpp (voir figure 9), on ne trouve que trois exemplaires de plus (avec de mauvais *e-value*, de l'ordre de 0,02 à 0,07), dont deux séquences pratiquement identiques. En d'autres mots, même avec les séquences mises à jour, l'alignement de ces quelques séquences est vraisemblablement insuffisant pour y trouver de la covariation et donc y repérer une structure conservée. Donc il faudra penser à créer une autre version de RiboGap qui regrouperait les métagénomés les plus importants.

7.5 La méthodologie In-line probing

Maintenant que nous avons sélectionné plusieurs candidats intéressants, la prochaine étape serait d'effectuer des tests en laboratoire via des expériences *in vitro* appelés « *In-line probing* ». Cette méthode utilise l'instabilité naturelle de l'ARN pour élucider les caractéristiques de la structure secondaire et les capacités de liaison au ligand des *riboswitchs* et ce en analysant les molécules d'ARN dérivées des régions intergéniques contenant les structures candidates (Regulski & Breaker, 2008). En d'autres termes, le *In-line probing* permet de déterminer les éventuels changements structurels de l'ARN en réponse à une liaison à un ligand et aussi de comparer la réponse à celle en présence d'autres ligands analogues pour évaluer la spécificité.

8 CONCLUSION

Le projet présenté a permis dans un premier temps de créer une base de données qui englobe à la fois plusieurs données génétiques de différentes sources (NCBI, Rfam et ODB), en plus de données dérivées de prédictions générées par différents outils tels que : RNIE pour les terminateurs Rho indépendant, RhoTermPredict pour les terminateurs Rho dépendants, bTSSfinder pour les promoteurs, tRNAscan pour les ARNt, Infernal pour les ARNnc et InterProScan pour l'annotation fonctionnelle des gènes.

Ensuite, nous avons utilisé cette même base de données pour extraire les régions intergéniques associés à ppGpp et au *riboswitch* ykkC-yxkD. Ces régions intergéniques ont été compilées avec GraphClust pour effectuer un alignement structurel des séquences, permettant ainsi une facilité dans la sélection de candidats. L'étape suivante en laboratoire « *In-line probing* » permettra de vérifier si un changement de structures de ces candidats a lieu en présence de ppGpp, reflétant leur liaison au ligand.

Les travaux présentés ici utilisent une méthode basée sur la fonction des gènes comme point de départ d'une stratégie de génomique comparative 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 avec aisance. L'utilisation de la suite d'outil GraphClust a permis des analyses rapides d'un large ensemble de données. Des travaux précédents sur des éléments régulateurs dans la région 5'UTR du gène *mnmC* avait relevé le problème des faux-positifs liés à la conservation des promoteurs, ce qui nous avait mené à l'inclusion de prédictions de promoteurs dans ces nouvelles versions de RiboGap v2 et v2.1. Ainsi, grâce à l'outil bTSSfinder qui permet de faire la prédiction des promoteurs et de prévoir ainsi un chevauchement avec les potentiels motifs prédits, nous avons pris les devants pour permettre une évaluation des candidats à la lumière de boîtes conservées de promoteurs.

Finalement, nous avons sélectionné une liste de motifs à tester au laboratoire comme motifs ARN régulateurs potentiels. Des expériences d'*In-line probing* avec différentes concentrations de ppGpp (ou pppGpp) pourraient permettre de confirmer la découverte de nouveaux *riboswitches* impliqués dans la régulation via cette importante « alarmone » bactérienne.

9 BIBLIOGRAPHIE

- Alberts B, Johnson A, Walter P, Lewis J, Raff M & Roberts K (2008) *Molecular cell biology*. New York: Garland Science.
- Argaman L, Hershberg R, Vogel J, Bejerano G, Wagner EGH, Margalit H & Altuvia S (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Current Biology* 11(12):941-950.
- Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB, Popov I, Roma-Mateo C, Theodosiou A & Mitchell AL (2012) The PRINTS database: a fine-grained protein sequence annotation and analysis resource—its status in 2012. *Database* 2012.
- Ayoubi TA & Van De Yen WJ (1996) Regulation of gene expression by alternative promoters. *The FASEB Journal* 10(4):453-460.
- Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N & Jona I (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proceedings of the National Academy of Sciences* 101(17):6421-6426.
- Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S & Sonnhammer EL (2004) The Pfam protein families database. *Nucleic acids research* 32(suppl_1):D138-D141.
- Batey RT (2011) Recognition of S-adenosylmethionine by riboswitches. *Wiley Interdisciplinary Reviews: RNA* 2(2):299-311.
- Beaudoin J-D & Perreault J-P (2013) Exploring mRNA 3'-UTR G-quadruplexes: evidence of roles in both alternative polyadenylation and mRNA shortening. *Nucleic acids research* 41(11):5898-5911.
- Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J & Sayers EW (2012) GenBank. *Nucleic acids research* 41(D1):D36-D42.
- Breaker RR (2011) Prospects for riboswitch discovery and analysis. *Molecular cell* 43(6):867-879.
- Bru C, Courcelle E, Carrère S, Beausse Y, Dalmar S & Kahn D (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic acids research* 33(suppl_1):D212-D215.
- Canese K & Weis S (2013) PubMed: the bibliographic database. *The NCBI Handbook [Internet]*. 2nd edition, National Center for Biotechnology Information (US).
- Carter C & Houlihan D (2001) Protein synthesis. *Fish physiology* 20:31-75.
- Chan PP & Lowe TM (2019) tRNAscan-SE: searching for tRNA genes in genomic sequences. *Gene Prediction*, Springer. p 1-14.
- Chen S, Lesnik EA, Hall TA, Sampath R, Griffey RH, Ecker DJ & Blyn LB (2002) A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. *Biosystems* 65(2-3):157-177.
- Childs L, Nikoloski Z, May P & Walther D (2009) Identification and classification of ncRNA molecules using graph properties. *Nucleic acids research* 37(9):e66-e66.
- Ciampi MS (2006) Rho-dependent terminators and transcription termination. *Microbiology* 152(9):2515-2528.

- Colston SM, Fullmer MS, Beka L, Lamy B, Gogarten JP & Graf J (2014) Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using *Aeromonas* as a test case. *MBio* 5(6):e02136-02114.
- Coronel C & Morris S (2016) *Database systems: design, implementation, & management*. Cengage Learning,
- Dalebroux ZD & Swanson MS (2012) ppGpp: magic beyond RNA polymerase. *Nature Reviews Microbiology* 10(3):203-212.
- De Lay N, Schu DJ & Gottesman S (2013) Bacterial small RNA-based negative regulation: Hfq and its accomplices. *Journal of Biological Chemistry* 288(12):7996-8003.
- DeLong EF & Pace NR (2001) Environmental diversity of bacteria and archaea. *Systematic biology* 50(4):470-478.
- Di Salvo M, Puccio S, Peano C, Lacour S & Alifano P (2019) RhoTermPredict: an algorithm for predicting Rho-dependent transcription terminators based on *Escherichia coli*, *Bacillus subtilis* and *Salmonella enterica* databases. *BMC bioinformatics* 20(1):1-11.
- Downward J (2004) RNA interference. *Bmj* 328(7450):1245-1248.
- DuBois P (2008) *MySQL*. Pearson Education,
- e Silva SdA, Echeverrigaray S & Gerhardt GJ (2011) BacPP: bacterial promoter prediction—a tool for accurate sigma-factor specific assignment in enterobacteria. *Journal of theoretical biology* 287:92-99.
- 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.
- Fu L, Niu B, Zhu Z, Wu S & Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23):3150-3152.
- Garant J-M, Perreault J-P & Scott MS (2018) G4RNA screener web server: user focused interface for RNA G-quadruplex prediction. *Biochimie* 151:115-118.
- Gardner PP, Barquist L, Bateman A, Nawrocki EP & Weinberg Z (2011) RNIE: genome-wide prediction of bacterial intrinsic terminators. *Nucleic acids research* 39(14):5845-5852.
- Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S & Eddy SR (2009) Rfam: updates to the RNA families database. *Nucleic acids research* 37(suppl_1):D136-D140.
- Garrity GM, Bell JA & Lilburn T (2015) Proteobacteria phyl. nov. *Bergey's Manual of Systematics of Archaea and Bacteria* :1-1.
- Garst AD, Edwards AL & Batey RT (2011) Riboswitches: structures and mechanisms. *Cold Spring Harbor perspectives in biology* 3(6):a003533.
- Gu J & Bourne PE (2009) *Structural bioinformatics*. John Wiley & Sons,
- Gundavaram S, Birznieks G & Guelich S (2000) *CGI Programming with Perl*. O'reilly,
- Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK & Beck E (2012) TIGRFAMs and genome properties in 2013. *Nucleic acids research* 41(D1):D387-D395.
- Haller A, Altman RB, Soulière MF, Blanchard SC & Micura R (2013) Folding and ligand recognition of the TPP riboswitch aptamer at single-molecule resolution. *Proceedings of the National Academy of Sciences* 110(11):4188-4193.

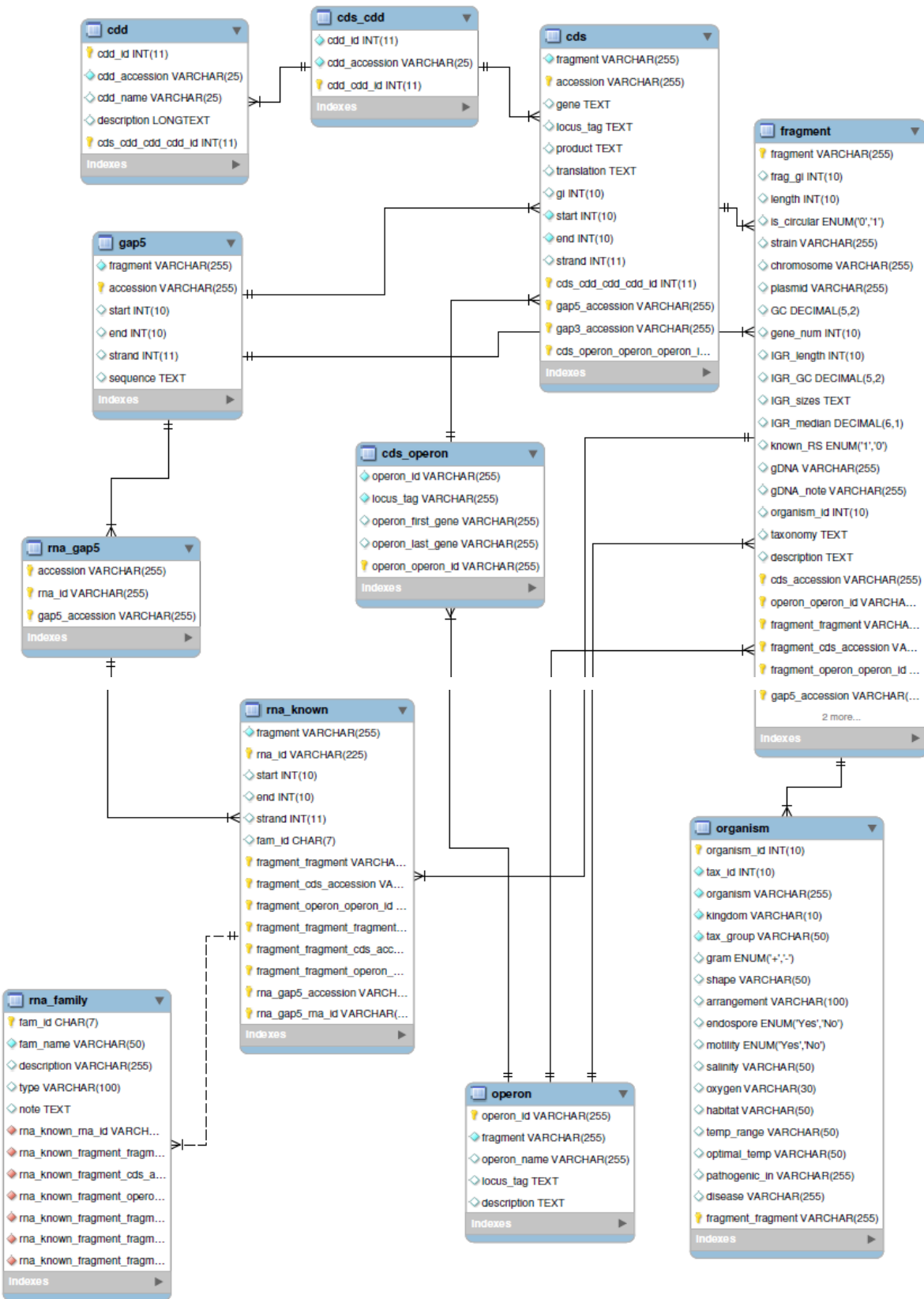
- Herbert KG, Spirollari J, Wang JT, Wang JT, Piel WH, Westbrook J, Barker WC, Hu ZZ & Wu CH (2007) Bioinformatic databases. *Wiley Encyclopedia of Computer Science and Engineering*.
- Heyne S, Costa F, Rose D & Backofen R (2012) GraphClust: alignment-free structural clustering of local RNA secondary structures. *Bioinformatics* 28(12):i224-i232.
- Huang Y, Niu B, Gao Y, Fu L & Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26(5):680-682.
- Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M & Sigrist CJ (2006) The PROSITE database. *Nucleic acids research* 34(suppl_1):D227-D230.
- Janssen S & Giegerich R (2015) The RNA shapes studio. *Bioinformatics* 31(3):423-425.
- Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S & Madden TL (2008) NCBI BLAST: a better web interface. *Nucleic acids research* 36(suppl_2):W5-W9.
- Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A & Nuka G (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9):1236-1240.
- Katoh K & Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics* 9(4):286-298.
- Krieg NR (2005) Identification of procaryotes. *Bergey's Manual® of Systematic Bacteriology*, Springer. p 33-38.
- Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP & Bork P (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic acids research* 30(1):242-244.
- Lewis TE, Sillitoe I, Dawson N, Lam SD, Clarke T, Lee D, Orengo C & Lees J (2018) Gene3D: extensive prediction of globular domains in proteins. *Nucleic acids research* 46(D1):D435-D439.
- Lipps HJ & Rhodes D (2009) G-quadruplex structures: in vivo evidence and function. *Trends in cell biology* 19(8):414-422.
- Lowe TM PP, Chan P. (TRNAscan-SE Search Server.
- Machtel P, Bąkowska-Żywicka K & Żywicki M (2016) Emerging applications of riboswitches—from antibacterial targets to molecular tools. *Journal of applied genetics* 57(4):531-541.
- Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY & Bryant SH (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic acids research* 30(1):281-283.
- Mattick JS (2005) The functional genomics of noncoding RNA. *Science* 309(5740):1527-1528.
- Mi H, Guo N, Kejariwal A & Thomas PD (2007) PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic acids research* 35(suppl_1):D247-D252.
- Miladi M, Sokhoyan E, Houwaart T, Heyne S, Costa F, Grüning B & Backofen R (2019) GraphClust2: annotation and discovery of structured RNAs with scalable and accessible integrative clustering. *GigaScience* 8(12):giz150.

- Nadiras C, Eveno E, Schwartz A, Figueroa-Bossi N & Boudvillain M (2018) A multivariate prediction model for Rho-dependent termination of transcription. *Nucleic acids research* 46(16):8245-8260.
- 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.
- 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.
- Nikolskaya AN, Arighi CN, Huang H, Barker WC & Wu CH (2006) PIRSF family classification system for protein functional and evolutionary analysis. *Evolutionary Bioinformatics* 2:117693430600200033.
- Noller HF (1984) Structure of ribosomal RNA. *Annual review of biochemistry* 53(1):119-162.
- Okuda S & Yoshizawa AC (2010) ODB: a database for operon organizations, 2011 update. *Nucleic acids research* 39(suppl_1):D552-D555.
- Pandurangan AP, Stahlhacke J, Oates ME, Smithers B & Gough J (2019) The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. *Nucleic acids research* 47(D1):D490-D494.
- Pruitt KD, Tatusova T & Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* 35(suppl_1):D61-D65.
- Regulski EE & Breaker RR (2008) In-line probing analysis of riboswitches. *Post-transcriptional gene regulation*, Springer. p 53-67.
- Reiss CW, Xiong Y & Strobel SA (2017) Structural basis for ligand binding to the guanidine-I riboswitch. *Structure* 25(1):195-202.
- Rhodes D & Lipps HJ (2015) G-quadruplexes and their regulatory roles in biology. *Nucleic acids research* 43(18):8627-8637.
- Rivas E, Clements J & Eddy SR (2017) A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature methods* 14(1):45-48.
- Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL, Wolf YI, Yin J & Koonin EV (2002) Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic acids research* 30(19):4264-4271.
- Rouleau S, Glouzon J-PS, Brumwell A, Bisailon M & Perreault J-P (2017) 3' UTR G-quadruplexes regulate miRNA binding. *Rna* 23(8):1172-1179.
- Rouleau SG, Garant J-M, Bolduc F, Bisailon M & Perreault J-P (2018) G-Quadruplexes influence pri-microRNA processing. *RNA biology* 15(2):198-206.
- Salamov VSA & Solovyevand A (2011) Automatic annotation of microbial genomes and metagenomic sequences. *Metagenomics and its applications in agriculture, biomedicine and environmental studies*. Hauppauge: Nova Science Publishers :61-78.
- Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, Connor R, Fiorini N, Funk K & Hefferon T (2019) Database resources of the national center for biotechnology information. *Nucleic acids research* 47(Database issue):D23.

- Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M & Federhen S (2010) Database resources of the national center for biotechnology information. *Nucleic acids research* 39(suppl_1):D38-D51.
- Schuler GD, Epstein JA, Ohkawa H & Kans JA (1996) [10] Entrez: Molecular biology database and retrieval system. *Methods in enzymology* 266:141-162.
- Shahmuradov IA, Mohamad Razali R, Bougouffa S, Radovanovic A & Bajic VB (2017) bTSSfinder: a novel tool for the prediction of promoters in cyanobacteria and Escherichia coli. *Bioinformatics* 33(3):334-340.
- Sherlock ME, Sudarsan N & Breaker RR (2018) Riboswitches for the alarmone ppGpp expand the collection of RNA-based signaling systems. *Proceedings of the National Academy of Sciences* 115(23):6052-6057.
- Shimoni Y, Friedlander G, Hetzroni G, Niv G, Altuvia S, Biham O & Margalit H (2007) Regulation of gene expression by small non-coding RNAs: a quantitative view. *Molecular systems biology* 3(1):138.
- Smale ST & Kadonaga JT (2003) The RNA polymerase II core promoter. *Annual review of biochemistry* 72(1):449-479.
- Sobell MG (2013) *A practical guide to Linux commands, editors, and shell programming*. Prentice Hall,
- Sudarsan N, Lee E, Weinberg Z, Moy R, Kim J, Link K & Breaker R (2008) Riboswitches in eubacteria sense the second messenger cyclic di-GMP. *Science* 321(5887):411-413.
- Superson AA, Phelan D, Dekovich A & Battistuzzi FU (2018) Using taxon resampling to identify species with contrasting phylogenetic signals: an empirical example in Terrabacteria. *BioRxiv* :369264.
- Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M & Ostell J (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic acids research* 44(14):6614-6624.
- Tran TT, Zhou F, Marshburn S, Stead M, Kushner SR & Xu Y (2009) De novo computational prediction of non-coding RNA genes in prokaryotic genomes. *Bioinformatics* 25(22):2897-2905.
- Tropp BE (2012) *Principles of molecular biology*. Jones & Bartlett Publishers,
- Vannutelli A, Belhamiti S, Garant J-M, Ouangraoua A & Perreault J-P (2020) Where are G-quadruplexes located in the human transcriptome? *NAR Genomics and Bioinformatics* 2(2):lqaa035.
- 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.
- 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.

10 ANNEXE I : DIAGRAMME COMPLET DE RIBOGAP

Le diagramme complet de la base de données RiboGap v1. Tiré de (Naghdí *et al.*, 2017)



11 ANNEXE II : LES GROUPES (*CLUSTERS*) RESULTATS

11.1 RiboGap v1

L'exécution de GraphClust avec les données de RiboGap v1 a permis d'avoir 12 *clusters*. Une fois que nous avons vérifié la conservation et la co-variation dans chaque cluster, quatre *clusters* ont été sélectionnés comme étant des candidats potentiels. Les quatre *clusters* présentés ici ont tous des caractéristiques suggérant une probabilité relativement élevée d'être des structures d'ARN conservées. Chacun sera brièvement commenté, suivi d'une figure récapitulative.

11.1.1 Cluster 11

Toutes les séquences du Cluster11 sont très conservées dans l'ensemble, avec une majorité de bases 100% conservées. Néanmoins, il y a plusieurs variations de séquences observées dans des régions prédites pour former des paires de base (pb) qui montrent des variations compatibles avec une pb (ex. G-C vs G-U) ou même des co-variations (ex. G-C vs A-U) (en jaune).

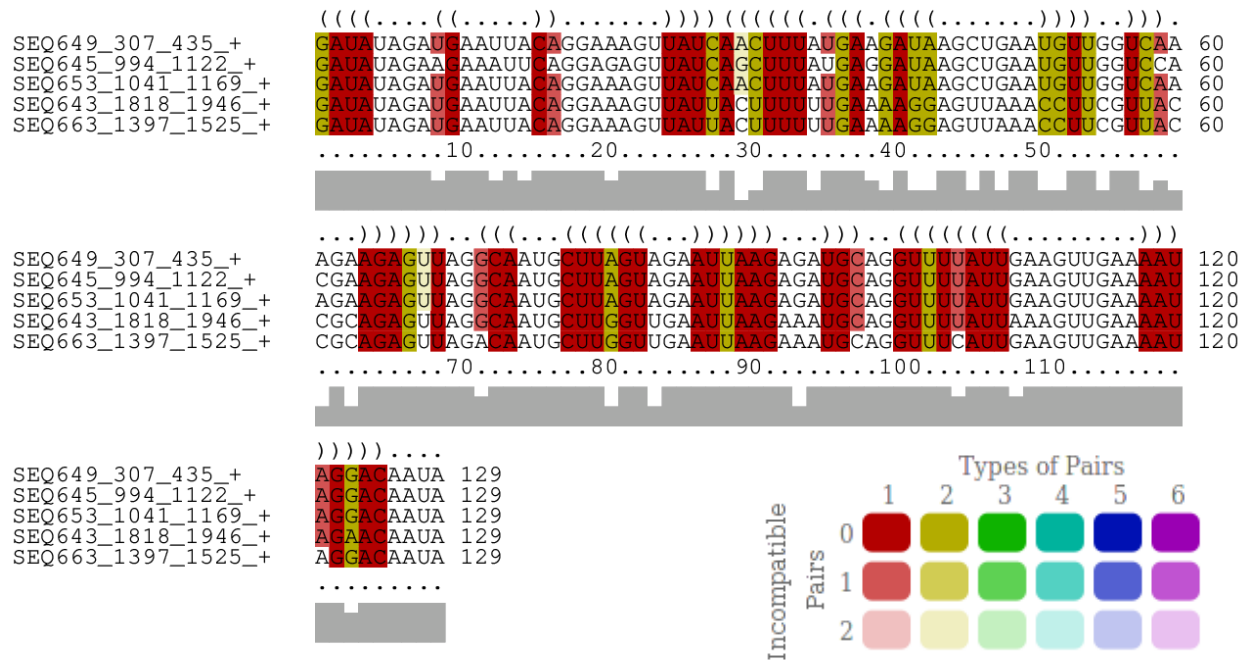


Figure 24 - Alignement structurel du Cluster 11

Les parenthèses = paires de base

Ombre grise = conservation

La légende utilisée dans cette figure s'applique à toutes les figures d'alignement des prochains *clusters*.

11.1.2 Cluster 1

Cluster1, il y a deux régions apparemment très conservées (7-43 et 60-89) (avec quelques variations compatibles avec les pb, mais aucune co-variation). Cependant, une 3^e région (152-180) avec une séquence beaucoup moins conservée semble néanmoins avoir une structure conservée, quoique flexible.

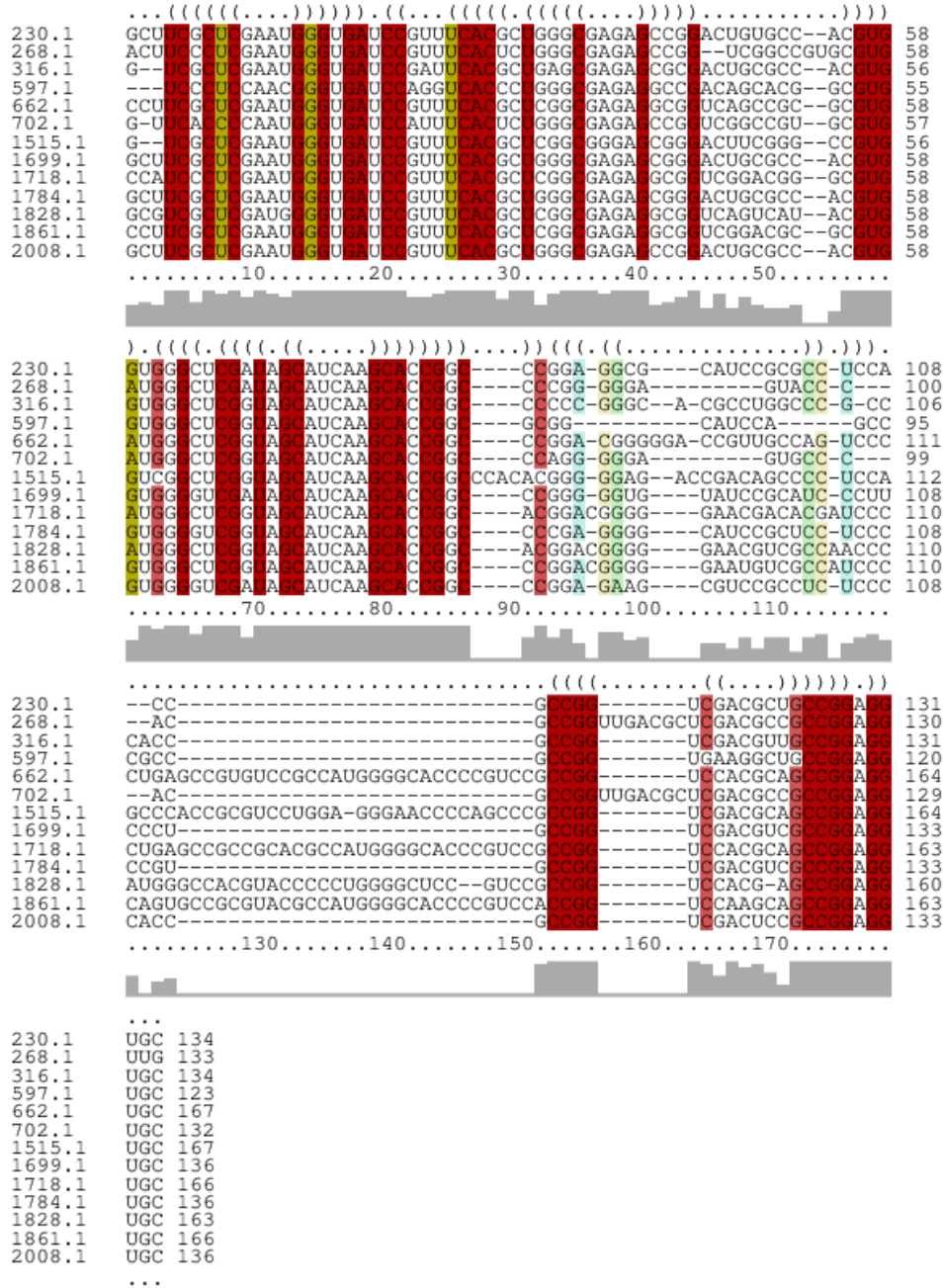


Figure 25 - Alignement structural du cluster 1

11.1.3 Cluster 8

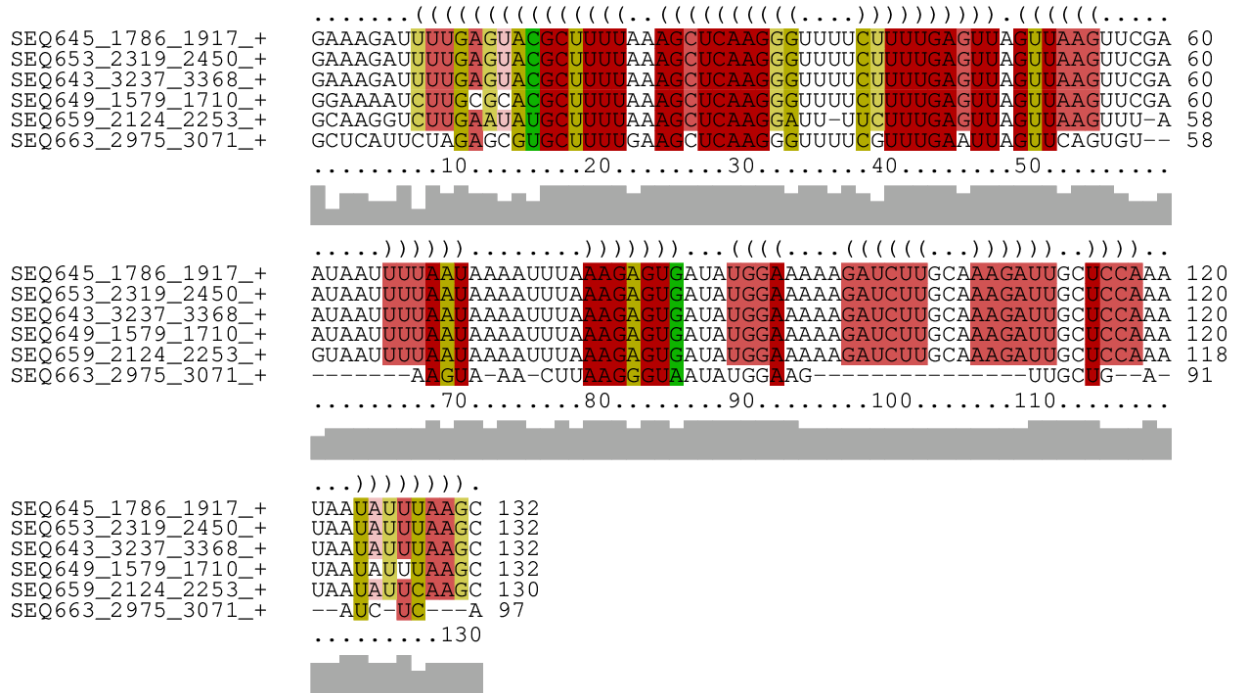


Figure 26 - Alignement structural du cluster 8

11.1.4 Cluster 14

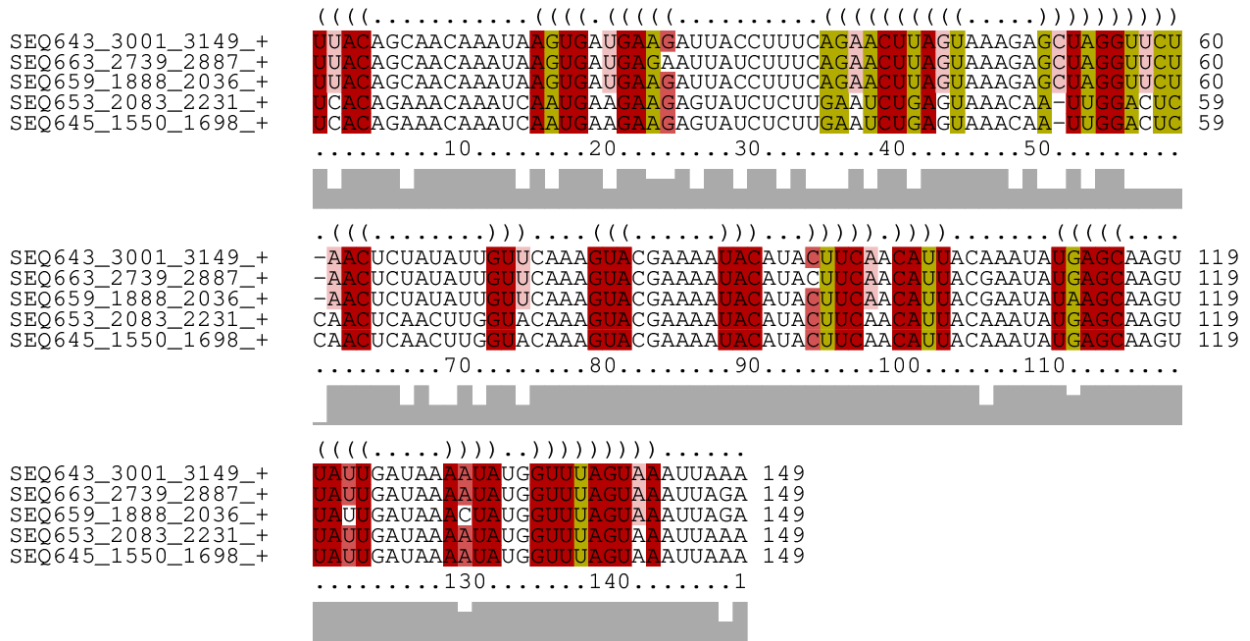


Figure 27–Alignement structurel du cluster 14

Les structures secondaires prédites pour les *clusters* 1, 8, 11 et 14 sont montrées dans la figure suivante.

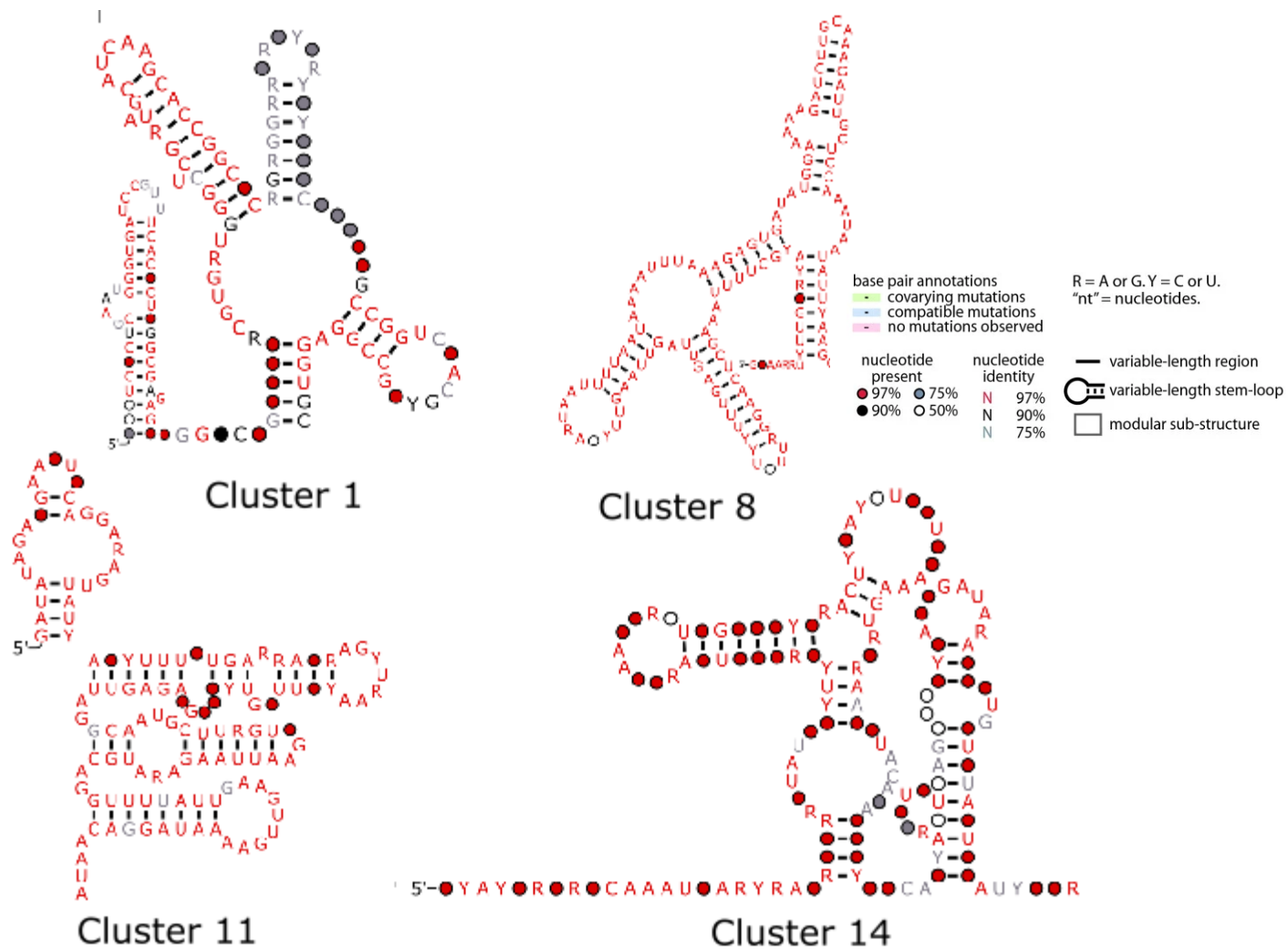
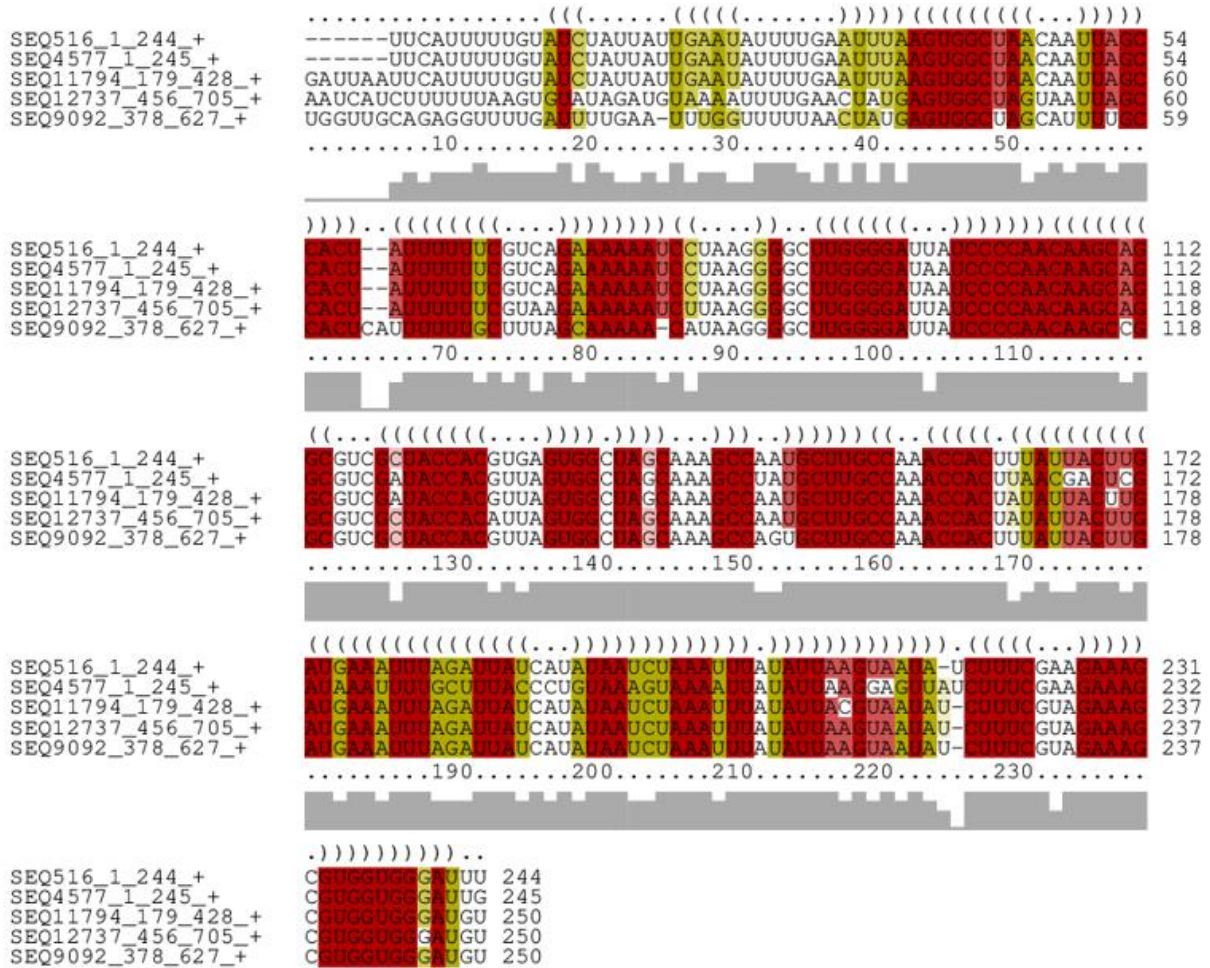


Figure 28 - Structures secondaires des *clusters* 1, 8, 11 et 14 générées avec R2R

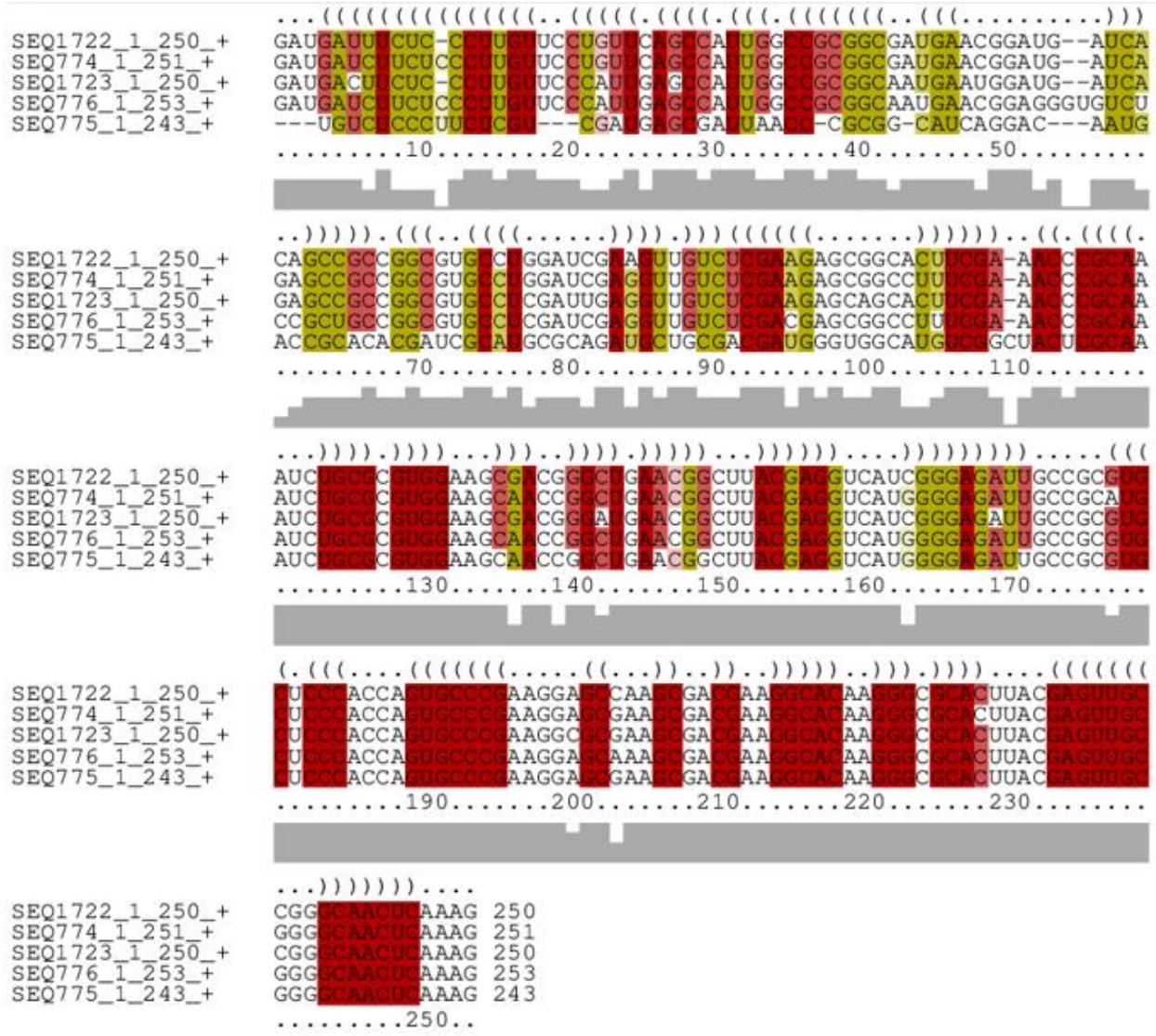
Cluster 15

Les différentes séquences ne sont pas entièrement alignées, mais l'existence de plusieurs régions conservées et de co-variations rend la probabilité assez élevée d'être des structures d'ARN conservées.



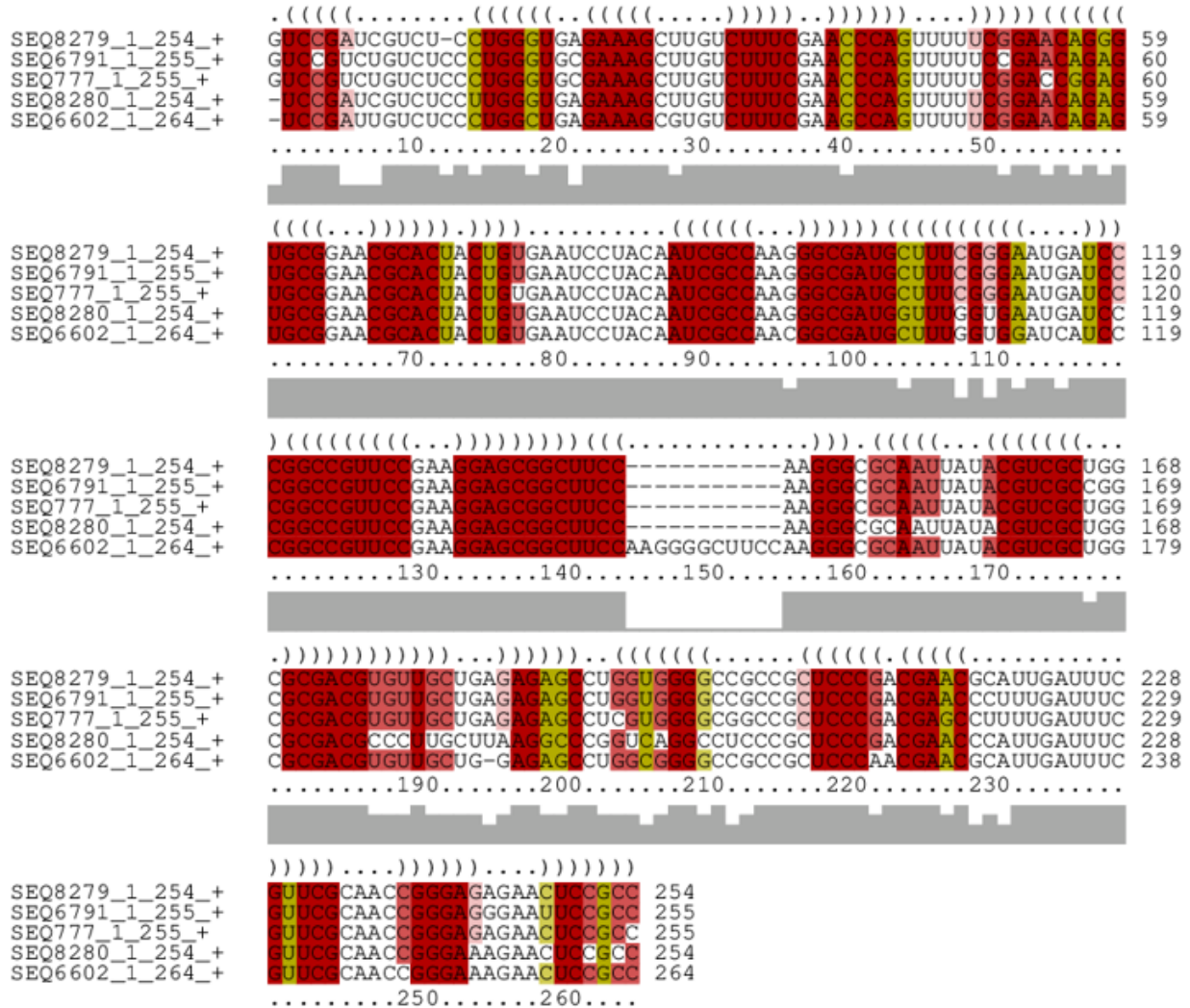
Cluster 19

Ce cluster montre beaucoup moins de conservations que les autres exemples. Mais la co-variation est très présente ce qui rend le cluster intéressant.



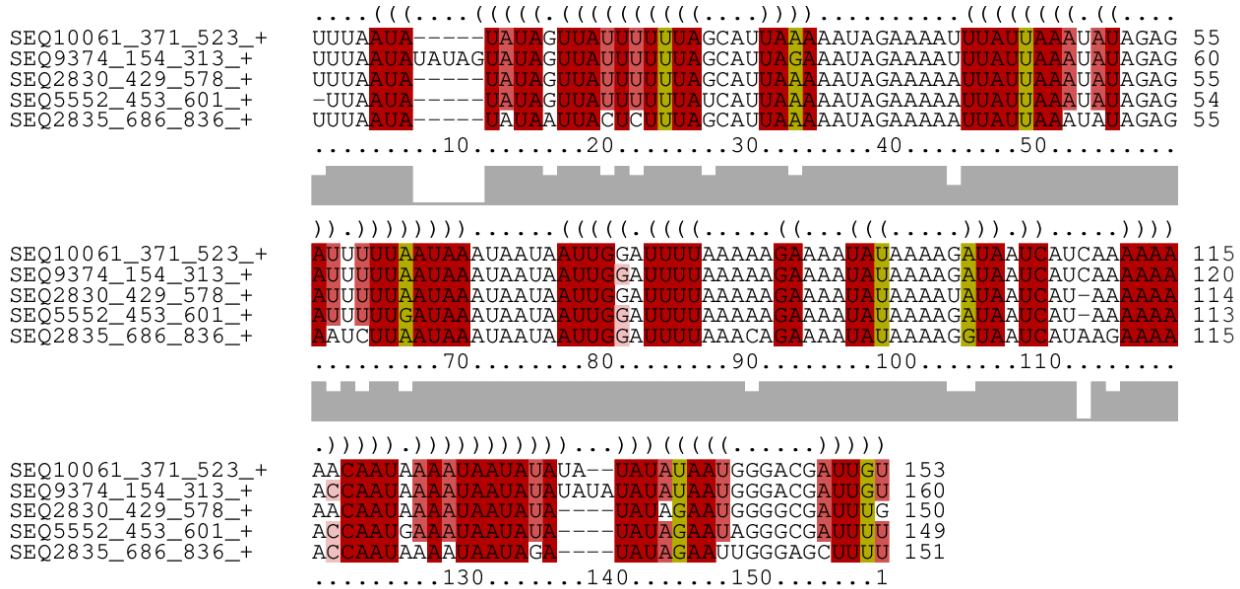
Cluster 20

Il est à noter que certaines prédictions peuvent parfois des portions où les structures prédites sont moins fiables, comme c'est le cas de la tige aux positions 164 à 168 et 188 à 192, mais plusieurs autres régions où les tiges prédites semblent avoir une haute probabilité de se former et où plusieurs pb montrent des variations compatibles supportant la structure prédite, comme par exemple la tige trouvée entre les positions 15 à 45 ci-dessous.

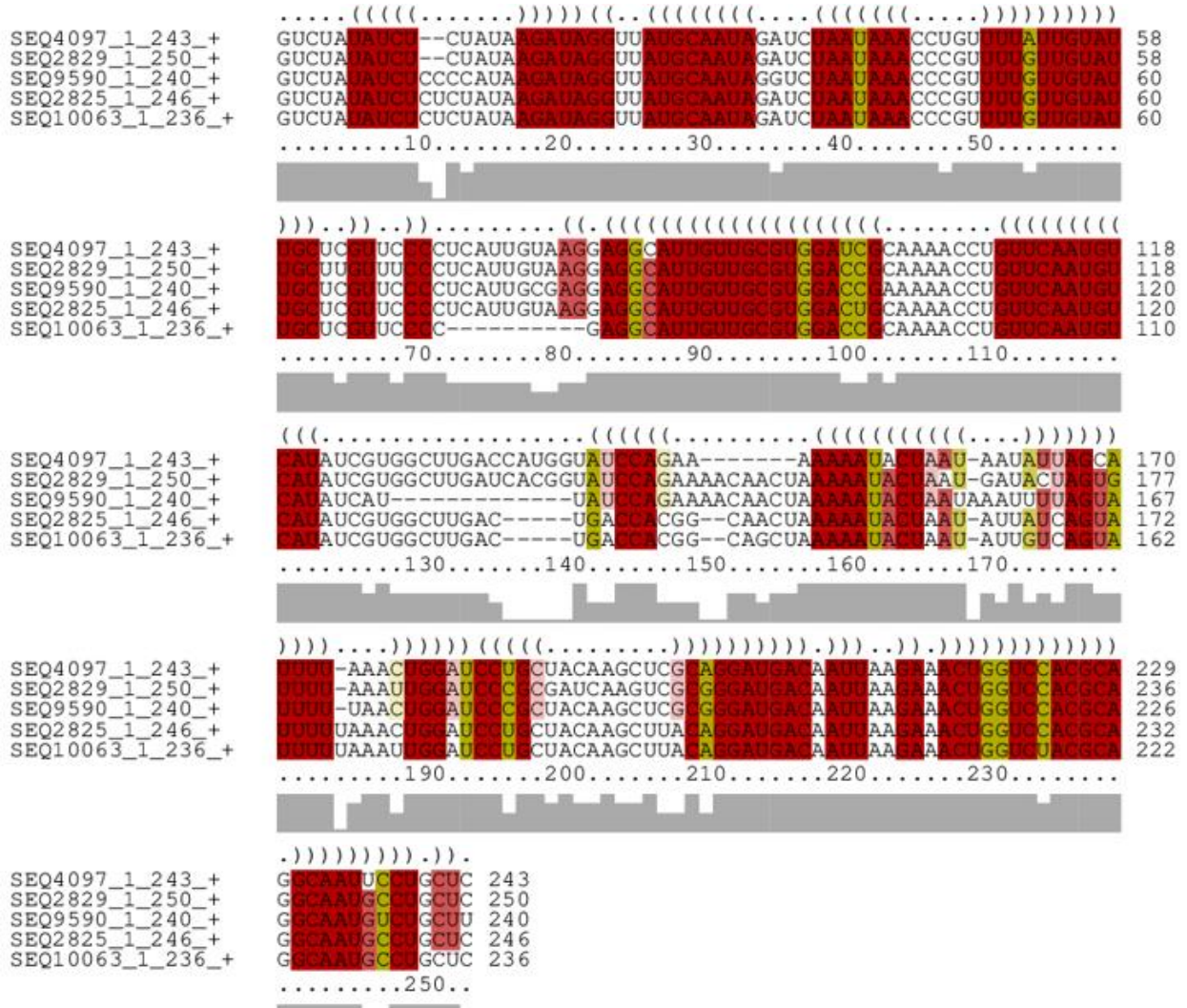


Cluster 26

Dans certains cas, une séquence trop simple, comme l'exemple du cluster 26, particulièrement riche en AU, peut plus facilement favoriser des compatibilités avec des structures multiples et peut donc mener à des alignements de structures que l'on doit considérer comme des faux-positifs.

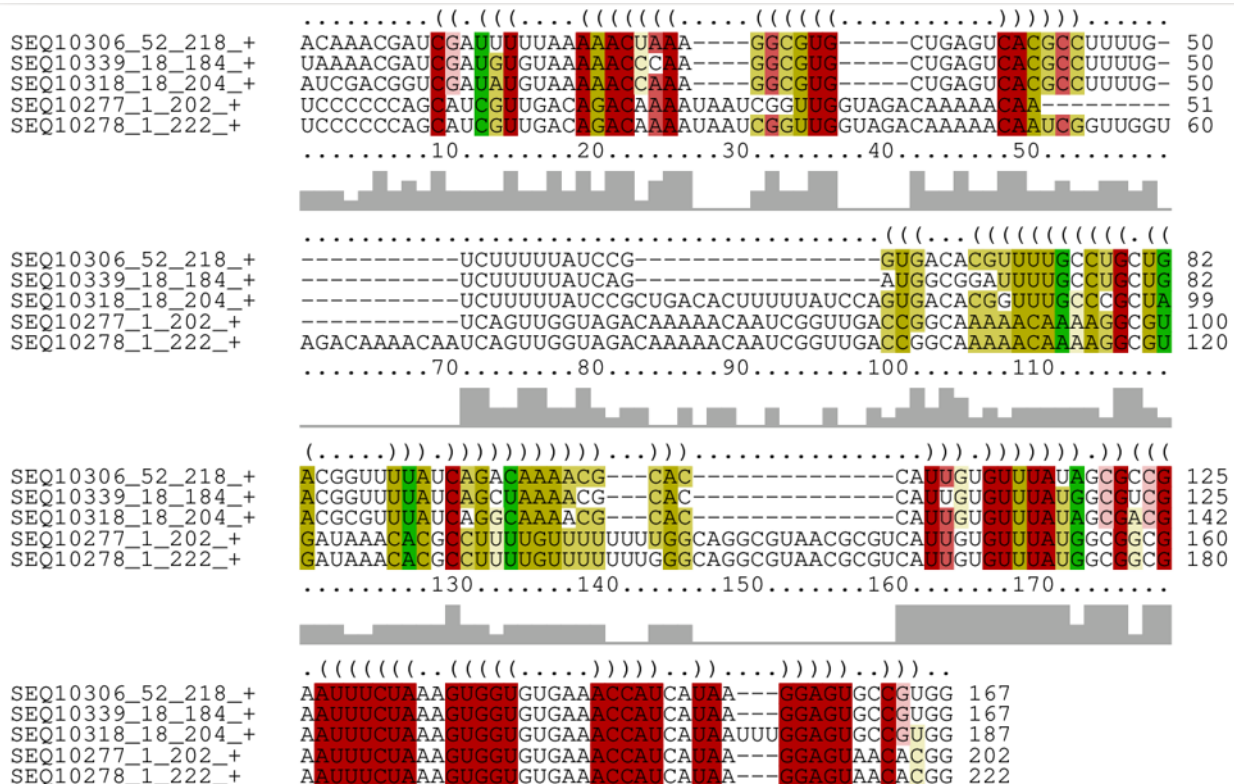


Cluster 27



Cluster 30

Selon toute vraisemblance, la partie de l'alignement qui montre le plus de co-variation contient un terminateur Rho-indépendant (avec les « U » des positions ~136 à 145 pour les deux dernières séquences et des positions ~166 à 171 pour les autres constituant la fin de ce terminateur).



Cluster 32

.....((((((((.....)))))).....((((((((.....)))))).....

SEQ10890_1_114_+ UGCUAACGGCGCGGAAUCGUCGCGCCAUAUUCCUCCUUUUCGCGCAUUCUUCUUUCCA 60
SEQ9081_1_114_+ UGCUAACGGCGCGGAAUCGUCGCGCCAUAUUCCUCUUUUCGCGCAUUCUUCUUUCCA 60
SEQ9400_1_114_+ UGCUAACGGCGCGGAAUCGUCGCGCCAUAUUCCUCCUGUUCGCGCAUUCUUCUUUCCA 60
SEQ9961_1_114_+ UGCUAACGGCGCGGAAUCGUCGCGCCAUAUUCCUCUUUUCGCGCAUUCUUCUUUCCA 60
SEQ11236_1_114_+ UGCUAACGGCGCGGAAUCGUCGCGCCAUAUUCCUCCUUUUCGCAUUCUUCUUUCCA 60

.....10.....20.....30.....40.....50.....

)).....((((((((.....)))))).....((((((((.....)))))).....

SEQ10890_1_114_+ CGUCCUAUUCGUCUUJGGJUAUAGUGUUUUCAUCAUAAAAGCAGGAGAACACA 114
SEQ9081_1_114_+ CGUCCUAUUCGUCUUJGGJUAUAGUGUUUUCAUCAUAAAAGCAGGAGAACACA 114
SEQ9400_1_114_+ CGUCAUUAUUCGUCUUJGGJUAUAGUGUUUUCAUCAUAAAAGCAGGAGAACACA 114
SEQ9961_1_114_+ CGUCAUUAUUCGUCUUJGGJUAUAGUGUUUUCAUCAUAAAAGCAGGAGAACACG 114
SEQ11236_1_114_+ UAUCCUGUCCGGCUUJGGJUAUAGUGUUUUCAUCAUAAAAGCAGGAGAACACG 114

.....70.....80.....90.....100.....110.....

Cluster 33

.....((((((((.....)))))).....((((((((.....)))))).....

SEQ5034_1_155_+ UCCGAAGAACAACGAGACAUACCAACGGCAAACCGGUGUGAUUUUAGGAGCAAGGUUUAU 60
SEQ204_1_157_+ UUAGUAGAACAGSCAAAGCAUACCAACGGCAAACCGGUAUGAUUUUAGGAGCAAGGUUUAU 60
SEQ2313_1_156_+ CCUGCAAGGCUGSCGGCAUGAACAGACGGCAGGCUUGUGAUUUAGGAGCAAGGUUUAU 60
SEQ5035_1_147_+ UCCG-----GCAGGAU-UACCAACGGCAGCGGUCUGAUUUUGG-AGAGCAAGGUUUAU 50
SEQ10988_1_146_+ UCUG-----GCAGGAU-UACCAACGGCAGCGGUUAUUUUGG-AGAGCAAGGUUUAU 50

.....10.....20.....30.....40.....50.....

(((.....)))))).....((((((((.....)))))).....

SEQ5034_1_155_+ GUUUCGGUAGACCGAAAC-ACCUUGC UUUGCUUAACGCCAAAGAA-AAUUUC CCGUUGG 118
SEQ204_1_157_+ GUUUCGGUAGACCGAAAC-ACCUUGC UUUGCUUAUCGCAAAGGGAAA-UUUC CCGUUGG 119
SEQ2313_1_156_+ GUUUCGGUAGACCGAAAC-ACCUUGC UUUGCUUGACAGCAAAGAAA-UUUC CCGUUGG 119
SEQ5035_1_147_+ GUUUCGGUAAAACCGAAACUACCUUGC UUUGCUUGACAGCAAAGAAA-UUUC CCGUUGG 110
SEQ10988_1_146_+ GUUUCGGUAAAACCGAAACUACCUUGC UUUGCUUGACAGCAAAGAAA-UUUC CCGUUGG 109

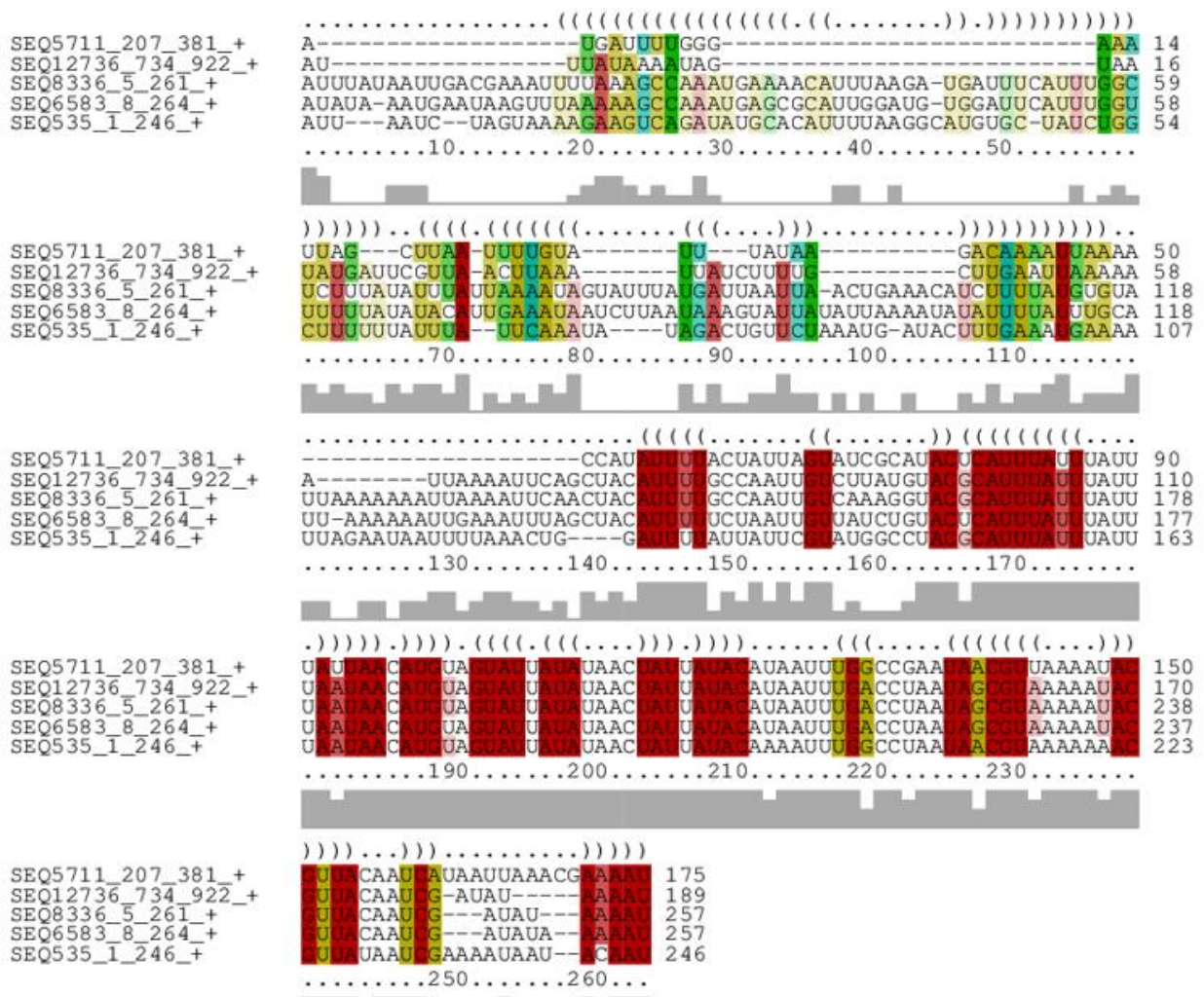
.....70.....80.....90.....100.....110.....

)).....((((((((.....)))))).....

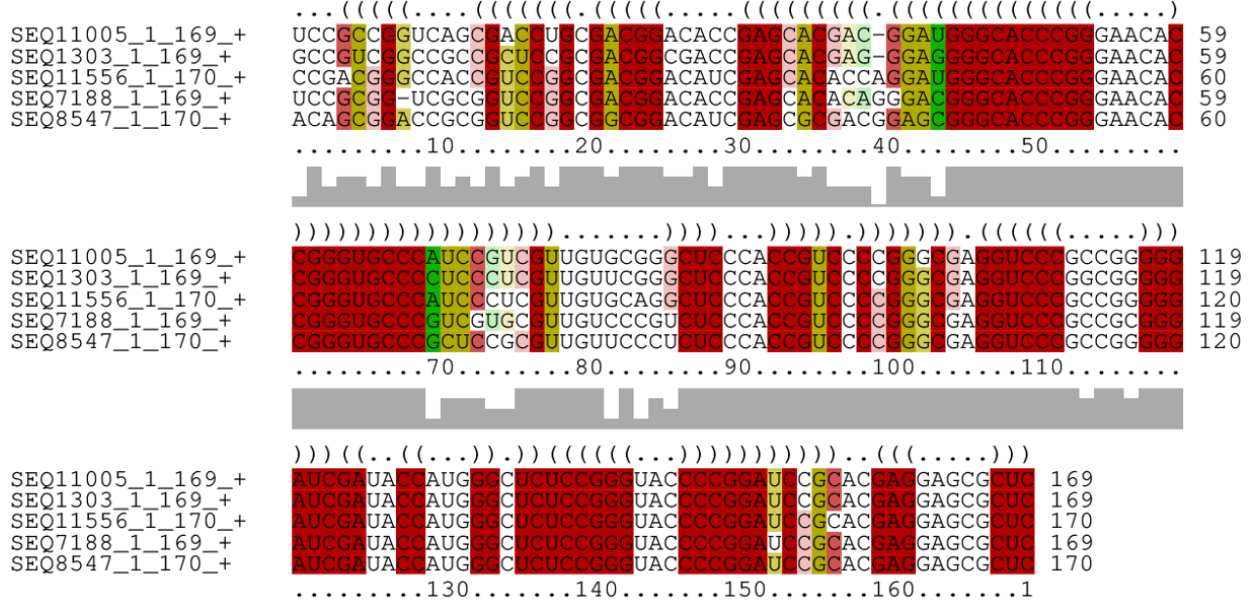
SEQ5034_1_155_+ UCCAAUUUUUAAGAUACCACCU-AAAAGCAAAAACAC 155
SEQ204_1_157_+ UCCAAUUUCUAAGAUACCACCUAAAAAGCAAAAACGU 157
SEQ2313_1_156_+ UCCAAUUUUUAAGAUACCACCUAAAA-GCAAAAACGU 156
SEQ5035_1_147_+ UCCAAUUUUUAAGAUACCAUUUAAA-UAAAAACCU 147
SEQ10988_1_146_+ UCCAAUUUUUAAGACACCAUUUAAA-GUAAAAACCU 146

Cluster 41

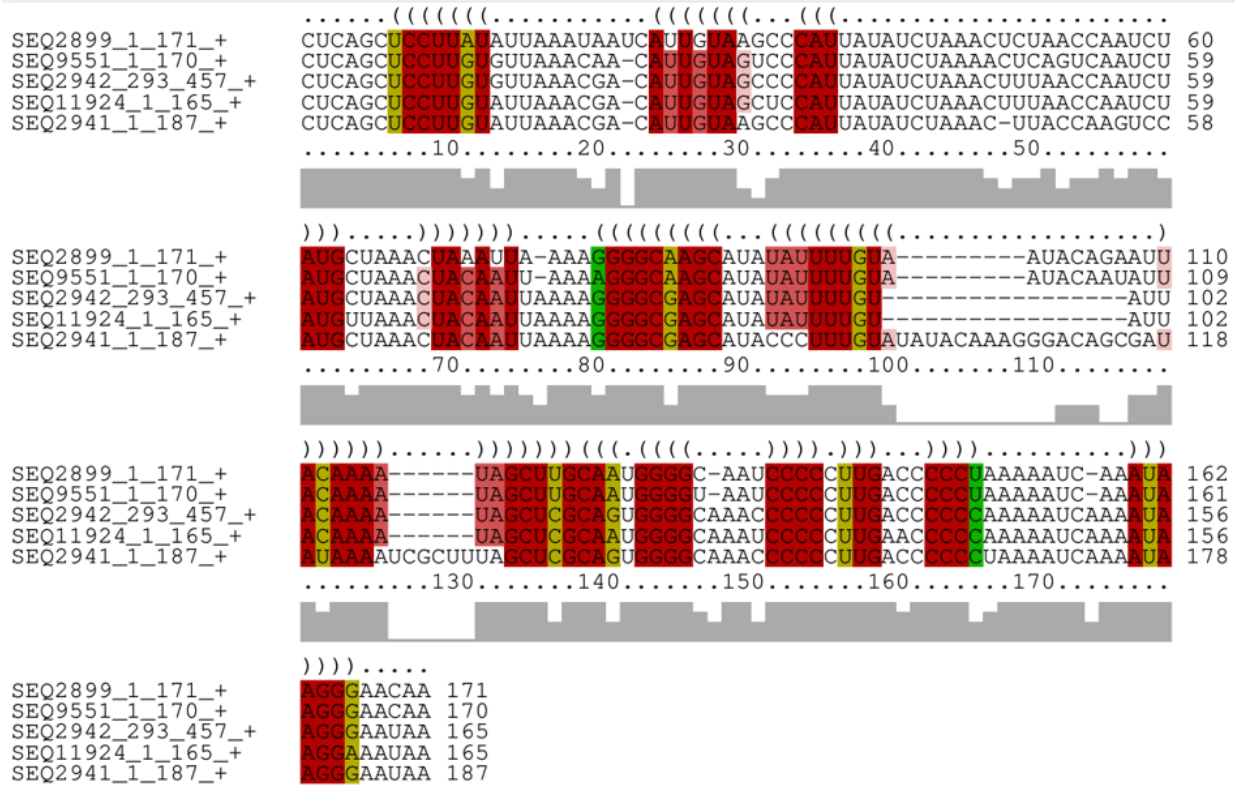
Ce cluster montre une potentielle grande variabilité et co-variation dans la première moitié de l'alignement et une conservation presque complète dans la deuxième moitié. Cette flagrante disparité peut laisser penser que seulement les régions IGR de la deuxième moitié s'alignent bien et que l'alignement de la première moitié est en quelque sorte « forcé » par GraphClust pour trouver une structure. Même si cette dernière hypothèse semble probable, le fait que les insertions (*gaps*) soient symétriques, relativement aux tiges, laissent croire que les deux tiges dans la première moitié de l'alignement pourraient être bien réelles.



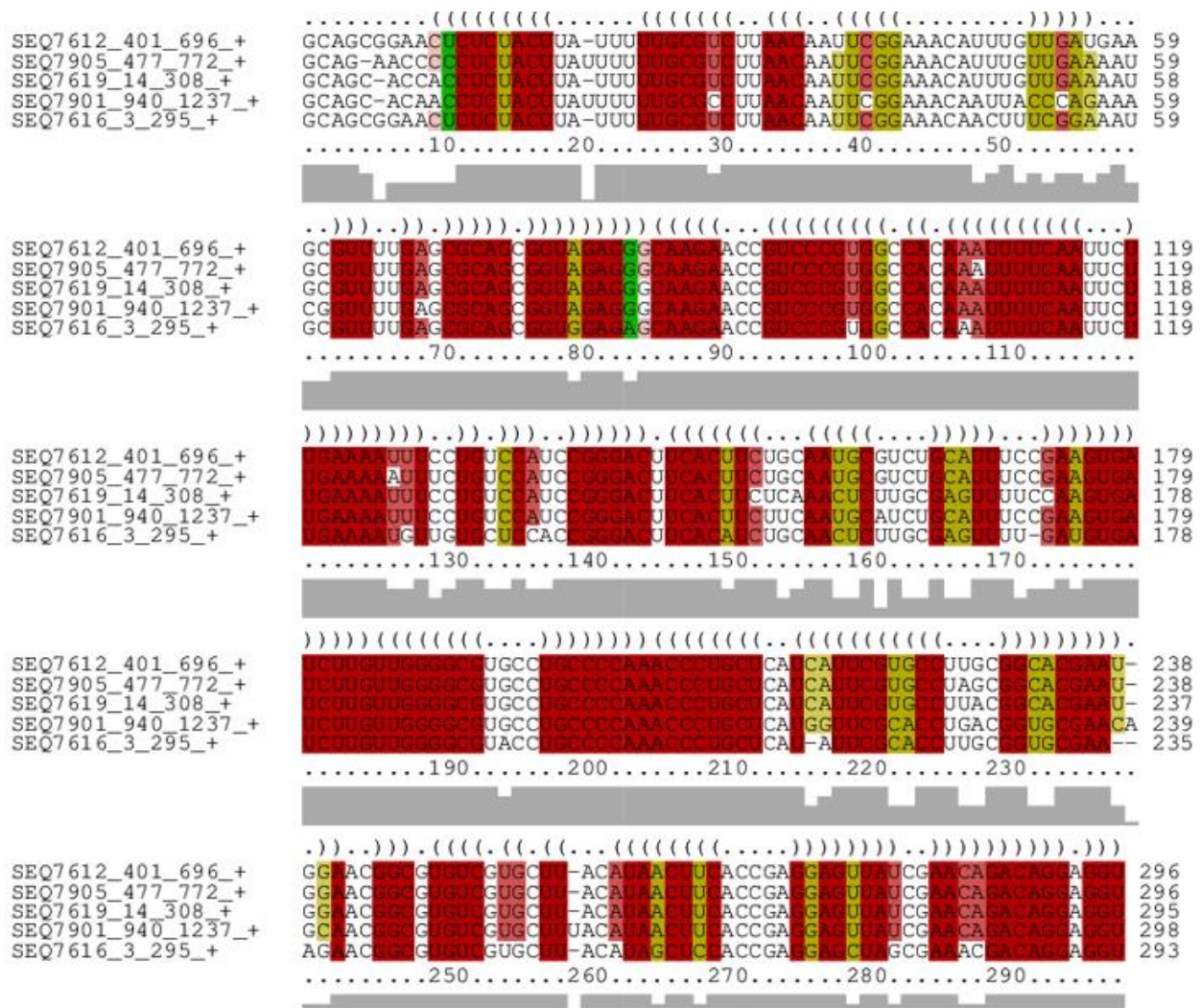
Cluster 45



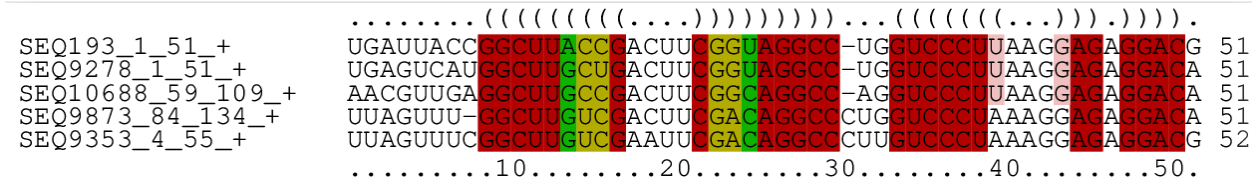
Cluster 47



Cluster 48



Cluster 49



Cluster 57

```

((.....(((.....)))).....)
SEQ11207_251_500_+    CCAACACCAGGGUUUUU GUAUACCGGCGGUGUAACCGAAACCCAAGGGUUUUGUAUGAAG 60
SEQ9808_503_752_+    CCAACACCAGGGUUUUU GUAUACCGGCGGUGUAACCGAAACCCAAGGGUUUUGUAUGAAG 60
SEQ9808_251_500_+    CCAACACCAGGGUUUUU GUAUACCGGCGGUGUAACCGAAACCCAAGGGUUUUGUAUGAAG 60
SEQ11207_17_248_+    AUUCCA-CAG-----AUAGUUC-----GCAACAGUGA-----GUCAAUUAUGAAG 42
SEQ9808_17_248_+    AUUCCA-CAG-----AUAGUUC-----GCAACAGUGA-----GUCAAUUAUGAAG 42
.....10.....20.....30.....40.....50.....

UAGGUUGGGUGUAACGUAAGGUAACCCAAGAACCAGGGUUUUGUAUAACGGGGUGGAAC 120
UAGGUUGGGUGUAACGGAGUGUAACCCAAGAACCAGGGUUUUGUAUAACGGGGUGGAAC 120
UAGGUUGGGUGUAACGGAGUGUAACCCAAGAACCAGGGUUUUGUAUAACGGGGUGGAAC 120
UAGGUUGGGUGUAACGUAAGGGAACCCAAGAACCAGGGUUUUGUAUAACGGGGUGGAAC 102
UAGGUUGGGUGUAACGUAAGGGAACCCAAGAACCAGGGUUUUGUAUAACGGGGUGGAAC 102
.....70.....80.....90.....100.....110.....

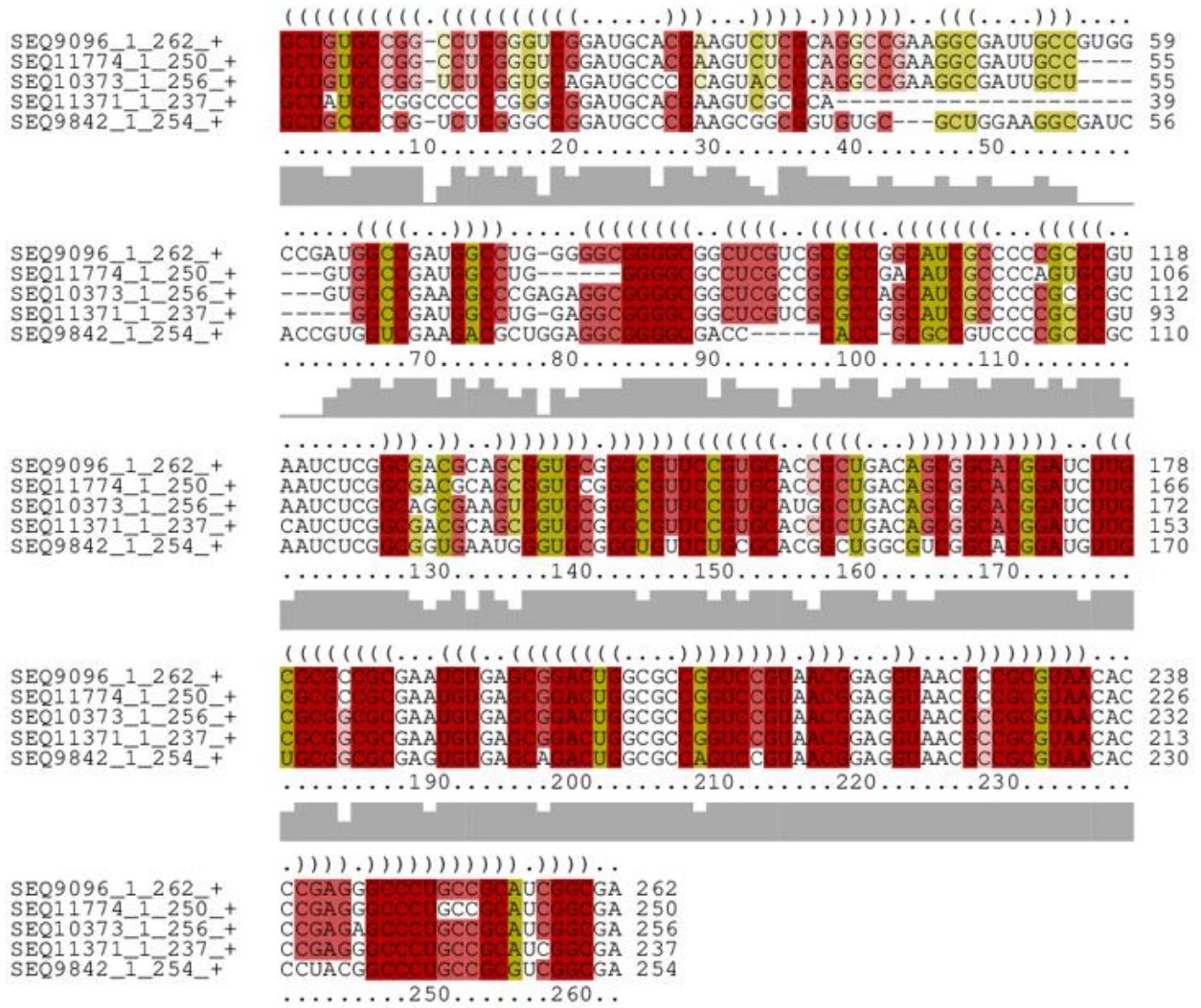
((.....(((.....)))).....)
SEQ11207_251_500_+    GAACACCAAGGGUUUGUAUGAAGUAGGUUGGGUGUAACGGAGGUGUAACCCAAACACCAAG 180
SEQ9808_503_752_+    GAACACCAAGGGUUUGUAUGAAGUAGGUUGGGUGUAACGGAGGUGUAACCCAAACACCAAG 180
SEQ9808_251_500_+    GAACACCAAGGGUUUGUAUGAAGUAGGUUGGGUGUAACGGAGGUGUAACCCAAACACCAAG 180
SEQ11207_17_248_+    GAACACCAAGGGUUUGUAUGAAGUAGGUUGGGUGUAACGGAGGUGUAACCCAAACACCAAG 162
SEQ9808_17_248_+    GAACACCAAGGGUUUGUAUGAAGUAGGUUGGGUGUAACGGAGGUGUAACCCAAACACCAAG 162
.....130.....140.....150.....160.....170.....

))))))...))..))..))..))..))..))..))..)).....((.....)
SEQ11207_251_500_+    GUUUUGUAUAAGGGCGGGGAACCGAACCAAGGGUUUGUAUAACGGCGGUUAACCGA 240
SEQ9808_503_752_+    GUUUUGUAUAAGGGCGGGGAACCGAACCAAGGGUUUGUAUAACGGCGGUUAACCGA 240
SEQ9808_251_500_+    GUUUUGUAUAAGGGCGGGGAACCGAACCAAGGGUUUUUUUAUGAAGUAGGUUGGGUGU 240
SEQ11207_17_248_+    GUUUUGUAUAAGGGCGGGUAACCGAACCAAGGGUUUGUAUGAAGUAGGUUGGGUGU 222
SEQ9808_17_248_+    GUUUUGUAUAAGGGCGGGUAACCGAACCAAGGGUUUGUAUGAAGUAGGUUGGGUGU 222
.....190.....200.....210.....220.....230.....

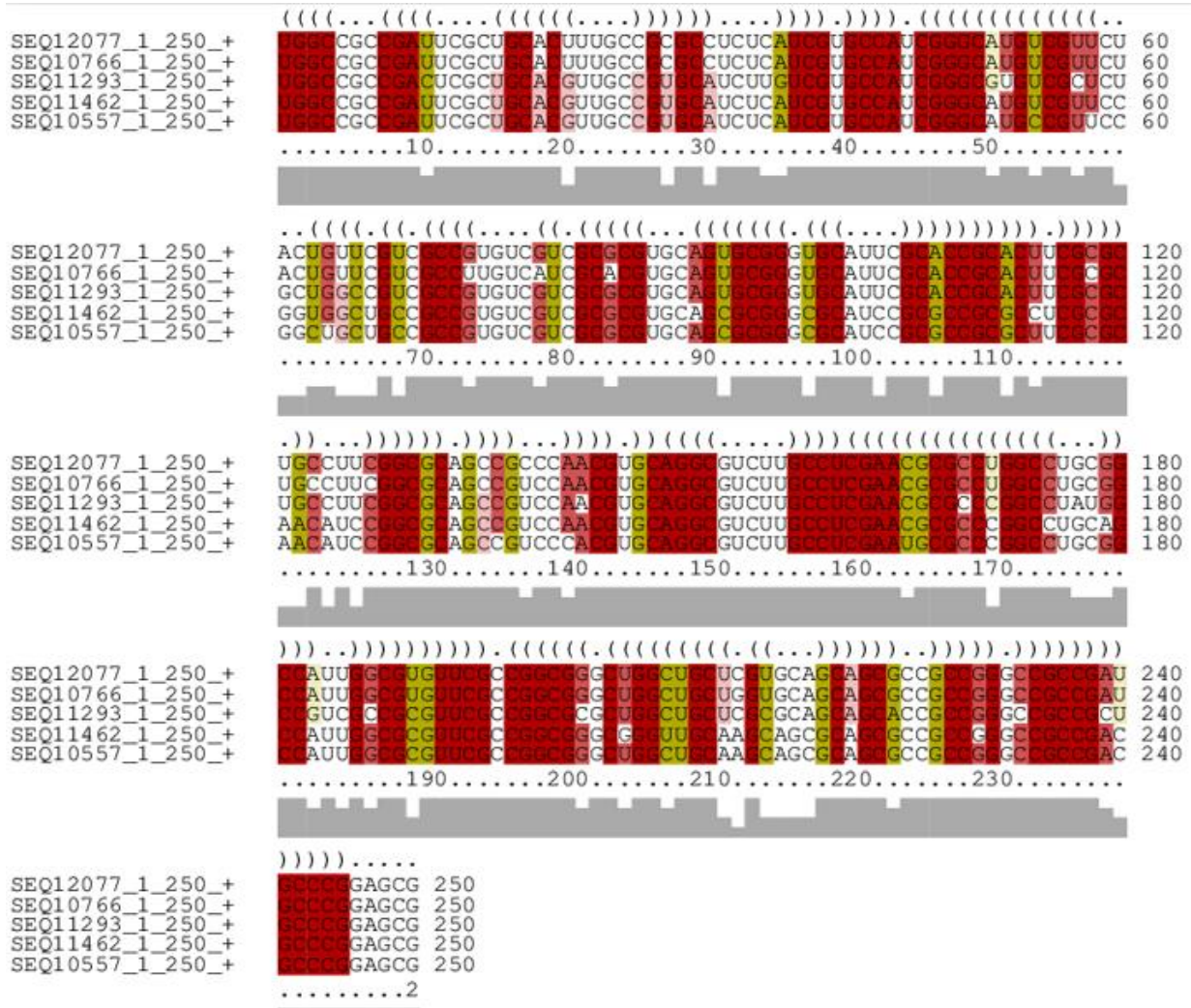
))))))..
SEQ11207_251_500_+    ACACCAGGG 250
SEQ9808_503_752_+    ACACCAGGG 250
SEQ9808_251_500_+    AAGGUAGUG 250
SEQ11207_17_248_+    AACGGAGUU 232
SEQ9808_17_248_+    AACGGAGUU 232
.....2
--  --  --

```

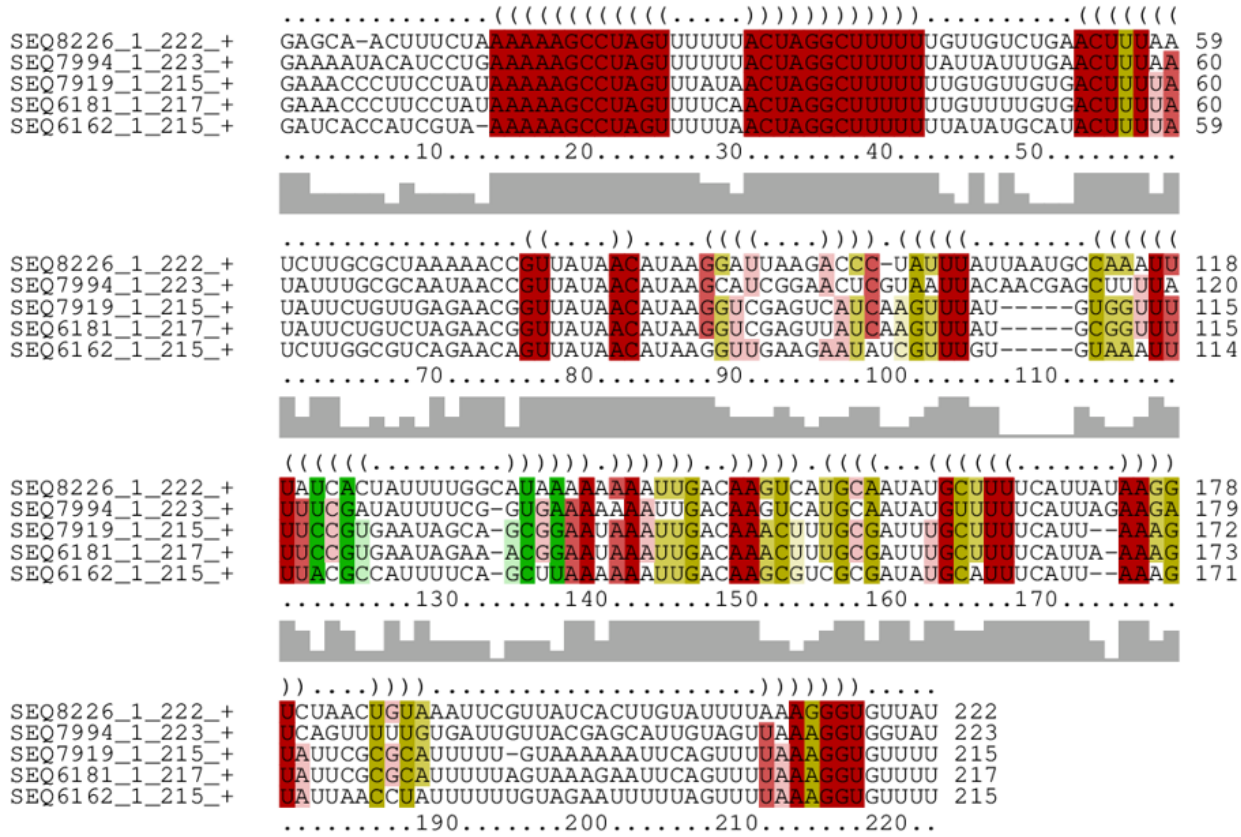
Cluster 58



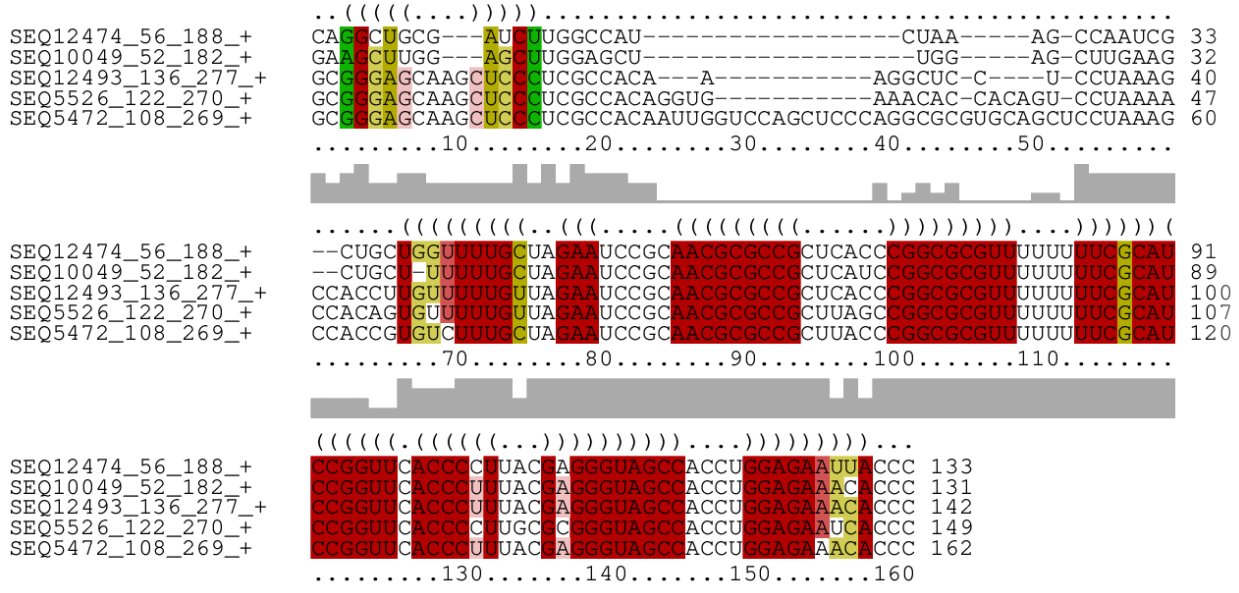
Cluster64



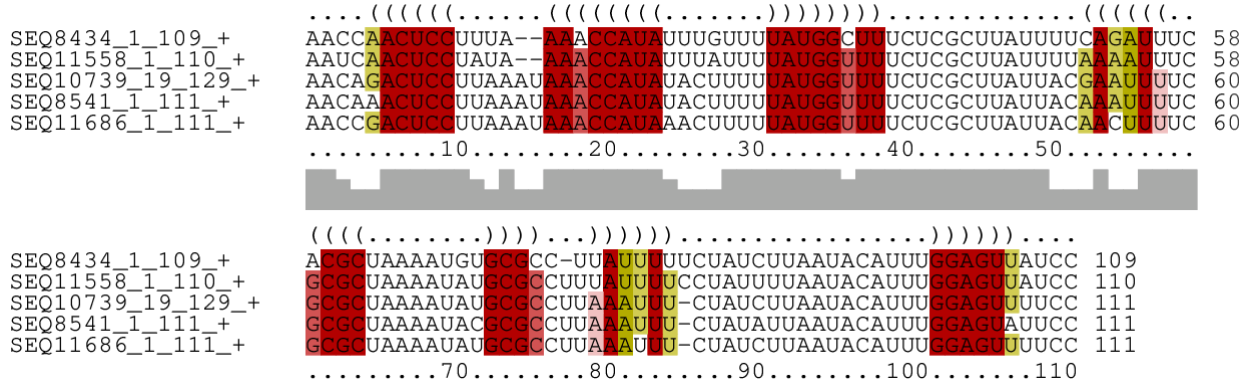
Cluster 68



Cluster 72

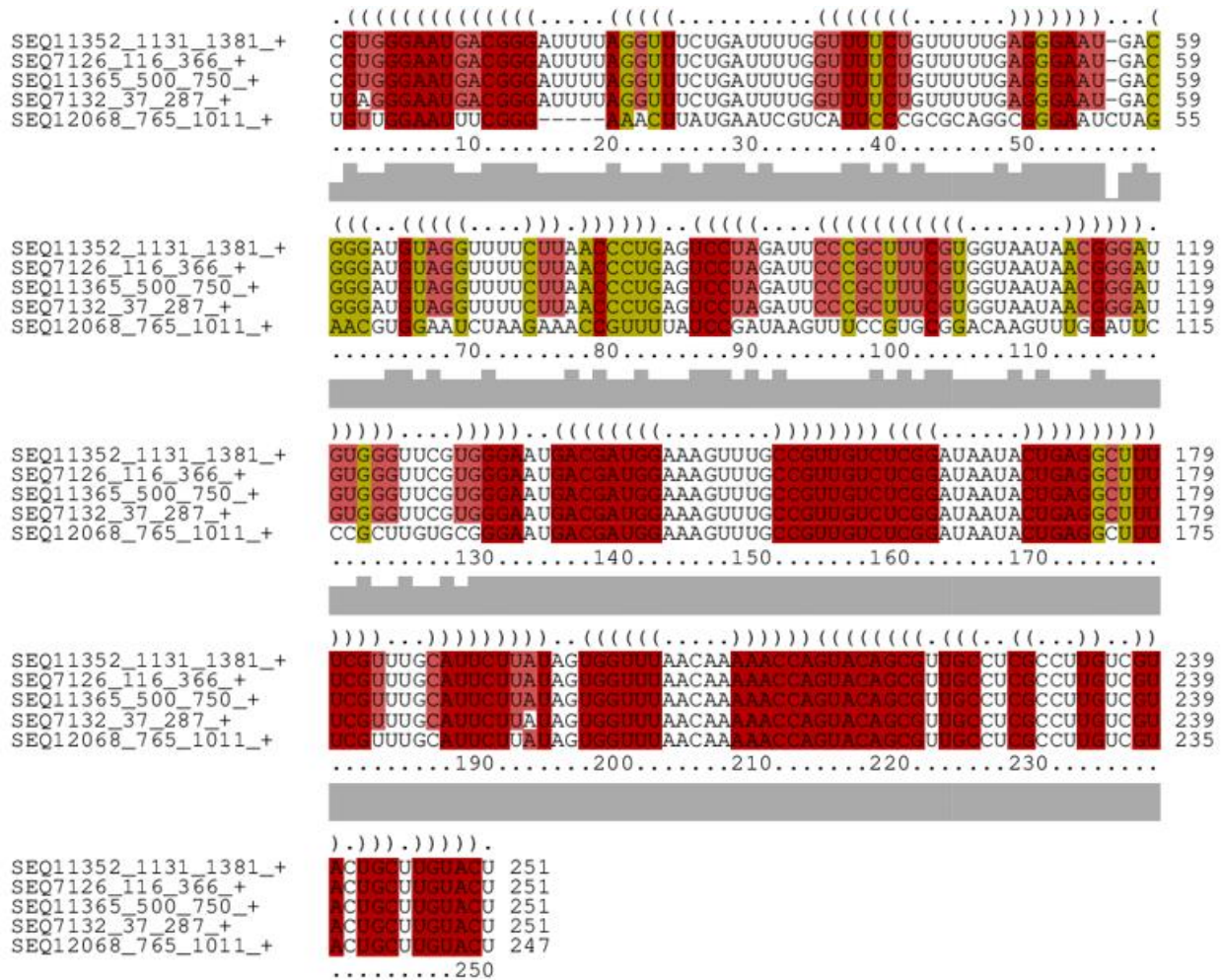


Cluster 74



Cluster 78

Ceci est un exemple où toute la soi-disant co-variation est due à une seule séquence (la dernière de l'alignement). Bien qu'avec la conservation élevée dans la 2^e moitié de la séquence il est évident que ces séquences soient apparentées, la première partie (en particulier des positions 20 à 130) de l'alignement montre une identité inférieure à 40% pour la dernière séquence, ce qui est très faible pour une séquence nucléotidique et suggère que ce soit une insertion ou autre forme de réarrangement local faisant en sorte que cette portion des séquences n'a pas une origine évolutive commune, ce qui rendrait la co-variation caduque.



Cluster 83

```

.....(((((((.(.(.(((.(.(((.(.((((.....(((((((.....((
SEQ9163_1_107_+ AAUUAUUUUAAUUCAUAAUCUUCUUGCGUUAGUUUAAGUUAAUUAUUAUUUUUUUA 60
SEQ9986_1_107_+ AAUUAUUUUAAUUCAUAAUCUUCUUGCAUUAUUUUUAAGUUAAUUAUUAUUUUUA 60
SEQ2818_1_108_+ AAUUAUUUUAAUUCAUAAUCUUCUUGCGUUAUUUUUAAGUCAUUAUUAUUUA-UUA 59
SEQ11019_1_106_+ AAUUAUUUUAAUUCAUAAUCUUCUUGCGUUAUUUUUAAGUGAUUAUUAUUUUUA 60
SEQ10799_1_102_+ AAUUAUUUUAAUUCAUAAUCUUCUUGCGUUAUUUUUAAGUUAAU-----UA 54
.....10.....20.....30.....40.....50.....

```

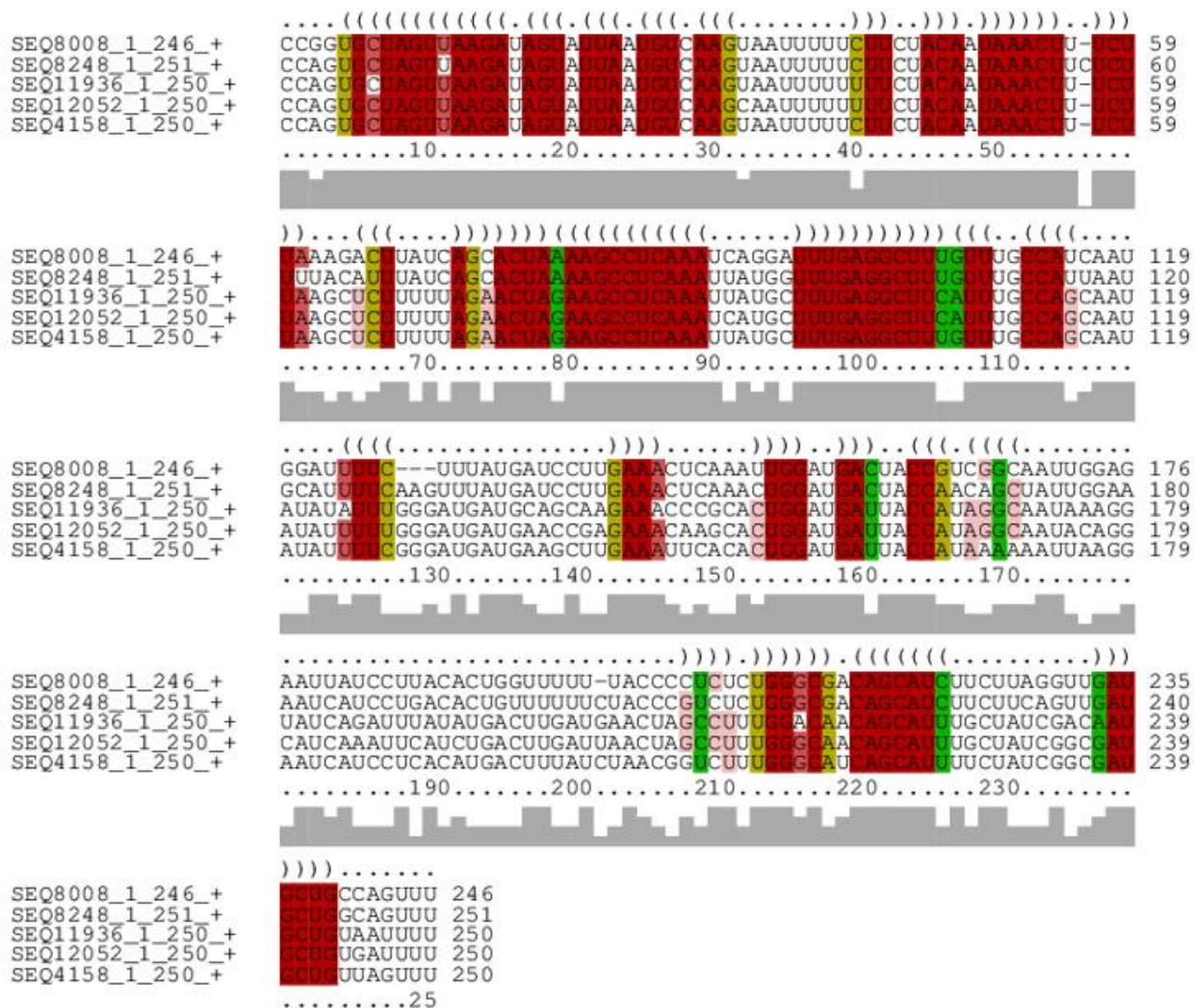


```

.....))))).....))))).....))))).....))))).....))))).....)))))
SEQ9163_1_107_+ AAAACCJUAUAA--ACAUAUAUACUAAAGCUAAUAUUGAGAUAUUUAU 107
SEQ9986_1_107_+ AAAACCJUAUAA--AUUAUAUAGUUAAGCUAAUAUUGAGAUAUUUAU 107
SEQ2818_1_108_+ AAAACCJUAUAAACAUAUAUAUAUUAAGCUAAUAUUGAGAUAUUUAU 108
SEQ11019_1_106_+ AAAACCJUAUU--AACUAUAUAUUAAGCUAAUAUUGAGAUAUUUAU 106
SEQ10799_1_102_+ AAAACCJUAUAA--ACAUAUAUUAUUAAGCUAAUAUUGAGAUAUUUAU 102
.....70.....80.....90.....100.....

```

Cluster 84



La figure suivante montre quelques exemples de structures secondaires intéressantes appartenant aux *clusters* mentionnés plus haut.

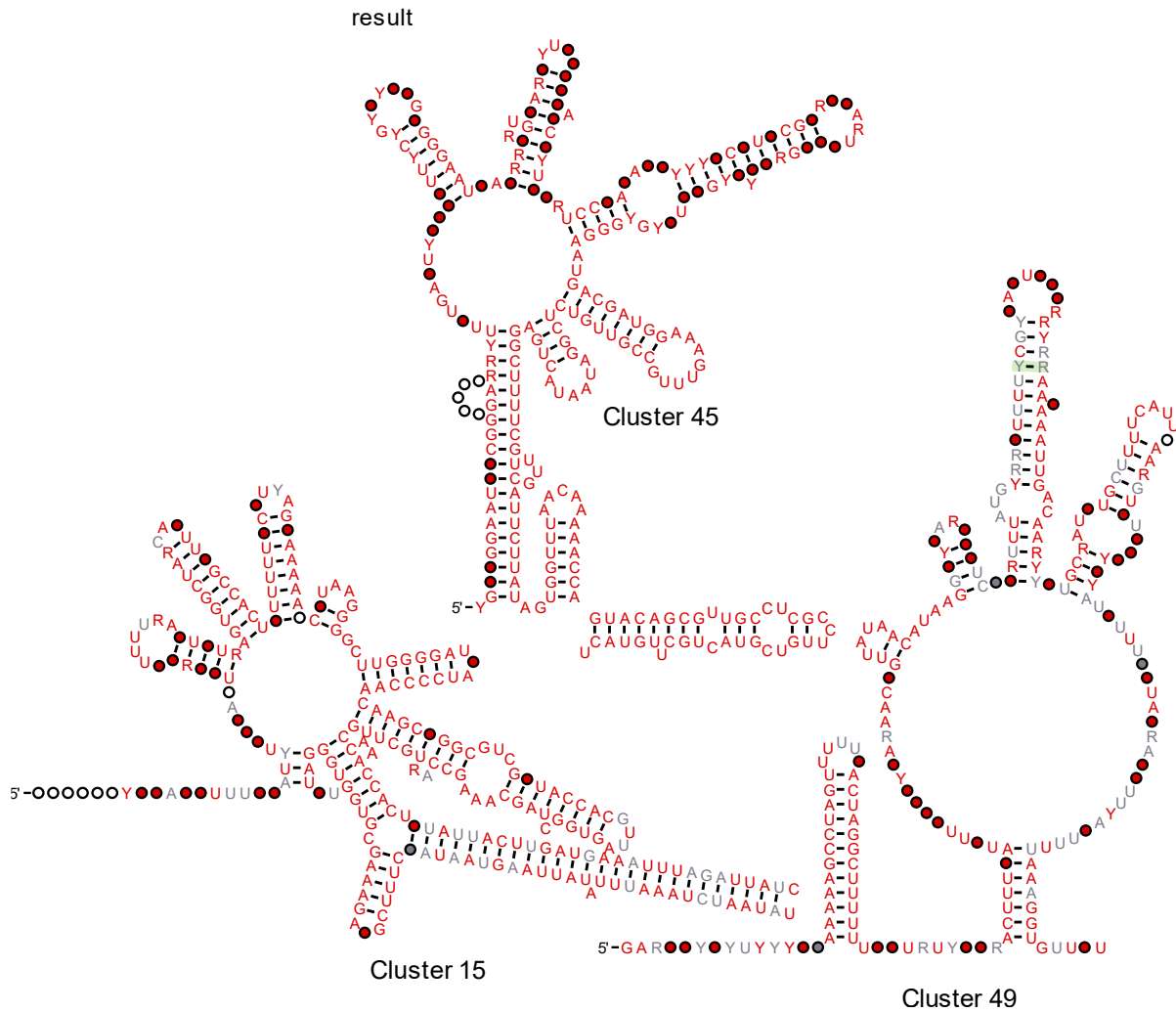


Figure 29 - Exemples de *clusters* intéressants pour la requête 1

11.2.1.2 Requête complexe

Pour rappel, la requête 2 correspond à la requête complète qui inclut tous les gènes connus régulés par le *riboswitch* ppGpp. Après avoir exécuté GraphClust sur l'ensemble des séquences, nous avons fait une sélection de cluster intéressant en analysant les alignements structuraux. La liste de ces *clusters* retenus sont cités ci-dessous :

Cluster 10

```
((.....)).....((((.....((((.....((.....(.....
SEQ12068_211_460_+ ACCAACAACUUAUUUUCUACGGUCUAGAGAGAAACGAAUUAUUGUUUGCCCAGCC 60
SEQ3931_209_458_+ ACCAACAACUUAUUUUCACGGUCUAGAGAGAAACGAAUUAUUGUUUGCCCAGCC 60
SEQ17216_217_465_+ ACCAACAACUUAUUUUCACGGUCUAGAGAGAAACGAAUUAUUGUUUGCCCAGCC 60
SEQ18166_209_458_+ ACCAACAACUUAUUUUCACGGUCUAGAGAGAAACGAAUUAUUGUUUGCCCAGCC 60
SEQ16859_209_458_+ ACCAACAACUUAUUUUCACGGUCUAGAGAGAAACGAAUUAUUGUUUGCCCAGCC 60
.....10.....20.....30.....40.....50.....
|||||
)))))).....))))))..((((.....((((.....)))))).....
SEQ12068_211_460_+ UACUCAGCAACACACCAUCGCAACUCCGAGGCUUGGCAACGCGCGUAUAAAAUU 120
SEQ3931_209_458_+ UACUCAGCAACACACCAUCGCAACUCCGAGGCUUGGCAACGCGCGUAUAAAAUU 120
SEQ17216_217_465_+ UACUCAGCAACACACCAUCGCAACUCCGAGGCUUGGCAACGCGCGUAUAAAAUU 120
SEQ18166_209_458_+ UACUCAGCAACACACCAUCGCAACUCCGAGGCUUGGCAACGCGCGUAUAAAAUU 120
SEQ16859_209_458_+ UACUCAGCAACACACCAUCGCAACUCCGAGGCUUGGCAACGCGCGUAUAAAAUU 120
.....70.....80.....90.....100.....110.....
|||||
((((.....)))))).....((((.....)))))).....
SEQ12068_211_460_+ UUGUCUCUUUUUAAGCAGCACUUA-AAAUACC UUUAAUUUUUCUGAAUUAGAUACC 179
SEQ3931_209_458_+ UUGUCUCUUUUUAAGCAGCACUUG-AAAUCC UUUAAUUUUUCUGAAUUAGAUACC 179
SEQ17216_217_465_+ UUGUCUCUUUUUAAGCAGCACUUA-AAAUACC UUUAAUUUUUCUGAAUUAGAUACC 179
SEQ18166_209_458_+ UUGUCUCUUUUUAAGCAGCACUUCG-AAAUACC UUUAAUUUUUCUGAAUUAGAUACC 178
SEQ16859_209_458_+ UGUUCUCUUUUUAAGCAGCACUUUCAAAUCC UUUAAUUUUUCUGAAUUAGAUACC 179
.....130.....140.....150.....160.....170.....
|||||
.....((((.....)))))).....((((.....)))))).....
SEQ12068_211_460_+ AUAUAUA-AGAGAGACUUUGCUUUUAAA- AUAUCCAGASCAACCGCGCUACGAGAC 237
SEQ3931_209_458_+ AUAUAUA-AGAGAGAUUUUGCUUUUAAA- AUAUCCAGASCAACCGCGCUACGAGAC 237
SEQ17216_217_465_+ AUAUA-UA-ACGUAGACUUUGCUUUUAAA- AUAUCCAGASCAACCGCGCUACGAGAC 236
SEQ18166_209_458_+ AUAUAUAACCGAGAGACUUUGCUUUUAAA- AUAUCCAGASCAACCGCGCUACGAGAC 237
SEQ16859_209_458_+ AUAUAUAACUGAG--ACUUUGCUUUUCACAGAUACUUGACSCAACCGUGCUACGAGAU 237
.....190.....200.....210.....220.....230.....
|||||
)))))).....
SEQ12068_211_460_+ AAGGAGUUUCUC 250
SEQ3931_209_458_+ AAGGAGUUUCUC 250
SEQ17216_217_465_+ AAGGAGUUUCUC 249
SEQ18166_209_458_+ AAGGAGUUUCUC 250
SEQ16859_209_458_+ AAGGAGUUUCUC 250
```

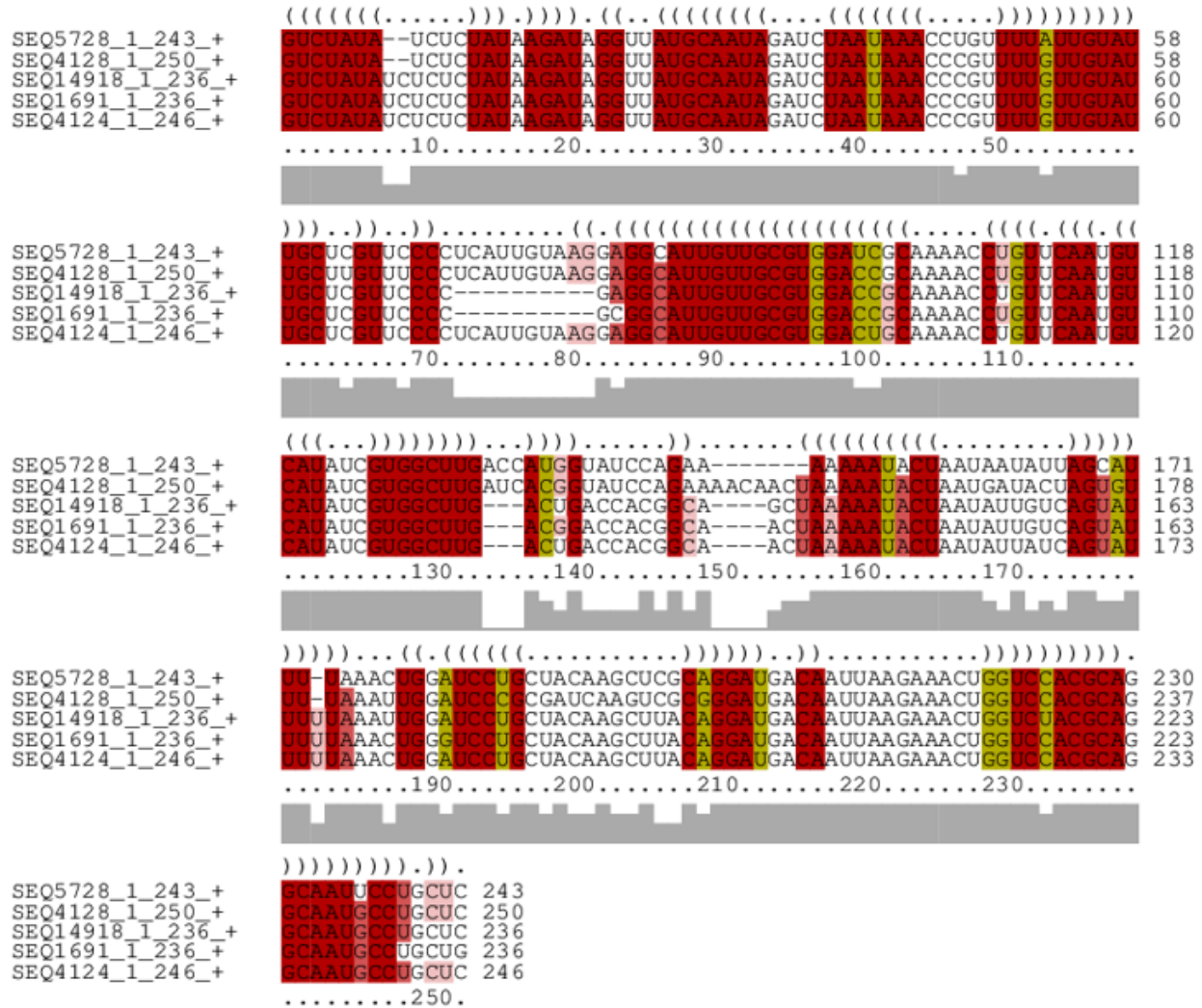
Cluster 12

```
.....((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((.....)))))))))))))))))))))))))))))))))))))))))))))))))..((
SEQ2544_78_204_+ AUGCGAU-----GCAAAAGAGC GACGAUACC AAUCGUCAAUCUUGCACC 45
SEQ10638_80_206_+ AUGCGAU-----GCAAAAGAGC GACGAUACC AAUCGUCAAUCUUGCACC 45
SEQ1148_56_182_+ AUGCGAU-----GCAAAAGAGC GACGAUACC AAUCGUCAAUCUUGCACC 45
SEQ2071_1_142_+ GCGCAACGGCCGGAUCCGAUCCGAAAAGAGC GACGAUACC AAUCGUCAAUCUUGCACC 60
SEQ923_1_142_+ GCGCGUCGACCGGACCUGAUCCGAAAAGAGC GACGAUACC AAUCGUCAAUCUUGCACC 60
.....10.....20.....30.....40.....50.....
.....((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((.....)))))))))))))))))))))))))))))))))))))))))))))))))..((
SEQ2544_78_204_+ UUUUAUGCUUAGGCAUAUUUAUUGCC CAUGUGUGCGCCAAUCGUU GAUUGGCGCGCUUUUA 105
SEQ10638_80_206_+ UUUUAUGCUUAGGCAUAUUUAUUGCC CAUGUGUGCGCCAAUCGUU GAUUGGCGCGCUUUUA 105
SEQ1148_56_182_+ UUUUAUGCUUAGGCAUAUUUAUUGCC CAUGUGUGCGCCAAUCGUU GAUUGGCGCGCUUUUA 105
SEQ2071_1_142_+ UUUUAUGCUUAGGCAUAUUUAUUGCC CAUGUGUGCGCCAAUCGUU GAUUGGCGCGCUUUUA 120
SEQ923_1_142_+ UUUUAUGCUUAGGCAUAUUUAUUGCC CAUGUGUGCGCCAAUCGUU GAUUGGCGCGCUUUUA 120
.....70.....80.....90.....100.....110.....
.....)))))))))..((
SEQ2544_78_204_+ UUUUUUUCG GAGUGC CUUUUGA 127
SEQ10638_80_206_+ UUUUUUUCG GAGUGC CUUUUGA 127
SEQ1148_56_182_+ UUUUUUUCG GAGUGC CUUUUGA 127
SEQ2071_1_142_+ UUUUUUGUCG GAGUGC CUUUUGA 142
SEQ923_1_142_+ UUUUUUGUCG GAGUGC CUUUUGA 142
.....130.....140
```

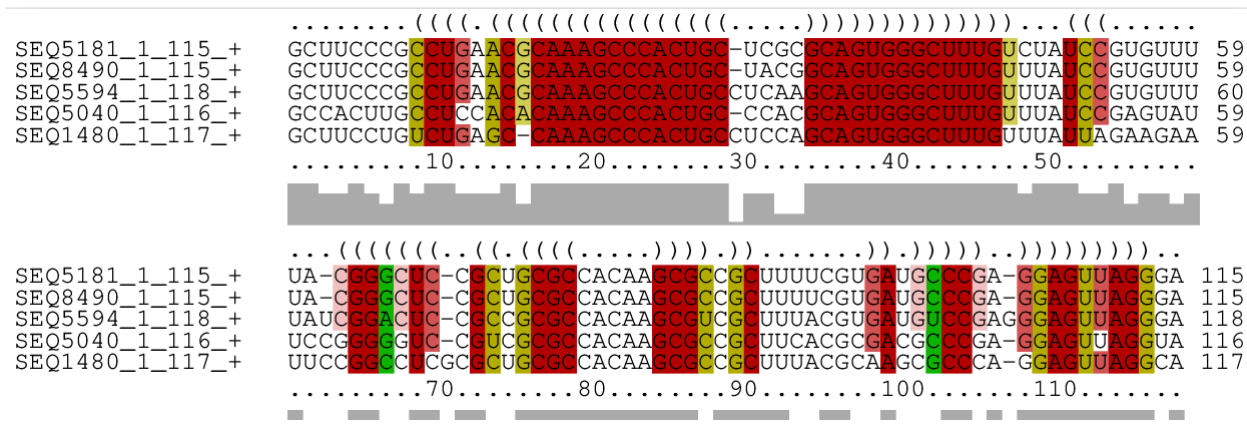
Cluster 27

```
.....((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((.....)))))))))))))))))))))))))))))))))))))))))))))))))..((
SEQ6294_1_183_+ GACGGAGUAC CGCCCGCGCACCCGCCCUGUCCGAGACC CGGACGGGGCGGUGCCUAUU 60
SEQ7589_2_184_+ AACGGAGUAC CGCCCGCGCACCCGCCCUGUCCGAGACC CGGACGGGGCGGUGCCUAUU 60
SEQ837_2_183_+ -ACGGAGUAC CGCCCGCGCACCCGCCCUGUCCGAGACC CGGACGGGGCGGUGCCUAUU 59
SEQ2084_2_183_+ AACGGAGUAC CGCCCGCGCACCCGCCCUGUCCGAAACC CGGACGGGGCGGUGCCUAUU 60
SEQ7619_1_182_+ GACGGAGUAC CGCCCGCGCACCCGCCCUGUCCGAAACC CGGACGGGGCGGUGCCUAUU 60
.....10.....20.....30.....40.....50.....
.....((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((.....)))))))))))))))))))))))))))))))))))))))))))))))))..((
SEQ6294_1_183_+ GUCGCGAUCUGCCCAUUGCGGGCGGGGGGGCC GAGACGUAACGGUGGUUUCG 120
SEQ7589_2_184_+ GUCGCGAUCGGCCCGUUGCUGACCAAAGGGCGGCCGAGACGACACCGUGCCUUUCG 120
SEQ837_2_183_+ GUCGCGAUCGGCCCGUUGCGGAGCCAAAGGGCGGUCCGAGACGACACCGUGCCUUUCG 119
SEQ2084_2_183_+ GCGCCGCGAUCUGCCCAUUGCGGGCGGAGCGGGGGGCCGAGACGACACCGUGGUUUCG 119
SEQ7619_1_182_+ AUCGCGAUCUGCCCAUUGCGGGCGGAGCGGGGGGCCGAGACGACACCGUGGUUUCG 119
.....70.....80.....90.....100.....110.....
.....((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((.....)))))))))))))))))))))))))))))))))))))))))))))))))..((
SEQ6294_1_183_+ CGGCGCCGAA GCGGUGAUC GGGCGACGGGAUUGGACAAGCGGCAAAGGGAGCUUCCCG 180
SEQ7589_2_184_+ CGGCGCCGAA GCGGUGAUC GGGCGACGGGACGCGACCAGCGGCAAAGGGAGCUUCCCG 180
SEQ837_2_183_+ CGGCGCCGAA GCGGUGAUC CGGGCGGGGAUUGGACAAGCGGCAAAGGGAGCUUCCCG 179
SEQ2084_2_183_+ CGGCGCCGAG GCGGUGAUC GGGCGACGGGAUUGGACCAGCGGCAAAGGGAGCUUCCCG 179
SEQ7619_1_182_+ CGGCGCCGCG GCGGUGAUC GGGCGACGGGAUUGGACCAGCGGCAAAGGGAGCUUCCCG 179
.....130.....140.....150.....160.....170.....
...
SEQ6294_1_183_+ CAC 183
SEQ7589_2_184_+ CAC 183
SEQ837_2_183_+ CAC 182
SEQ2084_2_183_+ CAC 182
SEQ7619_1_182_+ CAC 182
...
```

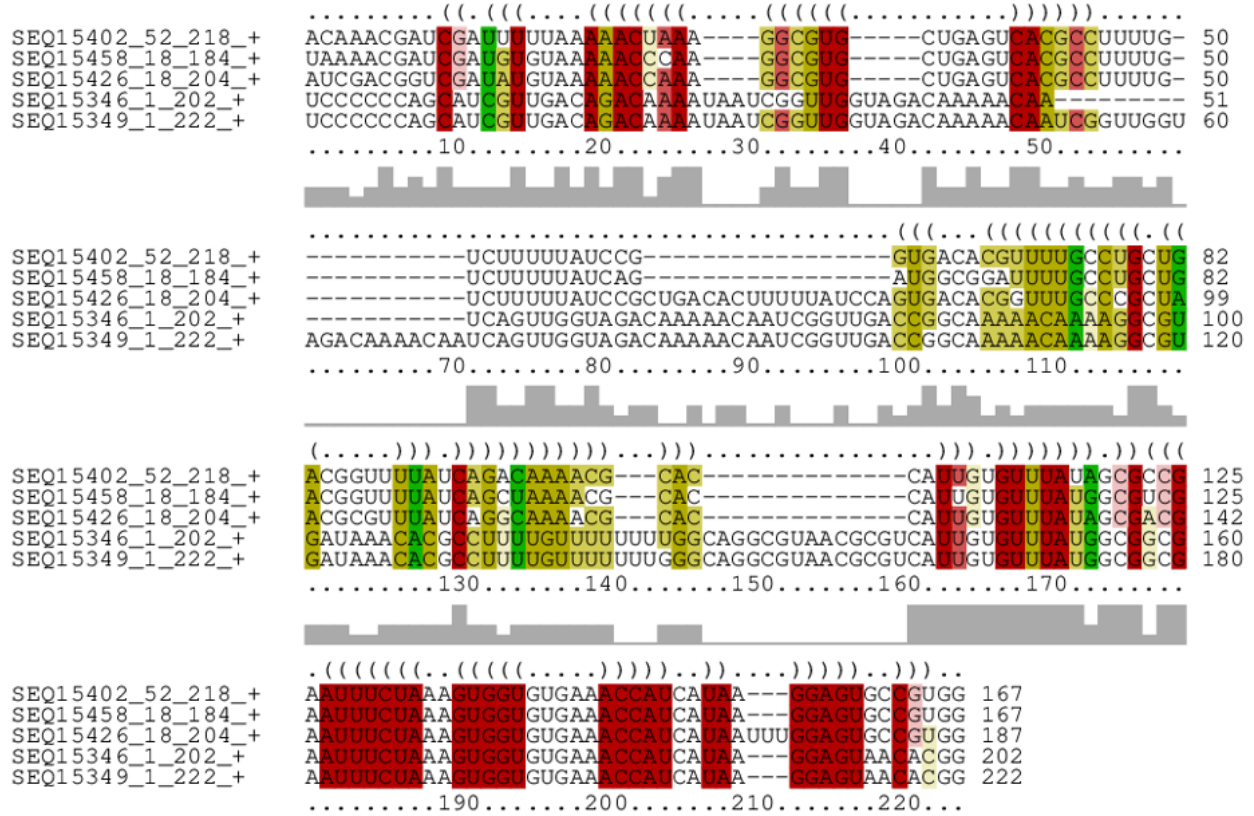

Cluster 64



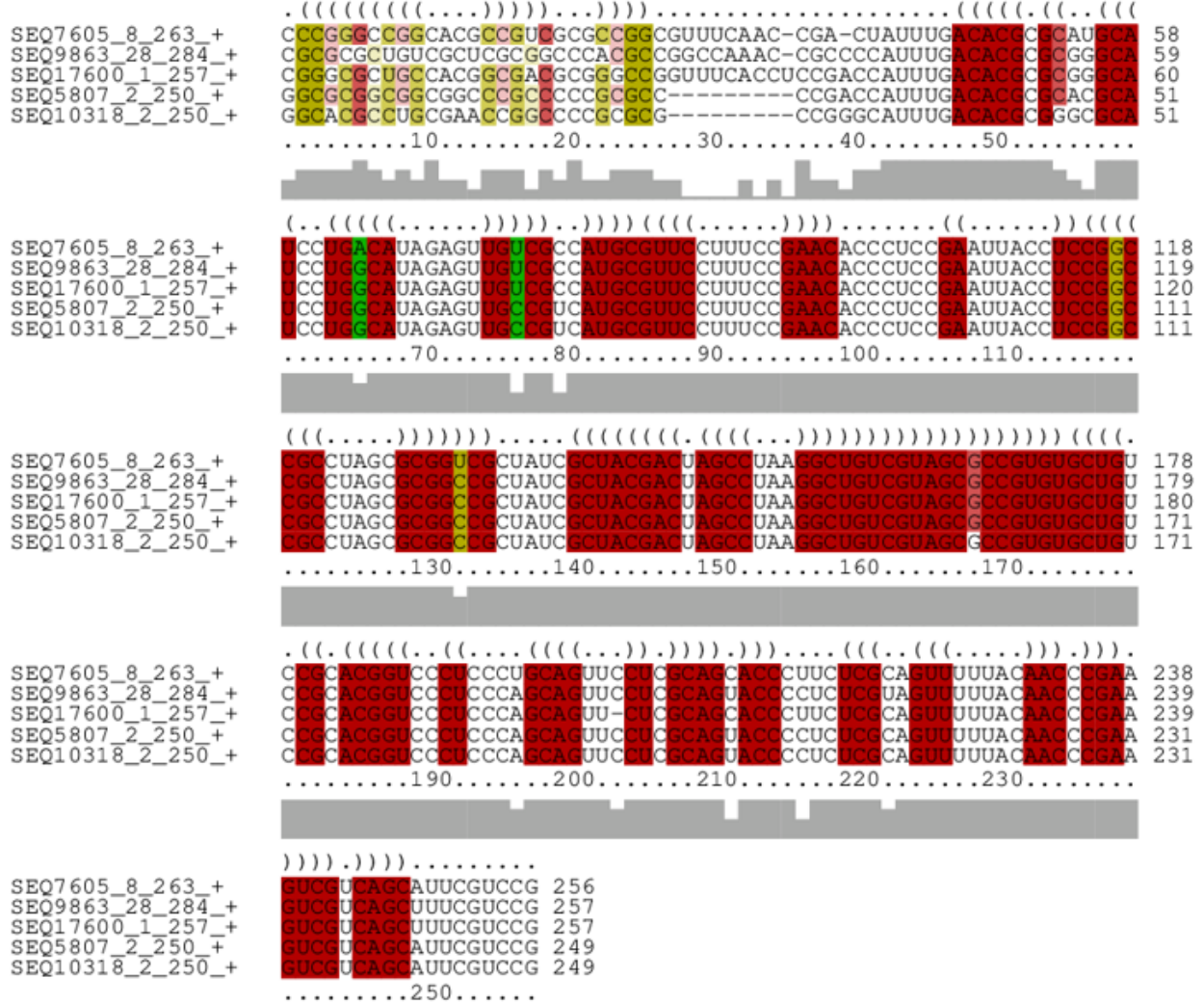
Cluster 66



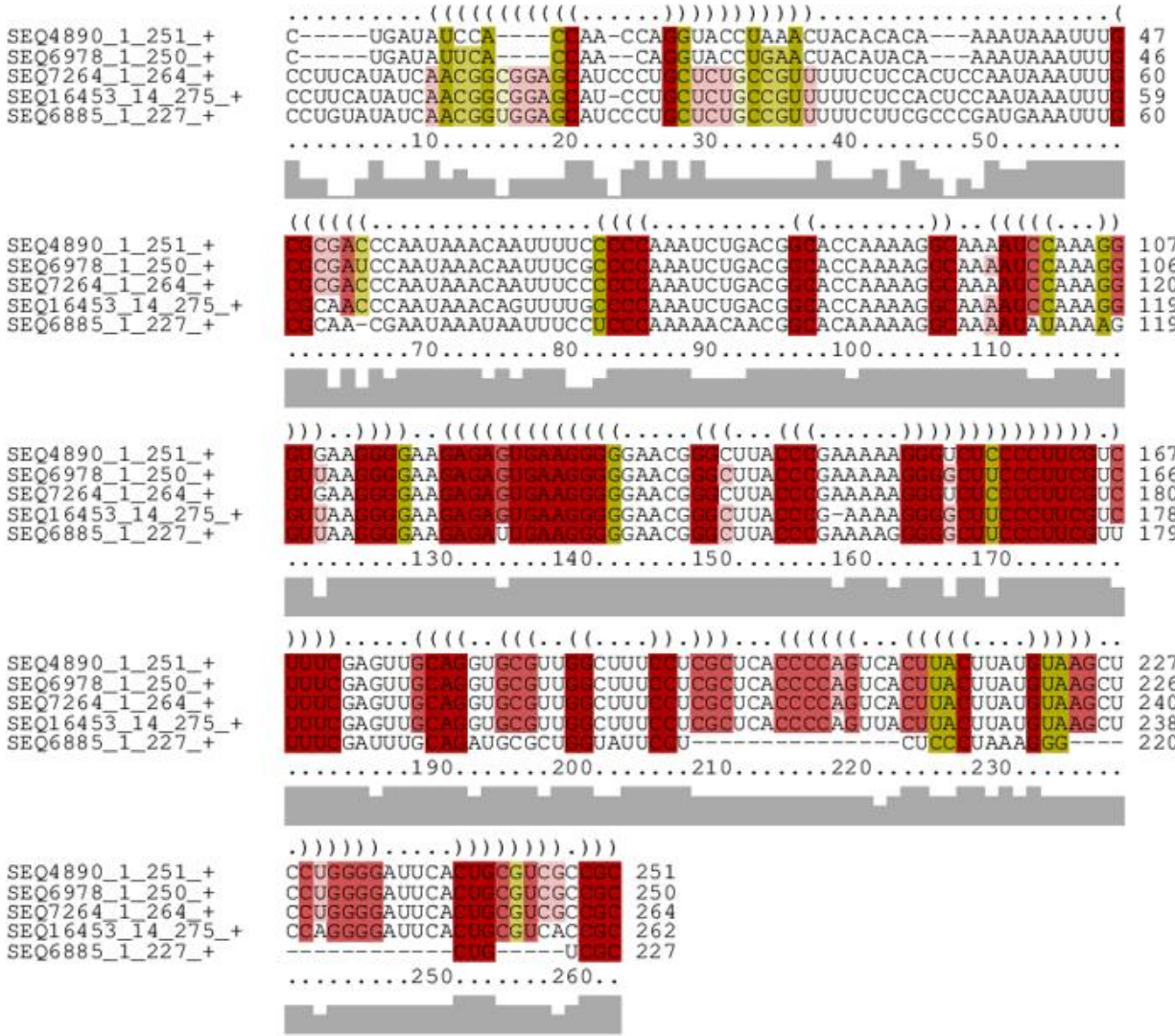
Cluster 70



Cluster 71



Cluster 73



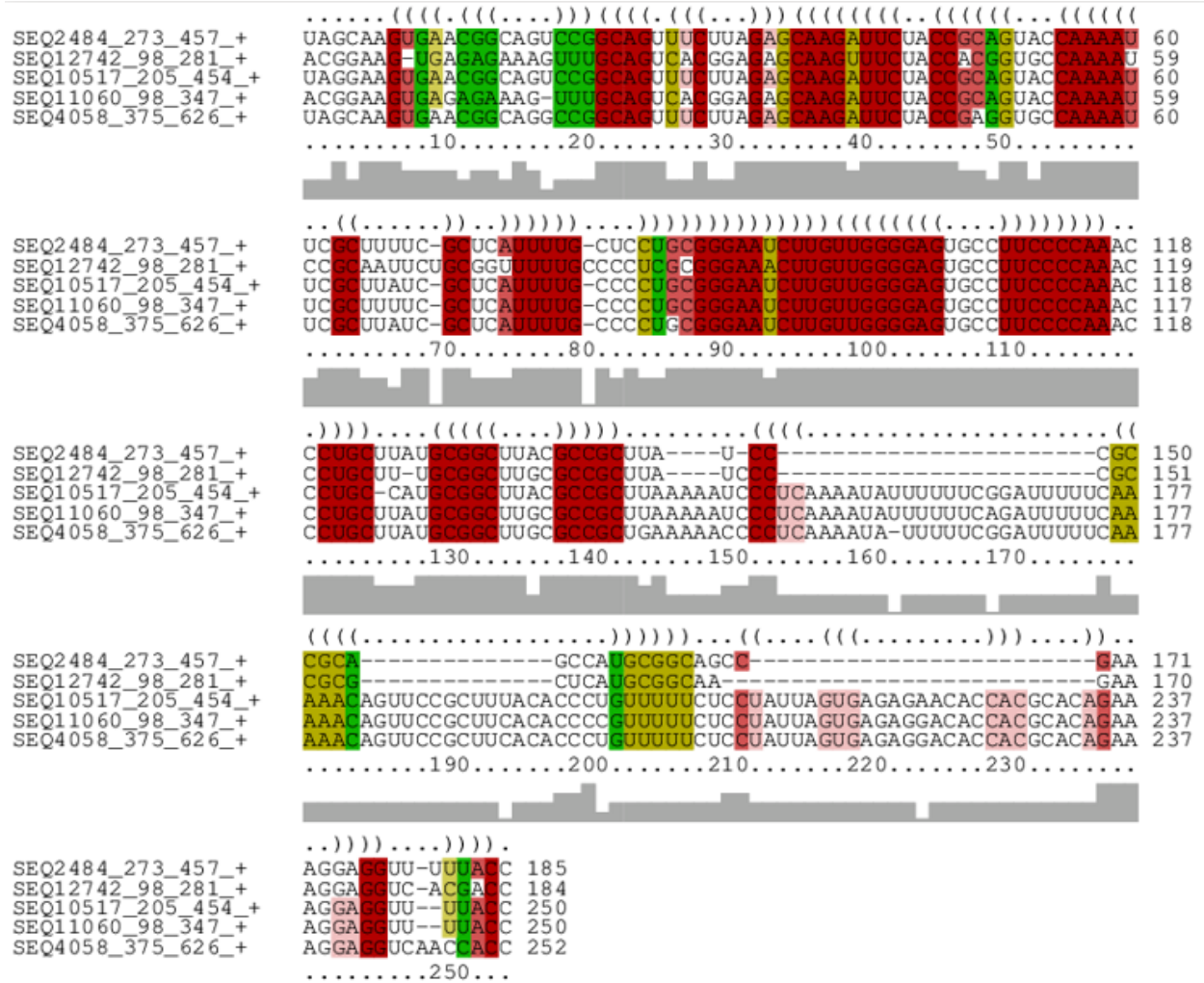
Cluster 74

```
.....((((((((.....)))))).....((((((((.....)))))).....
SEQ16552_1_114_+ UGCUAACGGCGCGGAAUCSUCGCGCCAUAUUCCUCCUUUCCCGCCAUUCUUCUUUCCA 60
SEQ13049_1_114_+ UGCUAACGGCGCGGAAUCSUCGCGCCAUAUUCCUCUUUCCCGUCAUUCUUCUUUCCA 60
SEQ13665_1_114_+ UGCUAACGGCGCGGAAUCSUCGCGCCAUAUUCCUCCUUUCCCGUCAUUCUUCUUUCCA 60
SEQ14725_1_114_+ UGCUAACGGCGCGGAAUCSUCGCGCCAUAUUCCUCCUUUCCCGUCAUUCUUCUUUCCA 60
SEQ17256_1_114_+ UGCUAACGGCGCGGAAUCSUCGCGCCAUAUUCCUCCUUUCCCAUCAUUCUUCUUUCUA 60
.....10.....20.....30.....40.....50.....
).....((((((((.....)))))).....
SEQ16552_1_114_+ CGUCCUAUUCGUCUUJGGJUAUAGUGUUUUCAUCAUAAAGCAGGAGAACACA 114
SEQ13049_1_114_+ CGUCCUAUUCUUCUUJGGJUAUAGUGUUUUCAUCAUAAAGCAGGAGAACACA 114
SEQ13665_1_114_+ CGUCAUAUUCGUCUUJGGJUAUAGUGUUUUCAUCAUAAAGCAGGAGAACACA 114
SEQ14725_1_114_+ CGUCAUAUCCGGCUUJGGJUAUAGUGUUUUCAUCAUAAAGCAGGAGAACACG 114
SEQ17256_1_114_+ UAUCCUGUCCGGCUUJGGJUAUAGUGUUUUCAUCAUAAAGCAGGAGAACACG 114
.....70.....80.....90.....100.....110..
```

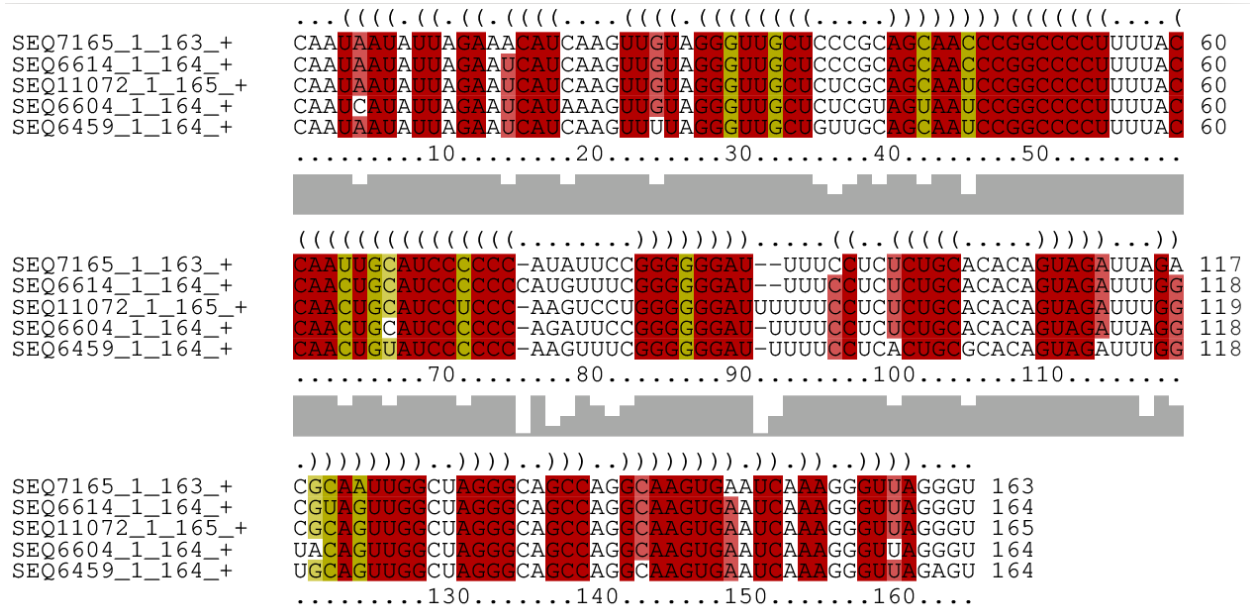
Cluster 75

```
.....((((((((.....)))))).....((((((((.....)))))).....
SEQ4473_251_500_+ UUGUGUUAGUUGUGAUUUUUAUGUUGUUUUUAUUAAAUCUGUUUUAAAUUUAGAGUU 60
SEQ4511_251_500_+ UUGUGUUAGUUGUGAUUUUUAUGUUGUUUUUAUUAAAUCUGUUUUAAAUUUAGAGUU 60
SEQ14963_247_501_+ UUGUGUUAGUUGUGAUUUUUAUGUUGUUUUUAUUAAAUCUGUUUUAAAUUUAGAGUU 60
SEQ8544_247_501_+ UUGUGUUAGUUGUGAUUUUUAUGUUGUUUUUAUUAAAUCUGUUUUAAAUUUAGAGUU 60
SEQ4442_247_501_+ UUGUGUUAGUUGUGAUUUUUAUGUUGUUUUUAUUAAAUCUGUUUUAAAUUUAGAGUU 60
.....10.....20.....30.....40.....50.....
((...((((((((.....)))))).....((((((((.....)))))).....
SEQ4473_251_500_+ UUGAGUUUAGAGGUUGUUUUAAAGCUUUUAGAGGUUAGUUUUUAAGAUAAAAGAUUA 113
SEQ4511_251_500_+ UUGAGUUUAGAGGUUGUUUUAAAGCUUUUAGAGGUUAGUUUUUAAGAUAAAAGAUUA 120
SEQ14963_247_501_+ UUGAGUUUAGAGGUUGUUUUAAAGCUUUUAGAGGUUAGUUUUUAAGAUAAAAGAUUA 120
SEQ8544_247_501_+ UUGAGUUUAGAGGUUGUUUUAAAGCUUUUAGAGGUUAGUUUUUAAGAUAAAAGAUUA 120
SEQ4442_247_501_+ UUGAGUUUAGAGGUUGUUUUAAAGCUUUUAGAGGUUAGUUUUUAAGAUAAAAGAUUA 120
.....70.....80.....90.....100.....110.....
))))))))).....((((((((.....)))))).....
SEQ4473_251_500_+ AAGAUAAAAGAUUUUAGAGUUUGUAGUUUAGUUUUAAAAGAUUAAGAUAAAAGAUUUU 173
SEQ4511_251_500_+ AAGAUAAAAGAUUUUAGAGUUUGUAGUUUAGUUUUAAAAGAUUAAGAUAAAAGAUUUU 180
SEQ14963_247_501_+ AAGAUAAAAGAUUUUAGAGUUU---GAGUUUGUAGUU---AGAGGUUAGGGUUAGAGGUU 176
SEQ8544_247_501_+ AAGAUAAAAGAUUUUAGAGUUU---GAGUUUGUAGUU---AGAGGUUAGGGUUAGAGGUU 176
SEQ4442_247_501_+ AAGAUAAAAGAUUUUAGAGUUU---GAGUUUGUAGUU---AGAGGUUAGGGUUAGAGGUU 176
.....130.....140.....150.....160.....170.....
AGAGUUUGUAGUUU---GUAGUUAGAGGUUAGAGGUUGUAGGUUGUAGGUUGUAGGUUUGAG 232
SEQ4511_251_500_+ AGAGUUUGUAGUUU---GUAGUUUGUAGUUAG-----AGGUUAGAGGUUAGAGGUUAGAG 232
SEQ14963_247_501_+ AGAGUUGUAGGUUAGUGGUUCUUGGUUCUUGGUUAGUGUUUAGUGUUUAGUGGUUAGUG 236
SEQ8544_247_501_+ AGAGUUGUAGGUUAGUGGUUCUUGGUUCUUGGUUAGUGUUUAGUGUUUAGUGGUUAGUG 236
SEQ4442_247_501_+ AGAGUUGUAGGUUAGUGGUUCUUGGUUCUUGGUUAGUGUUUAGUGUUUAGUGGUUAGUG 236
.....190.....200.....210.....220.....230.....
GUUGUAGGU---UGUAGGUUG 250
SEQ4511_251_500_+ GUUGUAGGU---UGUAGGUUG 250
SEQ14963_247_501_+ GUUGUGGUUGUGGUUGUGG 255
SEQ8544_247_501_+ GUUGUGGUUGUGGUUGUGG 255
SEQ4442_247_501_+ GUUGUGGUUGUGGUUGUGG 255
```

Cluster 78



Cluster 86



La figure suivante montre quelques structures secondaires des *clusters* mentionnés.

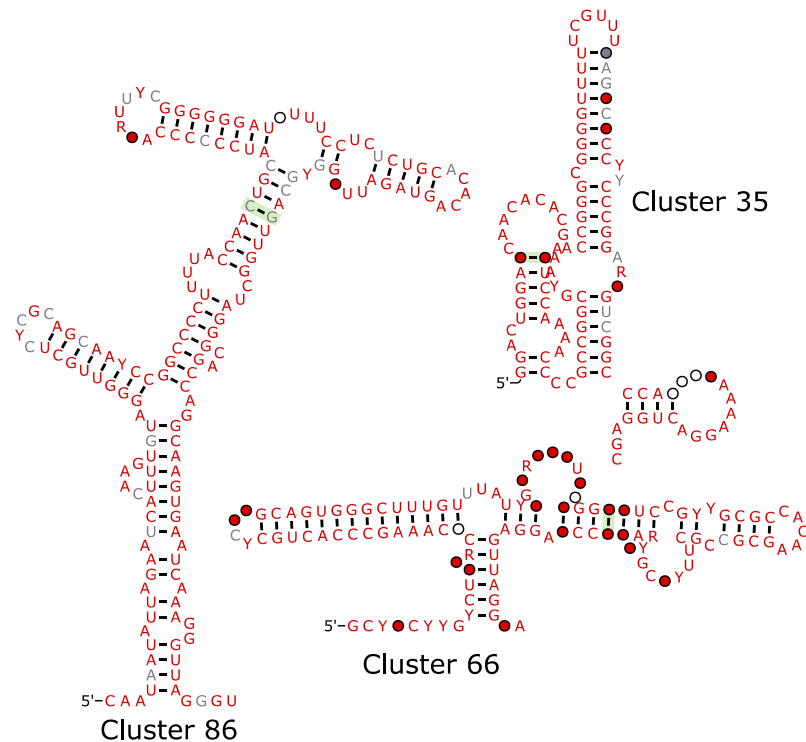


Figure 30 - Structures secondaires obtenues par GraphClust

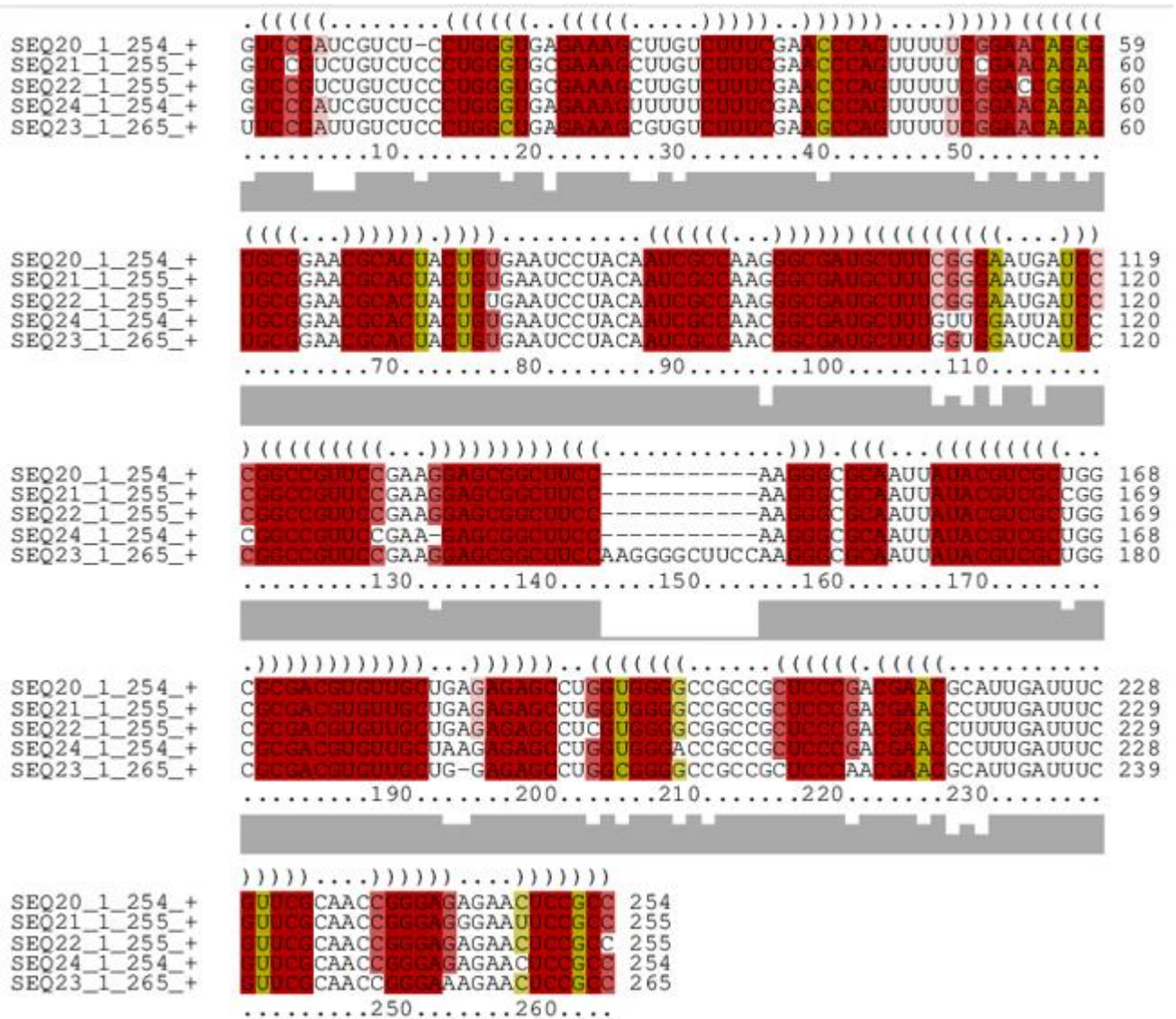
11.2.2 Recherche d'homologie avec Infernal

Nous avons exécuté GraphClust sur les séquences homologues trouvées à partir des *clusters* sélectionnés précédemment.

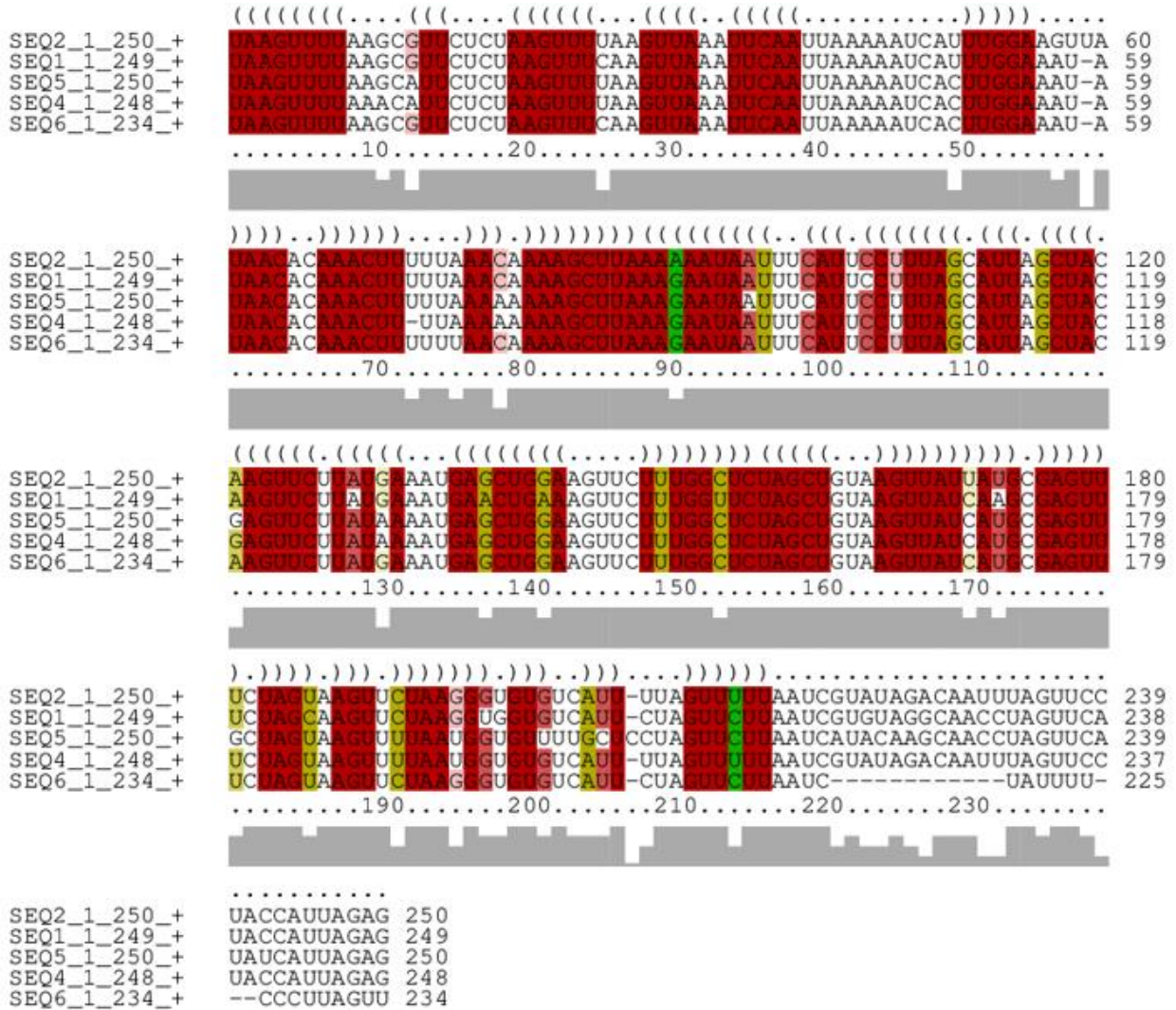
11.2.2.1 Requête simple

La liste qui va suivre concerne les *clusters* sélectionnés à partir des résultats de GraphClust exécutés sur les homologues des *clusters* obtenus de GraphClust.

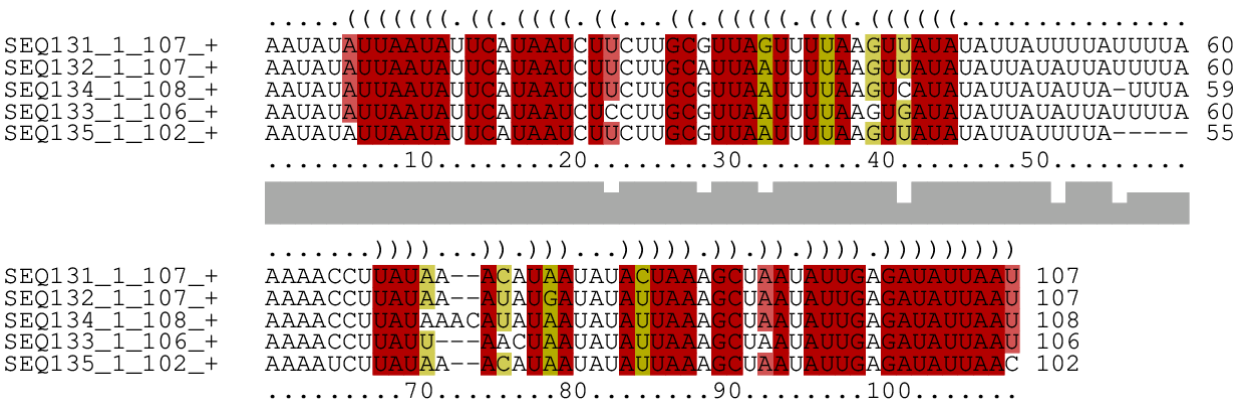
Cluster 2



Cluster 3



Cluster 4



Cluster 7

```

.....(((((((.....(((((((.....(((((((.....))))))))))
SEQ34_1_250_+ GUCUAUA--UCUCUAUAAGAUA GUUAU SCAAU GGUC JAU AAA CCCGU UUGUUGUA 58
SEQ35_1_250_+ GUCUAUA--UCUCUAUAAGAUA GUUAU SCAAU GAUC JAU AAA CCCGU UUGUUGUA 58
SEQ36_1_243_+ GUCUAUA--UCUCUAUAAGAUA GUUAU SCAAU GAUC JAU AAA CCUGU UUGUUGUA 58
SEQ33_1_246_+ GUCUAUAUCUCUCUAUAAGAUA GUUAU SCAAU GAUC JAU AAA CCCGU UUGUUGUA 60
.....10.....20.....30.....40.....50.....
[Secondary structure bar for positions 10-50]

))).....))..))..)))))((.....(((((((.....(((((((.....(((((((.....(((
SEQ34_1_250_+ UGCUUGUUC CCUAUUGCCAGG AGGCAUUGUUGCGUGGACC AAAAAUCUGU UCAAUG 118
SEQ35_1_250_+ UGCUUGUUC CCUAUUGUAAGG AGGCAUUGUUGCGUGGACC CAAAAUCUGU UCAAUG 118
SEQ36_1_243_+ UGCUUGUUC CCUAUUGUAAGG AGGCAUUGUUGCGUGGACC CAAAAUCUGU UCAAUG 118
SEQ33_1_246_+ UGCUUGUUC CCUAUUGUAAGG AGGCAUUGUUGCGUGGACC CAAAAUCUGU UCAAUG 120
.....70.....80.....90.....100.....110.....
[Secondary structure bar for positions 70-110]

((((.....))))))..))..))..((((.....))))))..))..))..((((.....))))))
SEQ34_1_250_+ CAUAUCGUGGCCUUGACCAGGUAUCCAGAAAACAACUAAAUAUAUAUAAAUUUAGUA 178
SEQ35_1_250_+ CAUAUCGUGGCCUUGACCAGGUAUCCAGAAAACAACUAAAUAUAUAUAAAUUUAGUA 178
SEQ36_1_243_+ CAUAUCGUGGCCUUGACCAGGUAUCCAGAA-----AAAAUAUAUAUAAAUUUAGCA 171
SEQ33_1_246_+ CAUAUCGUGGCCUUGACU---ACCAAGG---CAACUAAAUAUAUAUAAAUUUCAGUA 173
.....130.....140.....150.....160.....170.....
[Secondary structure bar for positions 130-170]

)).....))..((((.....))))))..))..))..((((.....))))))..))..))..))
SEQ34_1_250_+ UU-UAAACUGAUCUCCUACAAGCUCGAGGAGACAAUUAAGAAACGGUCCACGCAG 237
SEQ35_1_250_+ UU-UAAAUUGAUCUCCGGAUCAAGUCGAGGAGACAAUUAAGAAACGGUCCACGCAG 237
SEQ36_1_243_+ UU-UAAACUGAUCUCCUACAAGCUCGAGGAGACAAUUAAGAAACGGUCCACGCAG 230
SEQ33_1_246_+ UUUAACUGAUCUCCUACAAGCUUAGGAGACAAUUAAGAAACGGUCCACGCAG 233
.....190.....200.....210.....220.....230.....
[Secondary structure bar for positions 190-230]

))))))..))..
SEQ34_1_250_+ SCAAUGCCGUCU 250
SEQ35_1_250_+ SCAAUGCCGUC 250
SEQ36_1_243_+ SCAAUCCGUC 243
SEQ33_1_246_+ SCAAUGCCGUC 246
.....250.

```

Quelques structures secondaires sont présentées dans la figure suivante à titre d'exemple.

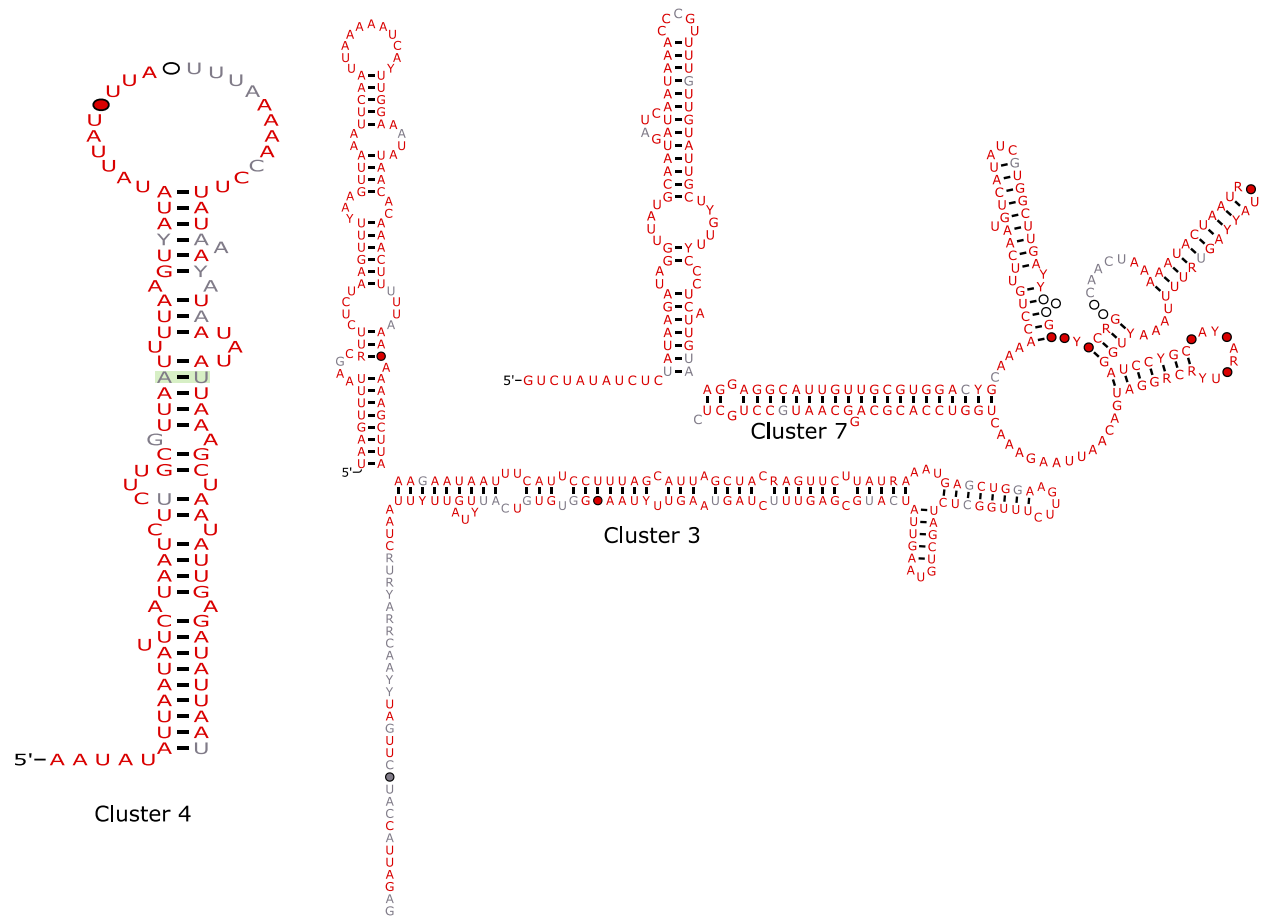
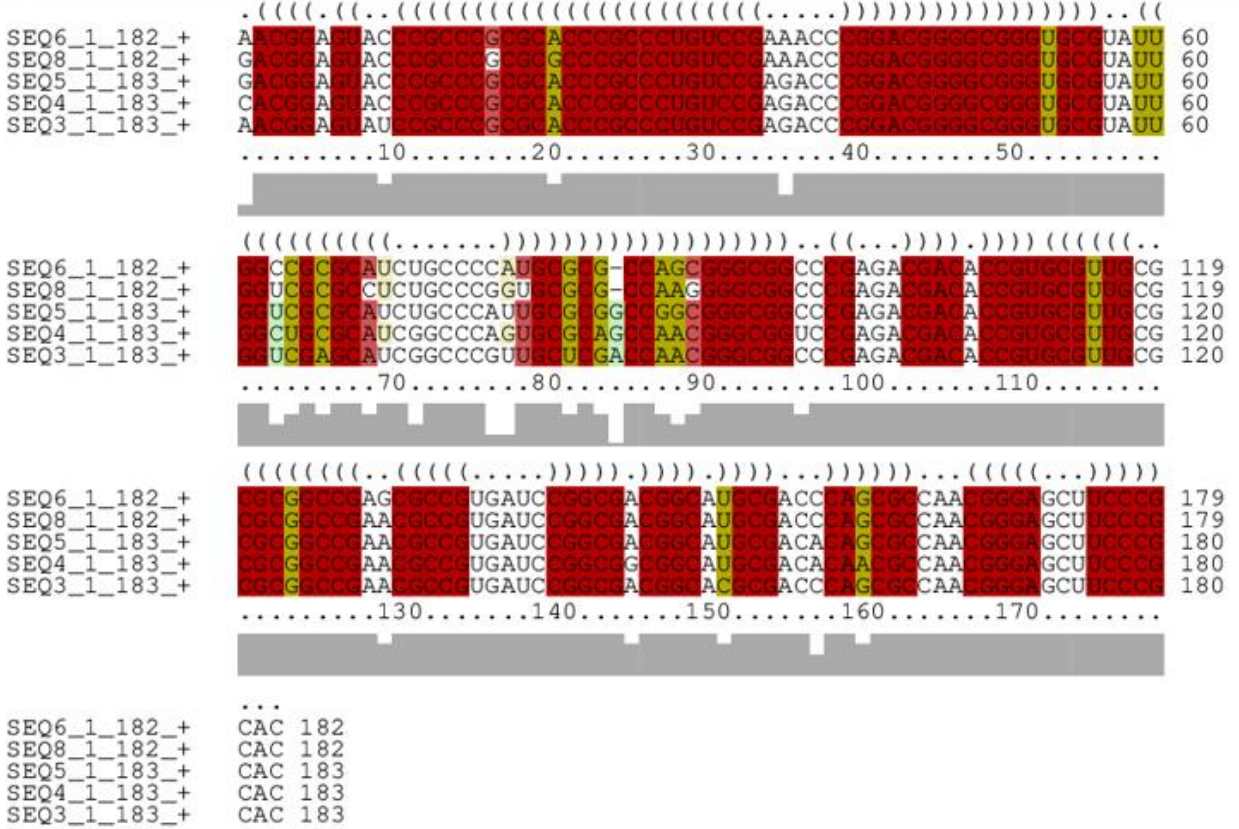


Figure 31 - Structures secondaires obtenues par GraphClust

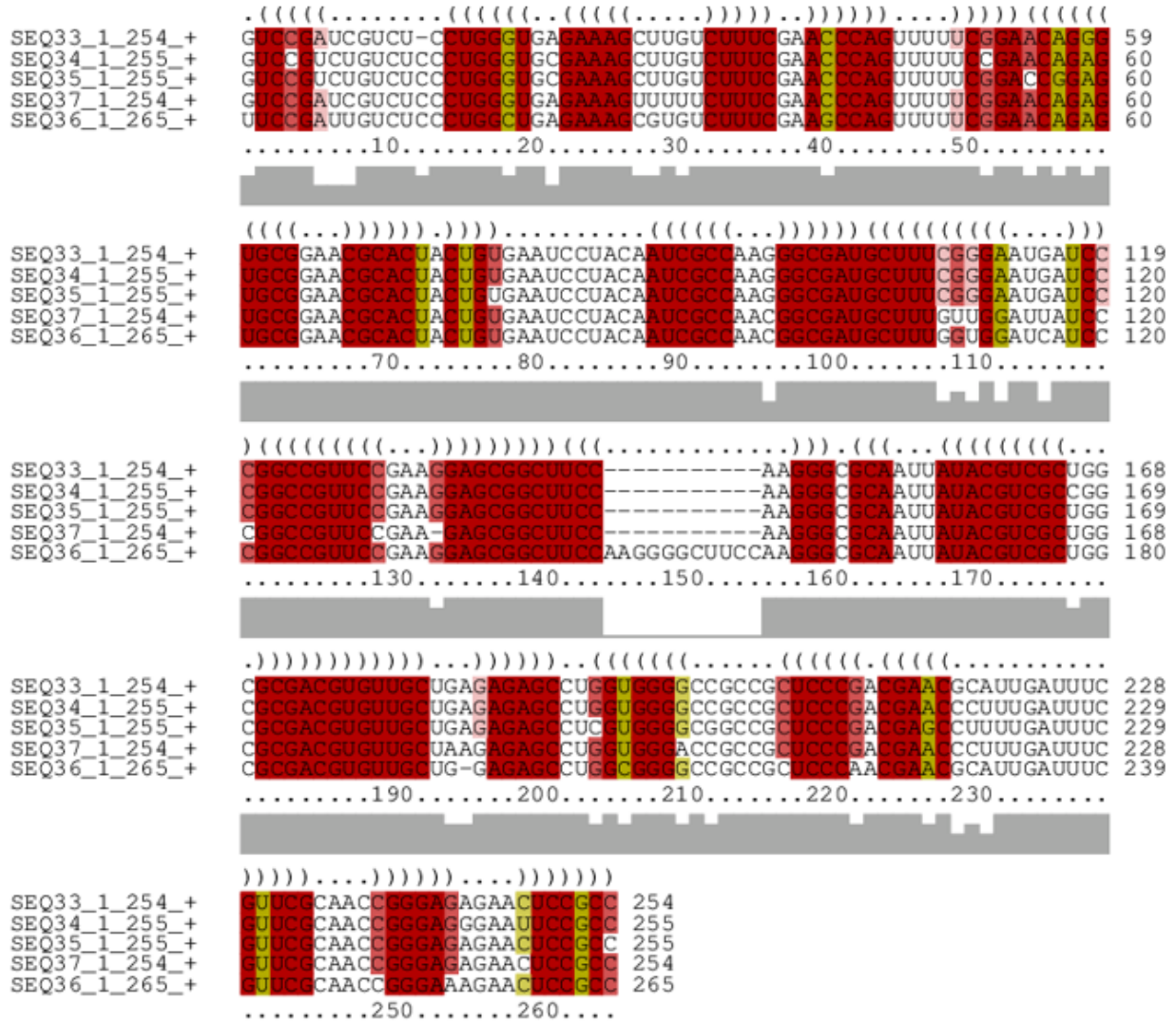
11.2.2.2 Requête complexe

La liste correspond aux *clusters* intéressants obtenus à partir des séquences de la deuxième requête. De même que les cas précédents, pour que ça soit des nouveaux ARNnc, il faut regarder la conservation et la co-variation existant dans les alignements structuraux de chaque cluster.

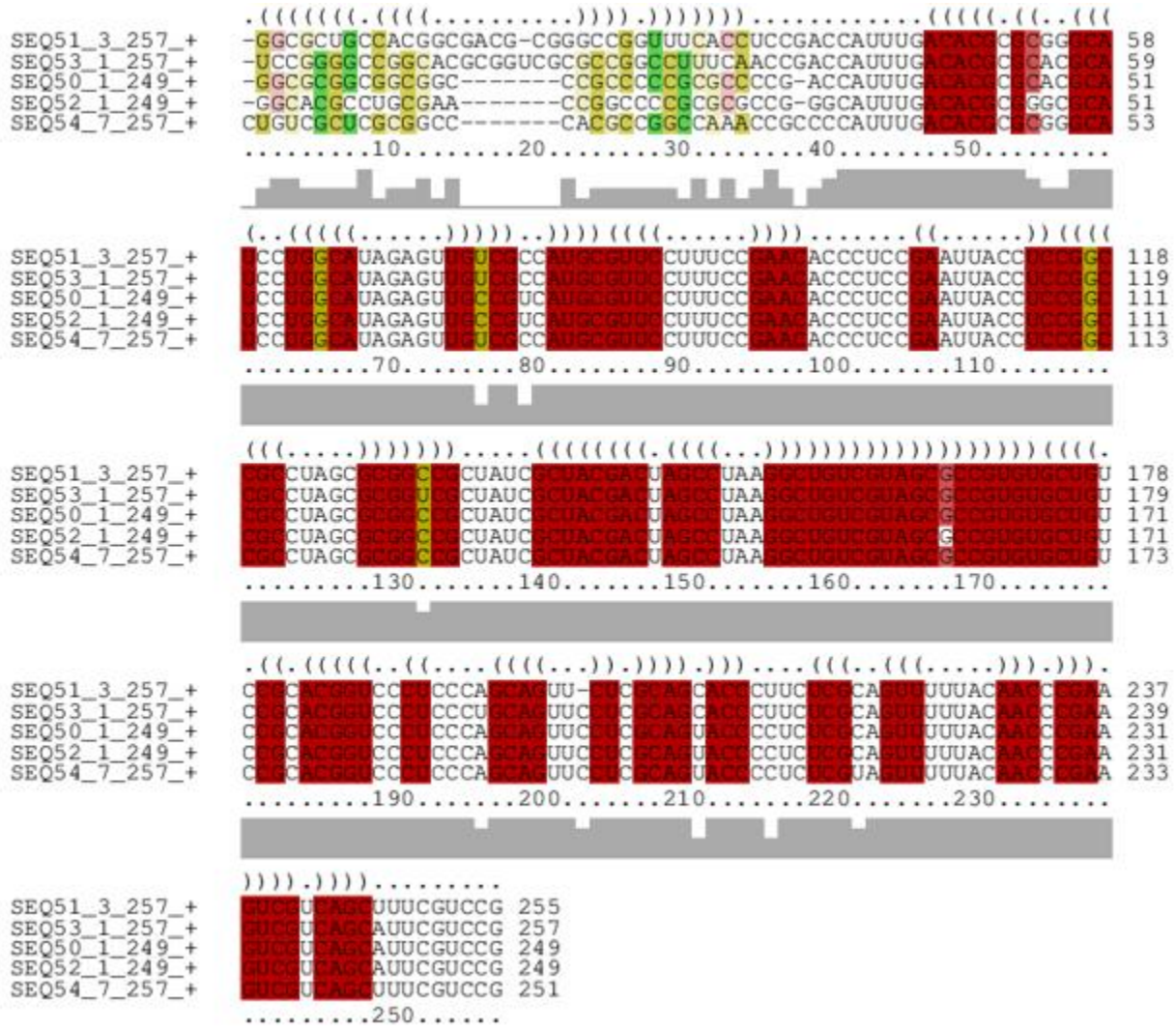
Cluster 1



Cluster 2



Cluster 4



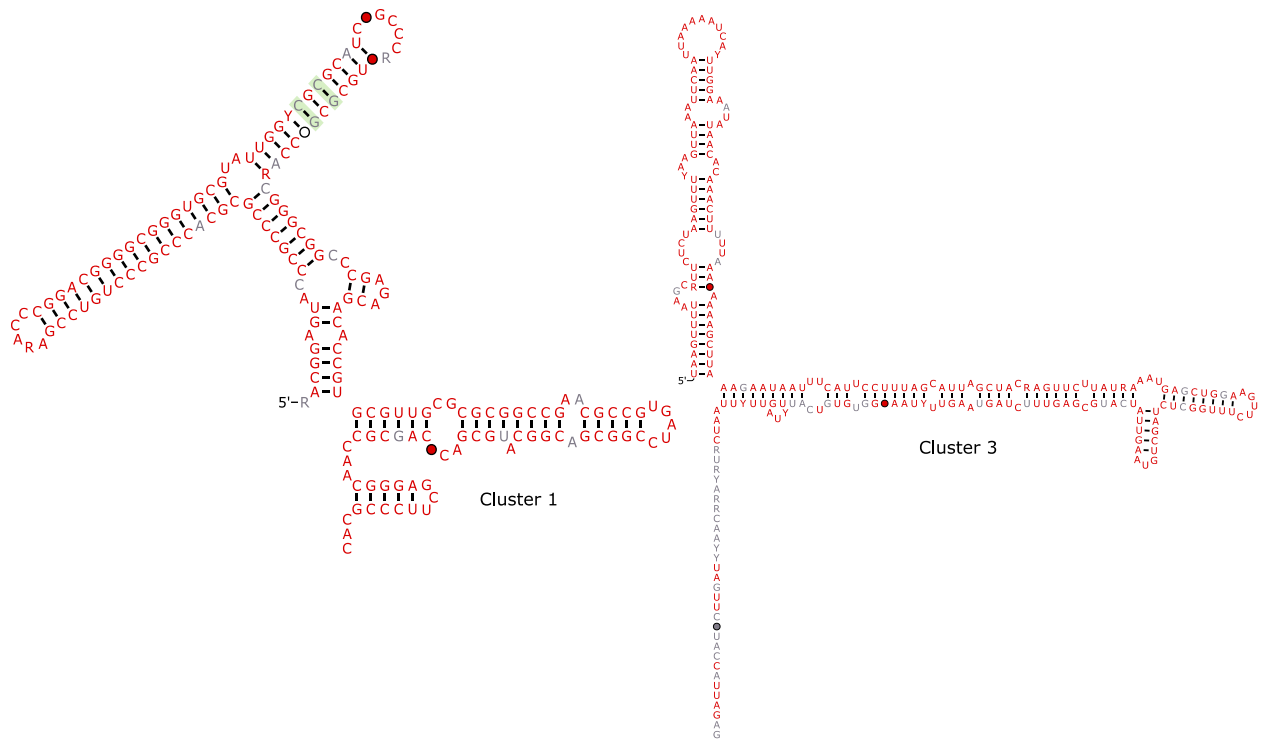


Figure 32 - Exemples de structures secondaires des séquences de la requête 2

11.3 RiboGap v2.1

Dans ce cas, nous exécutons les deux types de requêtes dans la base de données RiboGap v2.1 qui correspond aux génomes incomplets. Ceci nous permet d'avoir plus de candidats pour notre recherche de nouveaux ARNnc.

11.3.1 GraphClust



Une fois que nous avons nos deux ensembles de données qui correspondent respectivement aux séquences obtenues à partir de la requêtes 1 et 2, on exécute GraphClust.

11.3.1.1 Requête simple

Plusieurs candidats ont été retenus suite à notre sélection des résultats de GraphClust en se reposant sur les critères de co-variation et de conservations. L'alignement structural des candidats est décrit dans ce qui suit :

Cluster 11

```

.....(((((((.....(((((((.....((((.....)))))))))))))..))....))
SEQ129_1_132_+ UCUGACU UUAUGAC C CUUUGCAGG ---CGGUGGUUACCGUCUGUCAAUAGUUGC GUC 57
SEQ1514_1_132_+ UCCGGUUUUAUGAC C CUUUGACAGG ---CGGUGGUUACCGUCUGUCAAUAGUUGC GUC 57
SEQ11668_1_132_+ UCUGACU UUAUGAUG C CUUUGCAGG ---CGGCGGUUACCGUCUGUCAAUAGUUGC GUC 57
SEQ11099_1_132_+ UCCGGUUUUAUGAUG C CUUUGCAGG ---CGACGGUUACUGUCUGUCAAUAGUUGC GUC 57
SEQ7162_1_138_+ UUUUAUCGACGC GCGGCCGGGCAGGGUACUACCGGUUACGCUCUGUAAAGCUUUUAG GUC 60
.....10.....20.....30.....40.....50.....

))..(((.....((((.....((((.....))))))))))....))..
SEQ129_1_132_+ AUAUUCUGAUC CUAUUGAUGCAGCAUU GUGCGCCGAC CCAAGGUUGGCGCGUUU CUU 117
SEQ1514_1_132_+ AUAUUCUGAUC CUAUUGAUGCAGCAUU GUGCGCCGAC CCAAGGUUGGCGCGUUU CUU 117
SEQ11668_1_132_+ AUAUUCUGAUC CUAUUGAUGCAGCAUU GUGCGCCGAC CUCAGGUUGGCGCGUUU CUU 117
SEQ11099_1_132_+ AUAUUCUGAUC CUAUUGAUGCAGCAUU GUGCGCCGAC CCAAGGUUGGCGCGUUU CUU 117
SEQ7162_1_138_+ AUAUUCUGAUC CUAUUGAUGCAGCAUU GUGCGCCGAC CCAAGGUUGGCGCGUUU CUU 120
.....70.....80.....90.....100.....110.....

))..))....))....
SEQ129_1_132_+ UUUGUGAAUAUCG---C 132
SEQ1514_1_132_+ UUUGUGAAUAUCG---C 132
SEQ11668_1_132_+ UUUGUGAAUAUCG---C 132
SEQ11099_1_132_+ UUUGUGAAUAUCG---C 132
SEQ7162_1_138_+ UUUCGGCAUCCAUAC 138
.....130.....

```

Cluster 17

```
.....(((.....((((((((((((.....((((((((.....(((.....(((.....(((
SEQ219_1_255_+  AAGAAUC CUC CAUUUUUUUUUGUAUAUAUC UUCUACAAUC AUUAACA SCUAAG AAAUC 60
SEQ8477_1_255_+ AAGAAUC CUC CAUUUUUUUUUGUAUAUAUC UUCUACAAUC AUUAACA SCUAAG AAAUC 60
SEQ5136_1_255_+ AAGAAUC CUC CAUUUUUUUUUGUAUAUAUC UUCUACAAUC AUUAACA SCUAAG AAAUC 60
SEQ5925_1_255_+ AAGAAUC CUC CAUUUUUUUUUGUAUAUAUC UUCUACAAUC AUUAACA SCUAAG AAAUC 60
SEQ5270_1_264_+ AAGAAUC CUC CAUUUUUUUUUGUAUAUAUC UUCUACAAUC AUUAACA SCUAAG AAAUC 60
.....10.....20.....30.....40.....50.....

((((((((.....
SEQ219_1_255_+  CAAAAUUC AAGACUAAUUCUGAACGUUCGUUU GAAAAACGAGUGAAAAUA-----A 111
SEQ8477_1_255_+  CAAAAUUC AAGACUAAUUCUGAACGUUCGUUU GAAAAACGAGUGAAAAUA-----A 111
SEQ5136_1_255_+  CAAAAUUC AAGACUAAUUCUGAACGUUCGUUU GAAAAACGAGUGAAAAUA-----G 111
SEQ5925_1_255_+  CAAAAUUC AAGACUAAUUCUGAACGUUCGUUU GAAAAACGAGUGAAAAUA-----G 111
SEQ5270_1_264_+  CAAAAUUC AAGACUAAUUCUGAACGUUCGUUU GAAAAACGAGUGAAAAUA-----G 111
.....70.....80.....90.....100.....110.....

.....)))).))))).....))))).....))))).....))))).....)))))
SEQ219_1_255_+  CAAGAAAA CGAAGAAAACAAGAACGUUCACCSAAU CUGAACUUGGUUUGGAUU GUUUU 171
SEQ8477_1_255_+  CAAGAAAA CGAAGAAAACAAGAACGUUCACCSAAU CUGAACUUGGUUUGGAUU GUUUU 171
SEQ5136_1_255_+  CAAGAAAA CGAAGAAAACAAGAACGUUCACCSAAU CUGAACUUGGUUUGGAUU GUUUU 171
SEQ5925_1_255_+  UAAGAAAA UAAACAAAACAAGAACGUUCGCCSAAU CUGAACUUGGUUUGGAUU GUUUU 171
SEQ5270_1_264_+  AAAGAAAA CAAAAGAAAACAAGAACGUUCGCCSAAU CUGAACUUGGUUUGGAUU GUUUU 180
.....130.....140.....150.....160.....170.....

)))).))))).....))))).....))))).....))))).....)))))
SEQ219_1_255_+  AAUUAGU AGUCUUAUGUUAGAUAUCCUUUUU UUUUUUCGU GAAAA UGGUAGAAUA GUAGG 231
SEQ8477_1_255_+  AAUUAGU AGUCUUAUGUUAGAUAUCCUUUUU UUUUUUCGU GAAAA UGGUAGAAUA GUAGG 231
SEQ5136_1_255_+  AAUUAGU AGUCUUAUGUUAGAUAUCCUUUUU UUUUUUCGU GAAAA UGGUAGAAUA GUAGG 231
SEQ5925_1_255_+  AAUUAGU AGUCUUAUGUUAGAUAUCCUUUUU UUUUUUCGU GAAAA UGGUAGAAUA GUAGG 231
SEQ5270_1_264_+  AAUUAGU AGUCUUAUGUUAGAUAUCCUUUUU UUUUUUCGU GAAAA UGGUAGAAUA GUAGG 240
.....190.....200.....210.....220.....230.....

))))).....))))).....)))))
SEQ219_1_255_+  AAAUAUAAGAGGAAGA GUAAGAGAU 255
SEQ8477_1_255_+  AAAUAUAAGAGGAAGA GUAAGAGAU 255
SEQ5136_1_255_+  AAAUAUAAGAGGAAGA GUAAGAGAU 255
SEQ5925_1_255_+  AAAUAUAAGAGGAAGA GUAAGAGAU 255
SEQ5270_1_264_+  AAAUAUAAGAGGAAGA GUAAGAGAU 264
.....250.....260..
```


Cluster 36

```
... (((... (((((((...))))))))) (((... (((... (((((((... (((...
SEQ1260_1_272_+ CCAUUAUCUAUGCCAGUUAGCAUGAAGCCGCAUUGUGCAAGAUCCGCCCGUUGGUUG 60
SEQ5523_1_272_+ CCAUUAUCUAUGCCAGUUAGCAUGAAGCCGCAUUGUGCAAGAUCCGCCCGUUGGUUG 60
SEQ11608_13_283_+ CCAUUAUCUAUGCCAGUUAGCAUGAAGCCGCAUUGUGCAAGAUCCGCCCGUUGGUUG 60
SEQ20320_1_272_+ CCAUUAUCUAUGCCAGUUAGCAUGAAGCCGCAUUGUGCAAGAUCCGCCCGUUGGUUG 60
SEQ17271_1_272_+ CCAUUAUCUAUGCCAGUUAGCAUGAAGCCGCAUUGUGCAAGAUCCGCCCGUUGGUUG 60
.....10.....20.....30.....40.....50.....
[Bar chart showing sequence conservation from position 10 to 50]

...))))) (((((((... (((... (((... (((((((... (((...
SEQ1260_1_272_+ CAUGUGCCGAUGGCCUGGACCCGUAUAACAUCUUCGGUGUUGCCAGGUUCCGUA 120
SEQ5523_1_272_+ CAUGUGCCGAUGGCCUGGACCCGUAUAACAUCUUCGGUGUUGCCAGGUUCCGUA 120
SEQ11608_13_283_+ CAUGUGCCGAUGGCCUGGACCCGUAUAACAUCUUCGGUGUUGCCAGGUUCCGUA 120
SEQ20320_1_272_+ CAUGUGCCGAUGGCCUGGACCCGUAUAACAUCUUCGGUGUUGCCAGGUUCCGUA 120
SEQ17271_1_272_+ CAUGUGCCGAUGGCCUGGACCCGUAUAACAUCUUCGGUGUUGCCAGGUUCCGUA 120
.....70.....80.....90.....100.....110.....
[Bar chart showing sequence conservation from position 70 to 110]

((((... (((((((... (((... (((... (((((((... (((...
SEQ1260_1_272_+ CGCGAAAGCGGUCCUGUUCGUAGGACUCCUUGGCCUGGUUGUUGUCCAGGACUGA 179
SEQ5523_1_272_+ CGCGAAAGCGGUCCUGUUCGUAGGACUCCUUGGCCUGGUUGUUGUCCAGGACUGA 179
SEQ11608_13_283_+ CGCGAAAGCGGUCCUGUUCGUAGGACUCCUUGGCCUGGUUGUUGUCCAGGACUGA 179
SEQ20320_1_272_+ CGCGAAAGCGGUCCUGUUCGUAGGACUCCUUGGCCUGGUUGUUGUCCAGGACUGA 179
SEQ17271_1_272_+ CGCGAAAGCGGUCCUGUUCGUAGGACUCCUUGGCCUGGUUGUUGUCCAGGACUGA 180
.....130.....140.....150.....160.....170.....
[Bar chart showing sequence conservation from position 130 to 170]

...))))) (((((((... (((... (((... (((((((... (((...
SEQ1260_1_272_+ AUGGAGCAGGUCGUGAGUUGCUGUUGCGUUGGUAAGUAUAAAAGCCCGUUU 239
SEQ5523_1_272_+ AUGGAGCAGGUCGUGAGUUGCUGUUGCGUUGGUAAGUAUAAAAGCCCGUUU 239
SEQ11608_13_283_+ AUGGAGCAGGUCGUGAGUUGCUGUUGCGUUGGUAAGUAUAAAAGCCCGUUU 238
SEQ20320_1_272_+ AUGGAGCAGGUCGUGAGUUGCUGUUGCGUUGGUAAGUAUAAAAGCCCGUUU 239
SEQ17271_1_272_+ AUGGAGCAGGUCGUGAGUUGCUGUUGCGUUGGUAAGUAUAAAAGCCCGUUU 239
.....190.....200.....210.....220.....230.....
[Bar chart showing sequence conservation from position 190 to 230]

((((...))))) (((((((... (((... (((... (((((((... (((...
SEQ1260_1_272_+ CCGGCCCGGGGAAUCCAAAGGAAGAAGUUUC 272
SEQ5523_1_272_+ CCGGCCCGGGGAAUCCAAAGGAAGAAGUUUC 272
SEQ11608_13_283_+ CCGGCCCGGGGAAUCCAAAGGAAGAAGUUUC 271
SEQ20320_1_272_+ CCGGCCCGGGGAAUCCAAAGGAAGAAGUUUC 272
SEQ17271_1_272_+ CCGGCCCGGGGAAUCCAAAGGAAGAAGUUUC 272
.....250.....260.....270.
```

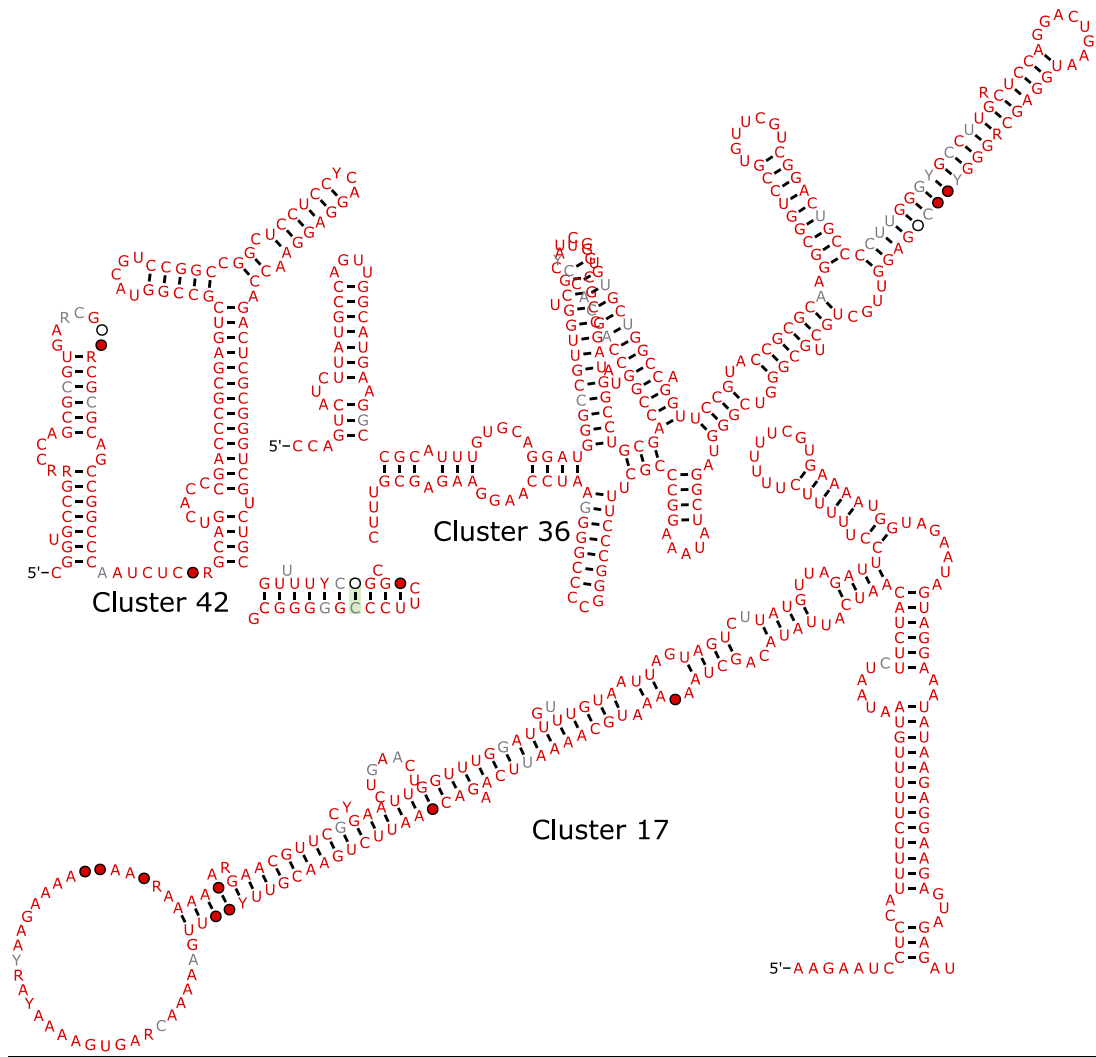



Figure 33 - Structure secondaires de quelques *clusters* sélectionnées

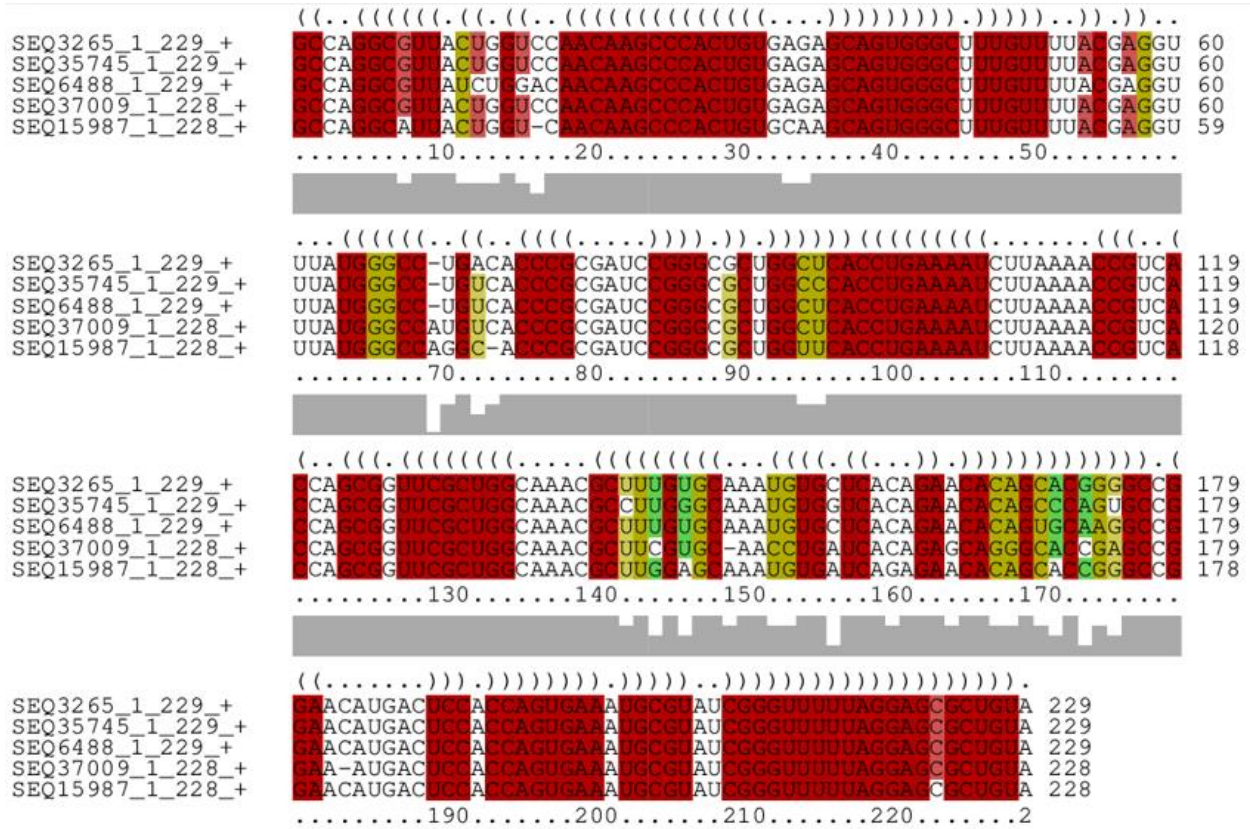
11.3.1.2 Requête complexe

Comme mentionné, GraphClust a été exécuté pour les deux ensembles de données de la requête 1 et 2. Et comme pour le cas précédent, nous avons également sélectionné des *clusters* qui respectaient parfaitement nos critères de recherche.

Cluster 19

```
... (((((((.....)))))). (((.....)))) (((..... ((. (
SEQ18109_198_447_+ UCAUUUCUCAUAAGUUGAGAAGAGCGAAAAAACGAGAAAAUGCGCCACAAAACUUGG 60
SEQ19872_198_447_+ UCAUUUCUCAUAAGUUGAGAAGAGCGAAAAAACGAGAAAAUGCGCCACAAAACUUGG 60
SEQ44639_198_447_+ UCAUUUCUCAUAAGUUGAGAAGAGCGAAAAAACGAGAAAAUGCGCCACAAAACUUGG 60
SEQ2694_198_447_+ UCAUUUCUCAUAAGUUGAGAAGAGCGAAAAAACGAGAAAAUGCGCCACAAAACUUGG 60
SEQ44998_198_447_+ UCAUUUCUCAUAAGUUGAGAAGAGCGAAAAAACGAGAAAAUGCGCCACAAAACUUGG 60
.....10.....20.....30.....40.....50.....
[Grey bar]
(((.....((((.....((((((((((((((((((((.....)))))).....)))))))))
SEQ18109_198_447_+ GAAAAAGUGUGUUUUUCAUUGAGUGAGAAUUGGGUAAUAUCCGAGUAUUGCUUUCUU 120
SEQ19872_198_447_+ GAAAAAGUGUGUUUUUCAUUGAGUGAGAAUUGGGUAAUAUCCGAGUAUUGCUUUCUU 120
SEQ44639_198_447_+ GAAAAAGUGUGUUUUUCAUUGAGUGAGAAUUGGGUAAUAUCCGAGUAUUGCUUUCUU 120
SEQ2694_198_447_+ GAAAAAGUGUGUUUUUCAUUGAGUGAGAAUUGGGUAAUAUCCGAGUAUUGCUUUCUU 120
SEQ44998_198_447_+ GAAAAAGUGUGUUUUUCAUUGAGUGAGAAUUGGGUAAUAUCCGAGUAUUGCUUUCUU 120
.....70.....80.....90.....100.....110.....
[Grey bar]
))))))))) (((((((.....)))))).....))))))))) (((.
SEQ18109_198_447_+ AUUUAUGAGAAUUAAGUGGAUUUGGACCUGA AAAAC CACAGUUCAGUAGGCACUCUG 180
SEQ19872_198_447_+ AUUUAUGAGAAUUAAGUGGAUUUGGACCUGA AAAAC CACAGUUCAGUAGGCACUCUG 180
SEQ44639_198_447_+ AUUUAUGAGAAUUAAGUGGAUUUGGACCUGA AAAAC CACAGUUCAGUAGGCACUCUG 180
SEQ2694_198_447_+ AUUUAUGAGAAUUAAGUGGAUUUGGACCUGA AAAAC CACAGUUCAGUAGGCACUCUG 180
SEQ44998_198_447_+ AUUUAUGAGAAUUAAGUGGAUUUGGACCUGA AAAAC CACAGUUCAGUAGGCACUCUG 180
.....130.....140.....150.....160.....170.....
[Grey bar]
. (((. (((..... (((. (((. (((.....)))))).....)))))) ..
SEQ18109_198_447_+ ACCUGCCUAAAUACCGCUCAAACUUGCGAGUUGUUACUAAAAGGAUGAGAAAAAGGAGGCA 240
SEQ19872_198_447_+ ACCUGCCUAAAUACCGCUCAAACUUGCGAGUUGUUACUAAAAGGAUGAGAAAAAGGAGGCA 240
SEQ44639_198_447_+ ACCUGCCUAAAUACCGCUCAAACUUGCGAGUUGUUACUAAAAGGAUGAGAAAAAGGAGGCA 240
SEQ2694_198_447_+ ACCUGCCUAAAUACCGCUCAAACUUGCGAGUUGUUACUAAAAGGAUGAGAAAAAGGAGGCA 240
SEQ44998_198_447_+ ACCUGCCUAAAUACCGCUCAAACUUGCGAGUUGUUACUAAAAGGAUGAGAAAAAGGAGGCA 240
.....190.....200.....210.....220.....230.....
[Grey bar]
.))))) .....
SEQ18109_198_447_+ AAGAGGUUUU 250
SEQ19872_198_447_+ AAGAGGUUUU 250
SEQ44639_198_447_+ AAGAGGUUUU 250
SEQ2694_198_447_+ AAGAGGUUUU 250
SEQ44998_198_447_+ AAGAGGUUUU 250
.....2
```


Cluster 35



Quelques exemples de structures secondaires obtenues de GraphClust sont indiqués dans la figure suivante.

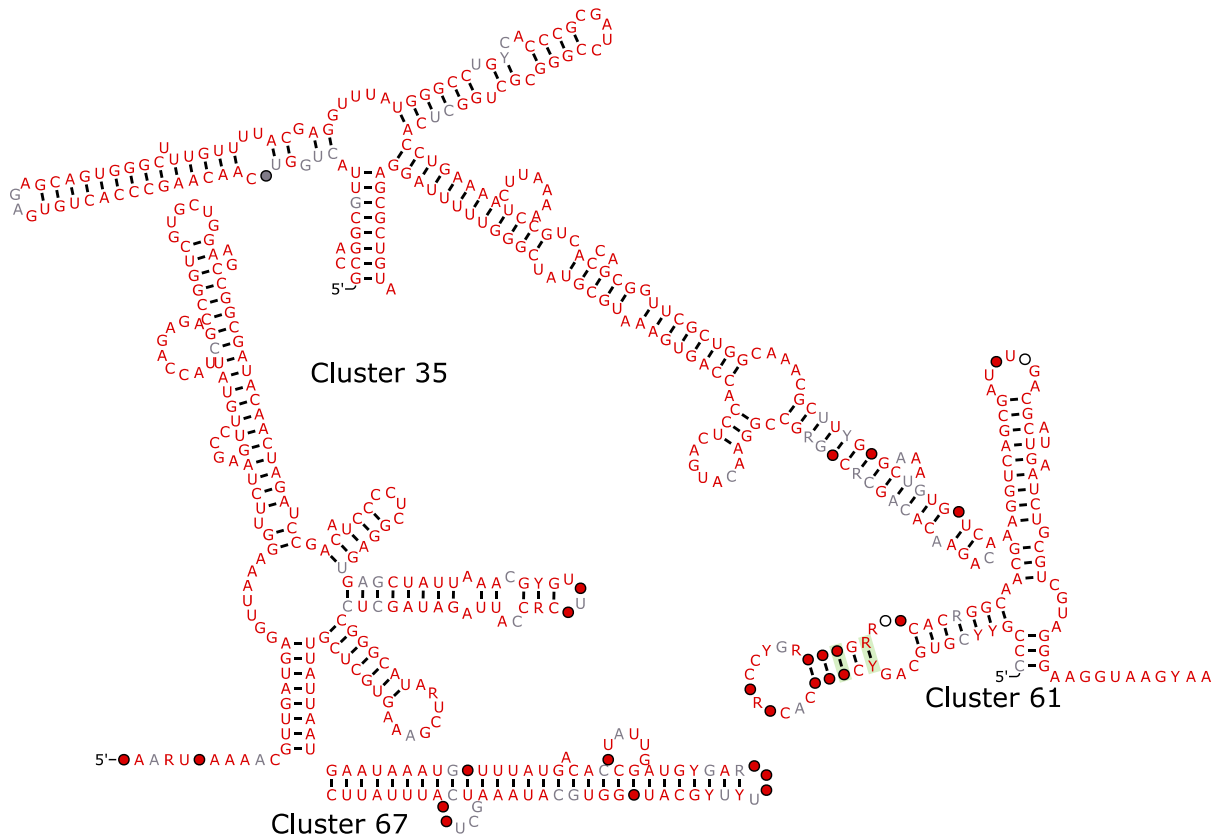


Figure 34 - Structures secondaires de candidats intéressants

11.3.2 Recherche d'homologues avec Infernal

11.3.2.1 Requête simple

Cluster 5

```

...(((...(((...))))))(((...(((...))))))(((...(((...))))))
SEQ273_1_271_+ CCAGUCAUC UUAUGCC AGUU SGCAUGA AGGCCGCAUU GUGCA GGAUGGGCCGUUGGU CG 60
SEQ271_1_272_+ CCAGUCAUC UUAUGCC AGUU SGCAUGA AGGCCGCAUU GUGCA GGAUGGGCCGUUGGU CG 60
SEQ274_1_272_+ CCAGUCAUC UUAUGCC AGUU SGCAUGA AGGCCGCAUU GUGCA GGAUGGGCCGUUGGU CG 60
SEQ275_1_272_+ CCAGUCAUC UUAUGCC AGUU SGCAUGA AGGCCGCAUU GUGCA GGAUGGGCCGUUGGU CG 60
SEQ276_1_271_+ CCAGUCAUC UUAUGCC AGUU SGCAUGA AGGCCGCAUU GUGCA GGAUGGGCCGUUGGU CG 60
.....10.....20.....30.....40.....50.....
[Bar chart showing conservation levels for positions 10-50]

.....))))) (((((((...(((...))))))))) ) (((((((...(((...)))))))))
SEQ273_1_271_+ CAUGU CGCCGAUGGCCUGCGACCG GUA CAGCACCG CGCG GGUUUU GGC CAGGUUCCGUAC 120
SEQ271_1_272_+ CAUGU CGCCGAUGGCCUGCGACCG GUA CCGGCACCUUCG GGUGCC GGC CAGGUUCCGUAC 120
SEQ274_1_272_+ CAUGU CGCCGAUGGCCUGCGACCG GUA CCAGUACCUUCG GGUCUG GGC CAGGUUCCGUAC 120
SEQ275_1_272_+ CAUGU CGCCGAUGGCCUGCGACCG GUA CAGCACCG GUCG GGUUCU GGC CAGGUUCCGUAC 120
SEQ276_1_271_+ CAUGU CGCCGAUGGCCUGCGACCG CUG UCGCACCG CUCG GGUUCC GGC CAGGUUCCGUAC 120
.....70.....80.....90.....100.....110.....
[Bar chart showing conservation levels for positions 70-110]

((((((((...(((...))))))))) ) (((((((...(((...)))))))))
SEQ273_1_271_+ CGCGCAAGGG GGUCC UGUUCGU CGGACUGCC-ACUUG GGCA CCUUG CAUCCAAGGACUGA 179
SEQ271_1_272_+ CGCGCAAGGG GGUCC UGUUCGU CGGACUGCC-CCUUG GGCGCCUUG CAUCCAAGGACUGA 179
SEQ274_1_272_+ CGCGCAAGGG GGUCC UGUUCGU CGGACUGCCUCCAAG GGCGCCUUG CAUCCAAGGACUGA 180
SEQ275_1_272_+ CGCGCAAGGG GGUCC UGUUCGU CGGACUGCCUCCAG GGCGCCUUG CAUCCAAGGACUGA 180
SEQ276_1_271_+ CGCGCAAGGG GGUCC UGUUCGU CGGACUGG-CUUUG GGUGCCUUG CAUCCAAGGACUGA 179
.....130.....140.....150.....160.....170.....
[Bar chart showing conservation levels for positions 130-170]

.....))))) ).....))))) ).....))))) ).....))))) ).....))))) ).....)))))
SEQ273_1_271_+ AUGGAGCAGGGUG-C AAGGUUUGC UGCUGGGGU CGGGUAGGCUAUAAA GGCCCGCUUUC 238
SEQ271_1_272_+ AUGGAGCGGGCGUGCGAGGUUGC UGCUGCCGGU CGGGUAGGCUAUAAA GGCCCGCUUUC 239
SEQ274_1_272_+ AUGGAGCAGGGUG-C GAGGUUUGC UGCUGCCGGU CGGGUAGGCUAUAAA GGCCCGCUUUC 239
SEQ275_1_272_+ AUGGAGCAGGGUG-C AAGGUUUGC UGCUGCCGGU CGGGUAGGCUAUAAA GGCCCGCUUUC 239
SEQ276_1_271_+ AUGGAGCAGGGCC-C GAGGUUUGC UGCUGCCGGU CGGGUAGGCUAUAAA GGCCCGCUUUC 238
.....190.....200.....210.....220.....230.....
[Bar chart showing conservation levels for positions 190-230]

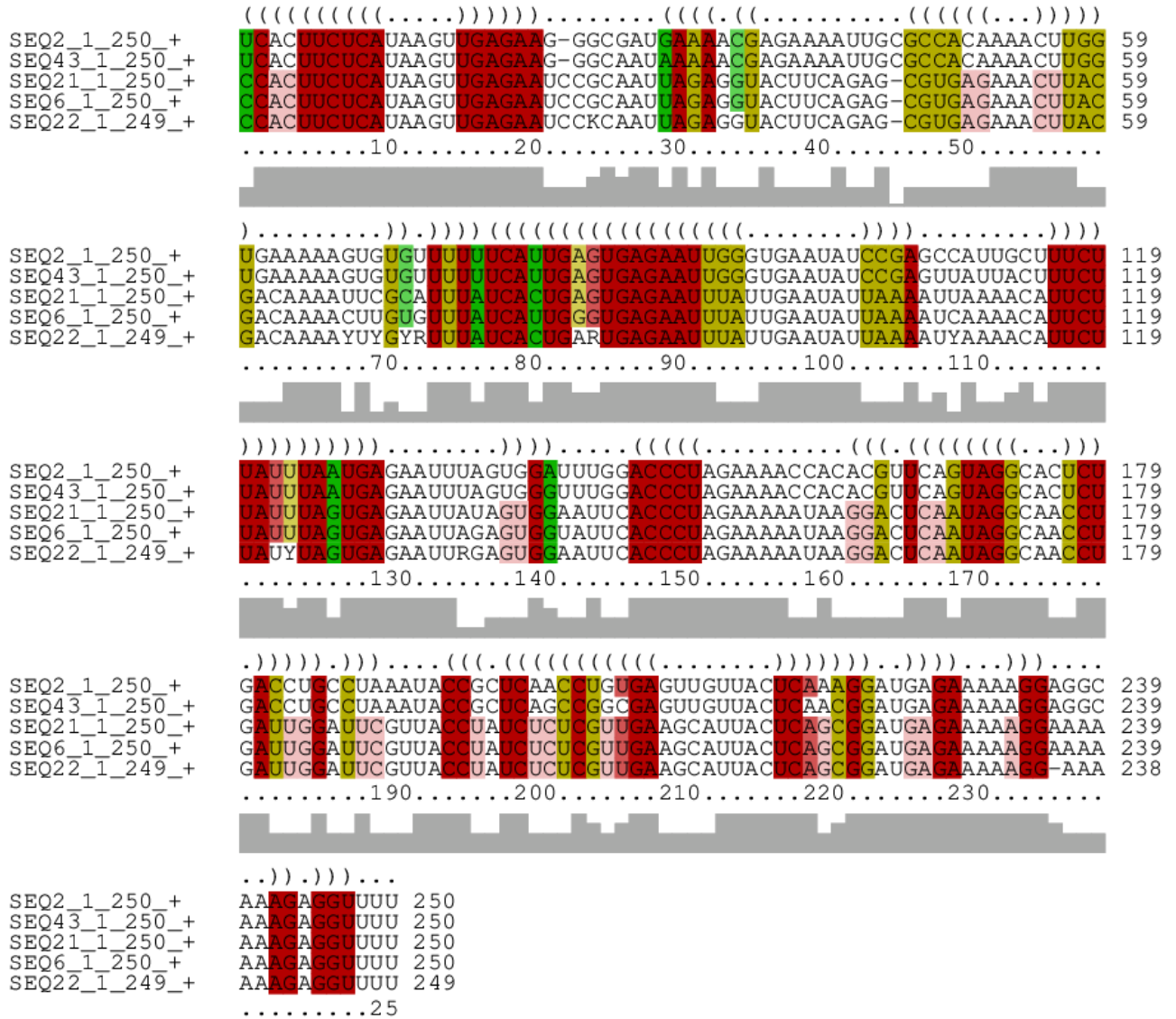
((((...))))) ).....))))) ).....))))) ).....)))))
SEQ273_1_271_+ CCGGGCC CCGGGGAAUCC AAGGAAGAGCGUUC 271
SEQ271_1_272_+ CCGGGCC CCGGGGAAUCC AAGGAAGAGCGUUC 272
SEQ274_1_272_+ CCGGGCC CCGGGGAAUCC AAGGAAGAGCGUUC 272
SEQ275_1_272_+ CCGGGCC CCGGGGAAUCC AAGGAAGAGCGUUC 272
SEQ276_1_271_+ CCGGGCC CCGGGGAAUCC AAGGAAGAGCGUUC 271
.....250.....260.....270.

```

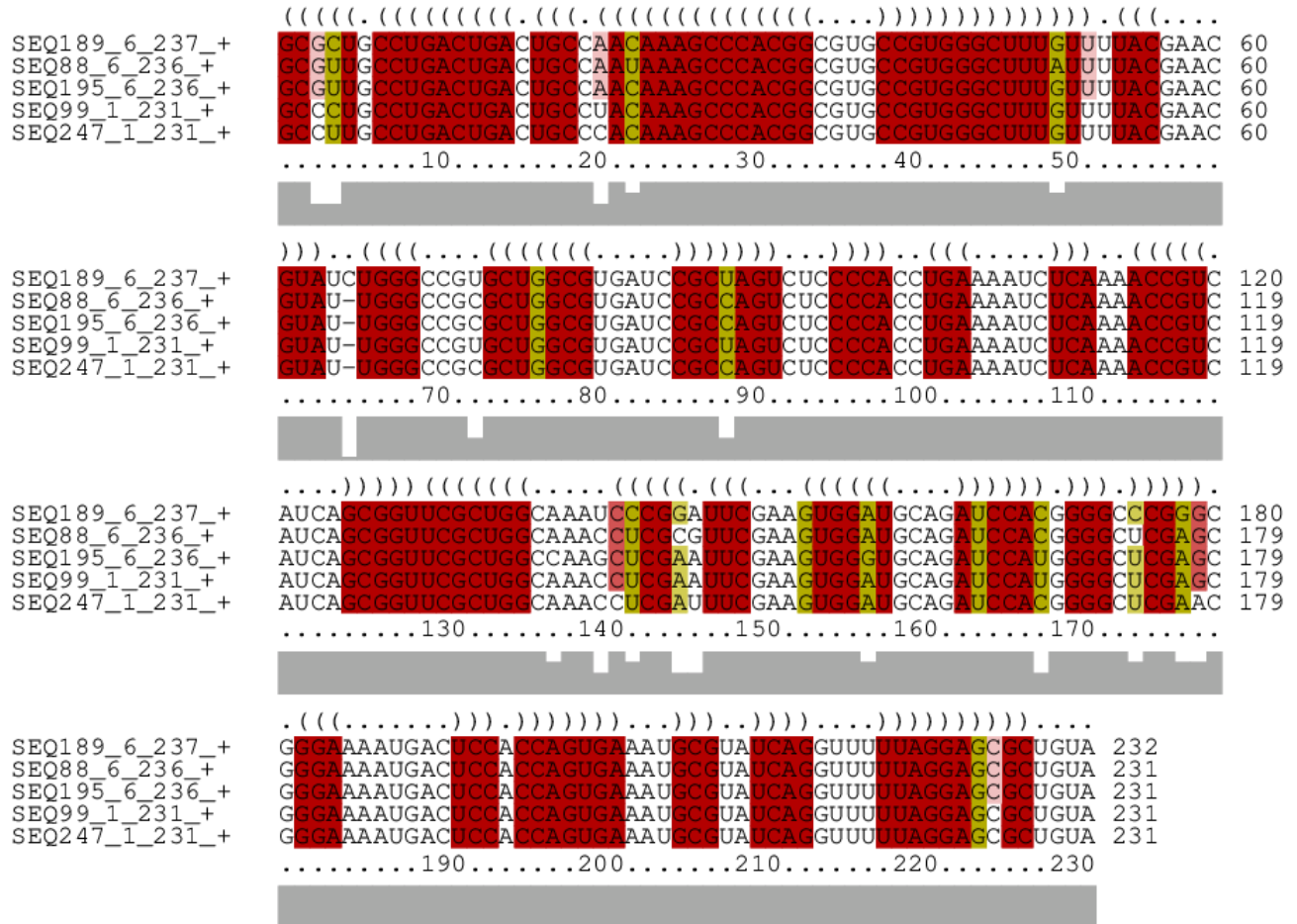
11.3.2.2 Requête complexe

Cluster

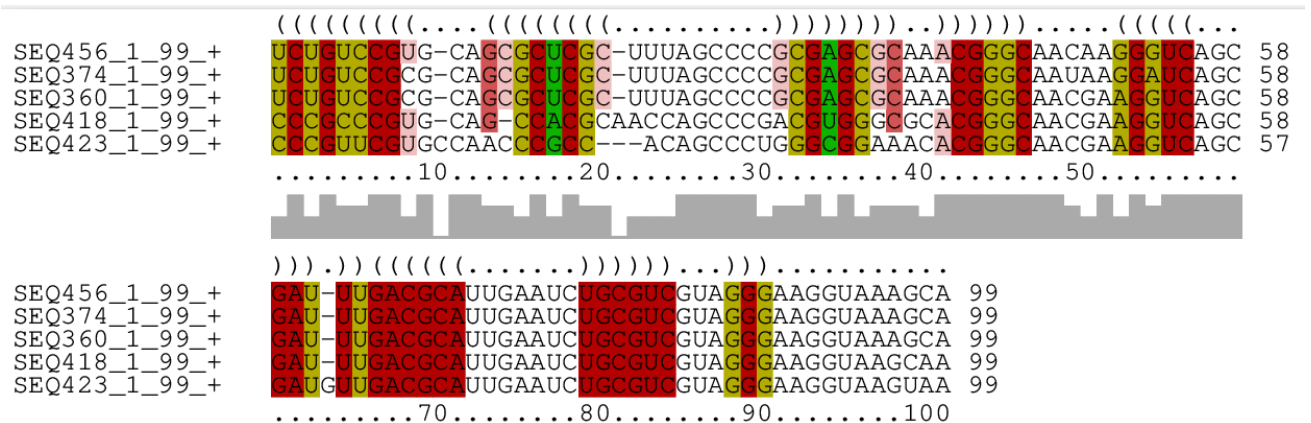
2



Cluster 3

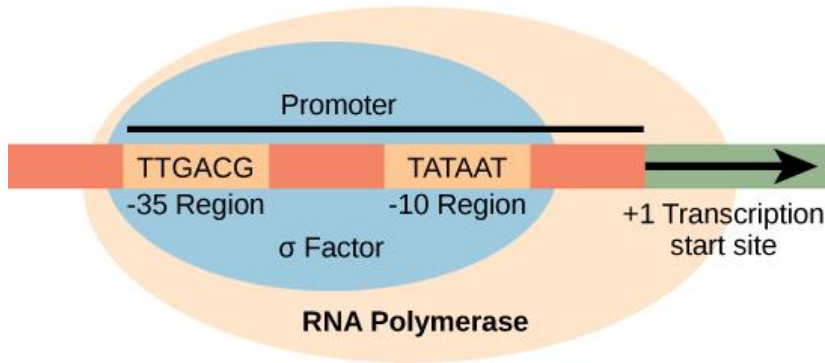


Cluster 4



12 ANNEXE III : DROITS DE REPUBLICATIONS DES IMAGES

Dans ce qui suit, nous présentons les droits de republications des images utilisés dans ce mémoire.



Attribution 4.0 International (CC BY 4.0)

This is a human-readable summary of (and not a substitute for) the [license](#). [Disclaimer](#).

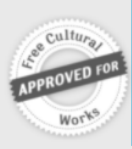
You are free to:

- Share** — copy and redistribute the material in any medium or format
- Adapt** — remix, transform, and build upon the material for any purpose, even commercially.

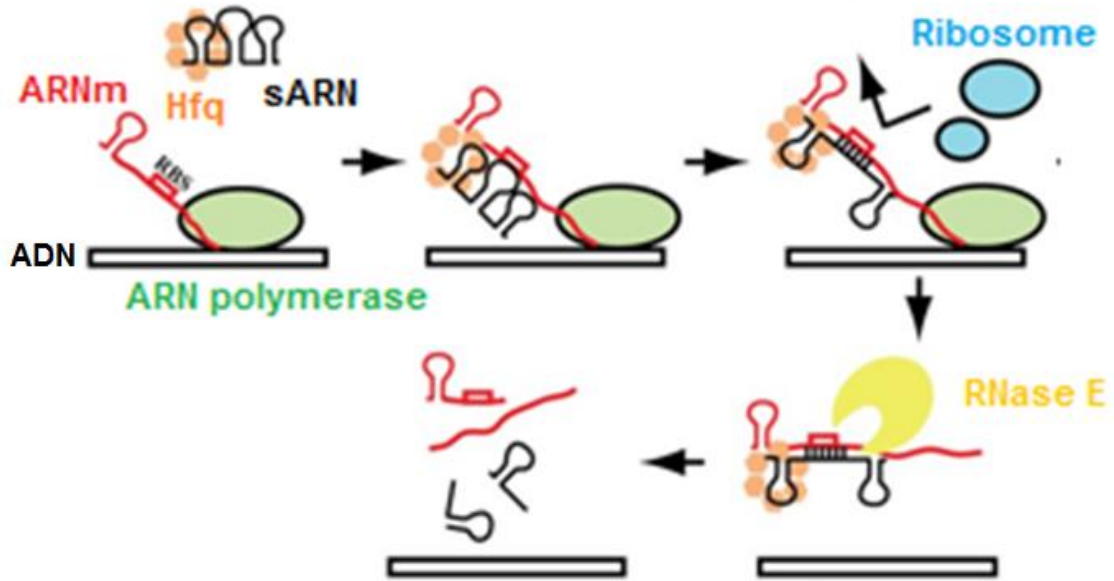
The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

- Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.



<https://creativecommons.org/licenses/by/4.0/>



<https://s100.copyright.com/AppDispatchServlet?publisherName=ELS&contentID=S0021925819334507&orderBeanReset=true&orderSource=Phoenix>

Publisher: Elsevier

Copyright © 1969, Elsevier

Creative Commons

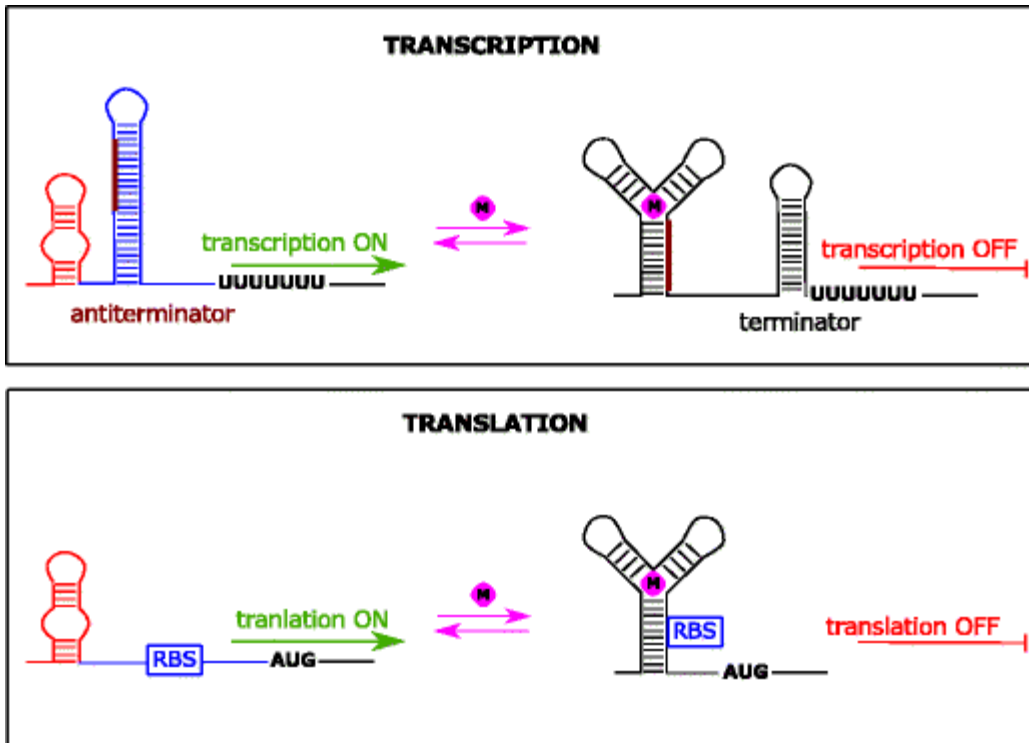
This is an open access article distributed under the terms of the [Creative Commons CC-BY](#) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

To request permission for a type of use not listed, please contact [Elsevier Global Rights Department](#).

Are you the [author](#) of this Elsevier journal article?

© 2021 Copyright - All Rights Reserved | [Copyright Clearance Center, Inc.](#) | [Privacy statement](#) | [Terms and Conditions](#)
 Comments? We would like to hear from you. E-mail us at customer-care@copyright.com



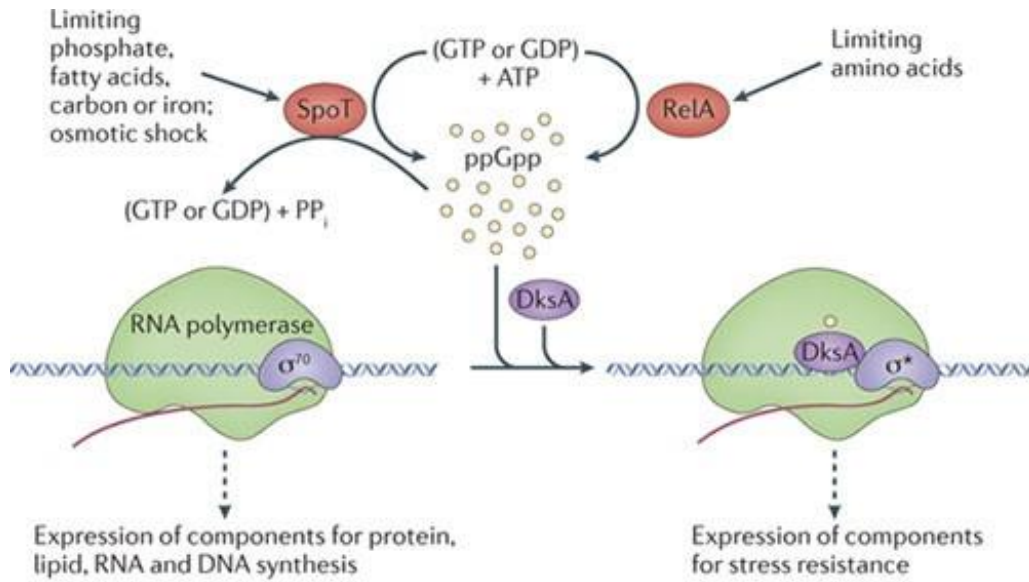
Additional information

Communicated by: Agnieszka Szalewska-Palasz

Rights and permissions

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

[Reprints and Permissions](#)



Nature Reviews | Microbiology

Your confirmation email will contain your order number for future reference.

License Number 5134270550376

License date Aug 22, 2021

[Printable Details](#)

Licensed Content

Licensed Content Publisher Springer Nature
 Licensed Content Publication Nature Reviews Microbiology
 Licensed Content Title ppGpp: magic beyond RNA polymerase
 Licensed Content Author Zachary D. Dalebroux et al
 Licensed Content Date Feb 16, 2012

Order Details

Type of Use Thesis/Dissertation
 Requestor type academic/university or research institute
 Format electronic
 Portion figures/tables/illustrations
 Number of figures/tables/illustrations 1
 High-res required no
 Will you be translating? no
 Circulation/distribution 30 - 99
 Author of this Springer Nature content no

About Your Work

Title EXTENSION DES FONCTIONNALITÉS DE LA BASE DE DONNÉES RIBOGAP POUR LA DÉCOUVERTE DE NOUVELLES STRUCTURES D'ARN NONCODANTS
 Institution name Institut national de la recherche scientifique
 Expected presentation date Sep 2021

Additional Data

Portions Figure 1

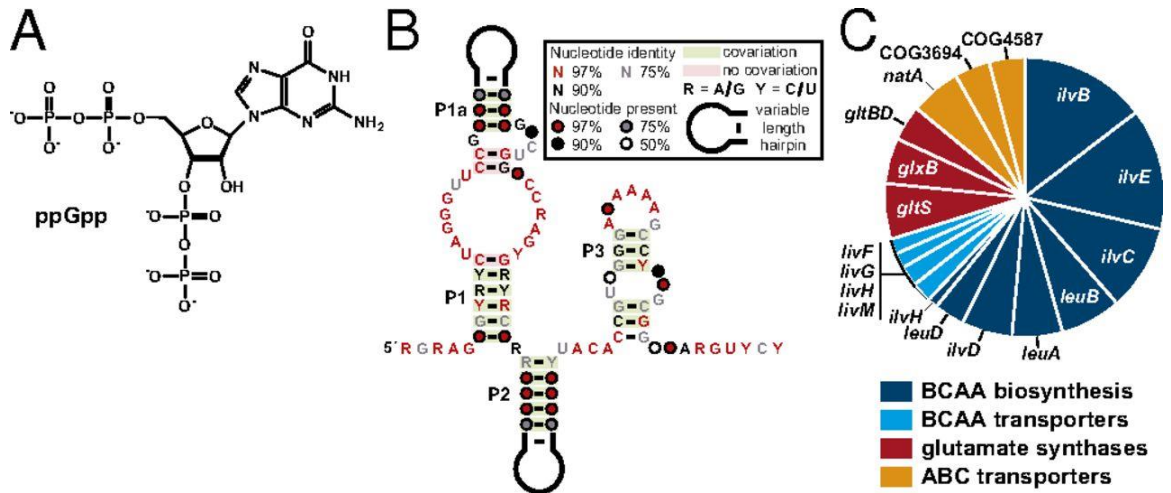
Requestor Location

Samia Djerroud
 5-11645 boul sainte colette

Tax Details

Requestor Location

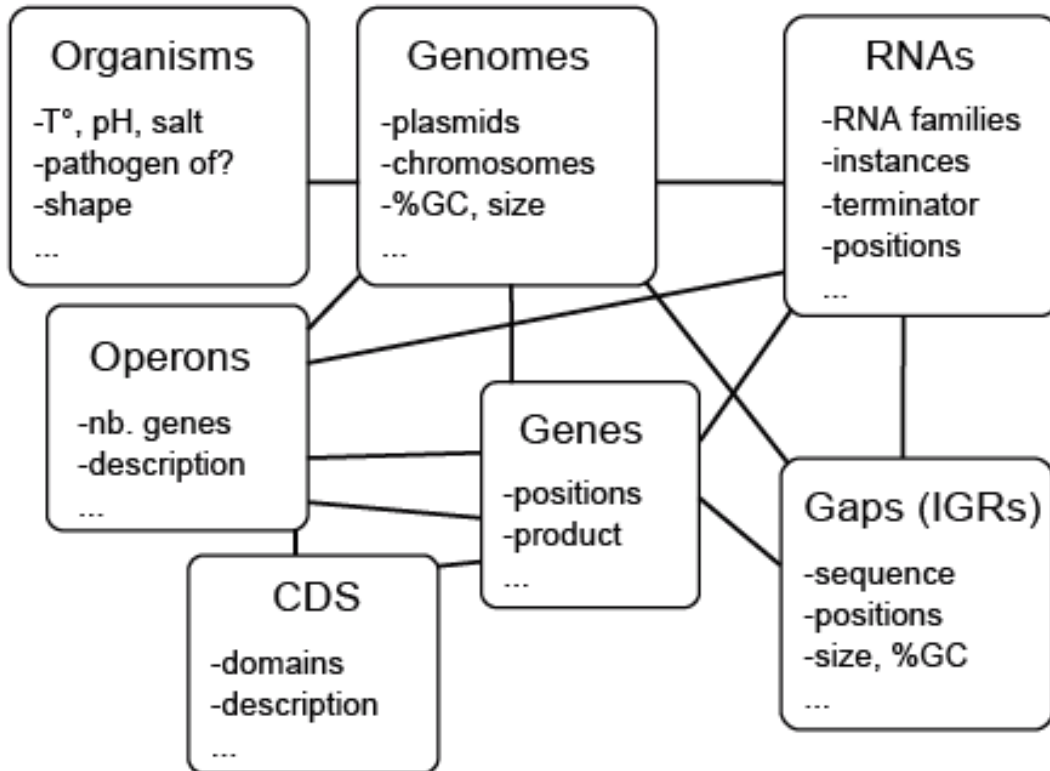
mONTREAL, QC H1G4T9
 Canada
 Attn: Samia Djerroud



PNAS authors need not obtain permission for the following cases:

1. to use their original figures or tables in their future works;
2. to make copies of their articles for their own personal use, including classroom use, or for the personal use of colleagues, provided those copies are not for sale and are not distributed in a systematic way;
3. to include their articles as part of their dissertations; or
4. to use all or part of their articles in printed compilations of their own works.

The full journal reference must be cited and, for articles published in volumes 90–105 (1993–2008), "Copyright (copyright year) National Academy of Sciences" must be included as a copyright note.



Order Completed

Thank you for your order.

This Agreement between Samia Djerroud ("You") and Elsevier ("Elsevier") consists of your order details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License number Reference confirmation email for license number

License date Oct. 27 2021

Licensed Content

Licensed Content Publisher	Elsevier
Licensed Content Publication	Methods
Licensed Content Title	Search for 5'-leader regulatory RNA structures based on gene annotation aided by the RiboGap database
Licensed Content Author	Mohammad Reza Naghdi, Katia Smail, Joy X. Wang, Fallou Wade, Ronald R. Breaker, Jonathan Perreault
Licensed Content Date	15 March 2017
Licensed Content Volume	117
Licensed Content Issue	n/a
Licensed Content Pages	11

About Your Work

Title	EXTENSION DES FONCTIONNALITÉS DE LA BASE DE DONNÉES RIBOGAP POUR LA DÉCOUVERTE DE NOUVELLES STRUCTURES D'ARN NONCODANTS
Institution name	Institut national de la recherche scientifique
Expected presentation date	Nov 2021

Order Details

Type of Use	reuse in a thesis/dissertation
Portion	cover image
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	Yes, including English rights
Number of languages	1

Additional Data

Portions	1
Specific Languages	english

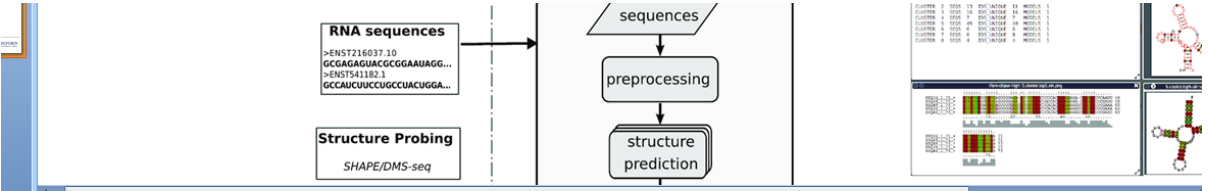
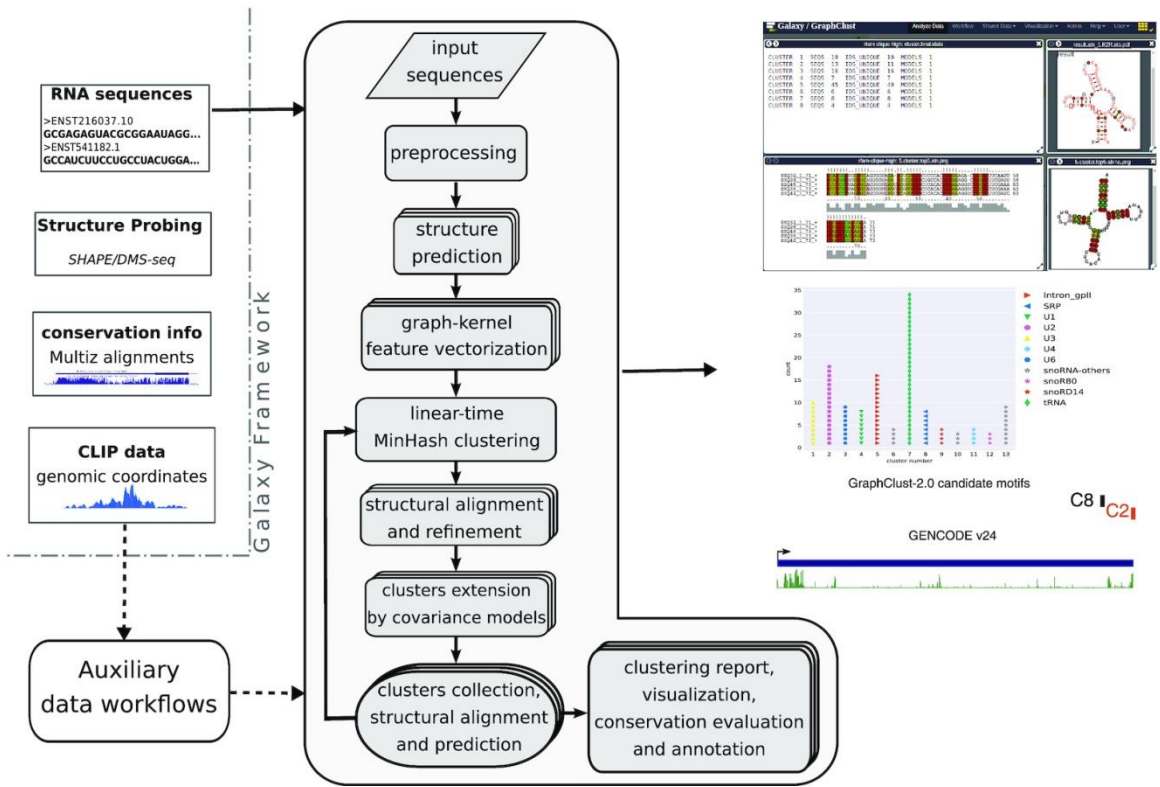


Figure 1 Overview of the GraphClust2 methodology. The flow chart represents the major clustering steps and is supplemented by graphical representations of the associated output data entries. The dashed arrows indicate optional data paths. Auxiliary workflows facilitate integrative clustering of experimental and genomic data including structure-probing raw reads or processed reactivities, genomic alignments and conservation information, and genomic intervals, e.g., from the CLIP experiments. On the right, a sample selection of the clustering outputs including the overview of the clusters, cluster alignment with LocARNA, RNAalfold consensus structure, and R2R [32] visualization and annotation of the cluster structure by R-scape. Clusters can also be visualized and annotated for the orthology structure conservation predictions.

Unless provided in the caption above, the following copyright applies to the content of this slide: © The Author(s) 2019. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

13 ANNEXE IV : MATERIEL SUPPLEMENTAIRE DE L'ARTICLE « RIBOGAP : A RELATIONAL DATABASE FOR PROKARYOTE GENOMICS »

13.1 Table des matières

1. THE FIRST VERSION OF THE DATABASE « RIBOGAP »	161
2. IMPROVEMENT OF THE DATABASE « RIBOGAP »	162
2.1. GENOMES	162
2.1.1 <i>Prokaryotic genomes list download</i>	164
2.1.2 <i>Complete genomes download</i>	165
2.1.3 <i>Incomplete genomes download</i>	166
2.2. CODING SEQUENCES	167
2.3. INTERGENIC REGIONS.....	171
2.3.1. <i>Noncoding RNA</i>	172
2.3.2. <i>Promoters</i>	176
2.3.3. <i>Terminators</i>	177
2.4. CONSERVED PROTEINS	179
3. DATA IN RIBOGAP V2	182
3.1. RIBOGAP V2'S DIAGRAM	182
3.2. COMPILATION OF PROMOTER PREDICTIONS.....	183
TABLE S8 - CALCULATION OF PROMOTERS PER GENES AND OPERONS ACCORDING TO TAXONOMY. ERREUR ! SIGNET NON DEFINI.	
4. THE USE OF THE DATABASE	187
4.1. FIND A PROMOTER, A TERMINATOR AND A RIBOSWITCH IN THE SAME INTERGENIC REGION ...	187
4.2. SEARCH THE EXPRESSION PLATFORMS OF ALL RIBOSWITCHES	191
4.3. DISCOVER NEW SMALL RNAS.....	200
4.4. MOTIF FINDER: G-QUADRUPLEX EXAMPLE	202
4.5. INTERGENIC SEQUENCES WITH CIS REGULATORY RNAS	205

13.2 The first version of the database « RiboGap »

The RiboGap database (<http://ribogap.iaf.inrs.ca/>) which currently exists has the following tables and functionalities.

cdd	Conserved domains	gap5	Sequence information for 5-prime-UTR
<ul style="list-style-type: none"> <input type="checkbox"/> cdd_id <input type="checkbox"/> cdd_accession <input type="checkbox"/> cdd_name <input type="checkbox"/> description 	ex: 116891 example: pfam05377 or cd06578 or TIGR01188... name of conserved domains like HemD nickel ABC transporter, nickel/metallophore periplasmic	<ul style="list-style-type: none"> <input type="checkbox"/> start <input type="checkbox"/> end <input type="checkbox"/> strand <input type="checkbox"/> sequence <input type="checkbox"/> size 	start position of 5 prime-UTR end position of 5 prime-UTR strand direction of the corresponding gene sequence of 5 prime-UTR in same strand as the gene size of 5 prime-UTR
cds	Coding sequence	operon	Operon
<ul style="list-style-type: none"> <input type="checkbox"/> accession <input type="checkbox"/> gene <input type="checkbox"/> DNA_seq <input type="checkbox"/> locus_tag <input type="checkbox"/> product <input type="checkbox"/> translation <input type="checkbox"/> gi <input type="checkbox"/> start <input type="checkbox"/> end <input type="checkbox"/> strand 	protein accession like NP_038276.1 gene name like rhlA DNA sequence of gene (No information available for now) ex:Marme_0002 ex:Mg transporter amino acid sequence protein gi like: 326793324 start position end position, strand direction of the gene relative to chromosome	<ul style="list-style-type: none"> <input type="checkbox"/> operon_id <input type="checkbox"/> operon_name <input type="checkbox"/> locus_tag <input type="checkbox"/> description 	CO143545(Conserved) / KO143545(Known) / PR143545 name of operon like virRS ex:Marme_0001, Marme_0002 arylesterase --> hypothetical protein --> peptidase
fragment	Chromosome information	organism	Organism
<ul style="list-style-type: none"> <input type="checkbox"/> DNA fragment <input type="checkbox"/> frag_gi <input type="checkbox"/> length <input type="checkbox"/> is_circular <input type="checkbox"/> strain <input type="checkbox"/> chromosome <input type="checkbox"/> plasmid <input type="checkbox"/> GC <input type="checkbox"/> gene_num <input type="checkbox"/> IGR_length <input type="checkbox"/> IGR_GC <input type="checkbox"/> IGR_median <input type="checkbox"/> organism_id <input type="checkbox"/> taxonomy <input type="checkbox"/> description 	Refseq accession number like NC_000913 gi of Refseq accession like 158421624 length of chromosome true/false strain information like "Newman" chromosome if exist in database plasmid if exist in database GC % like 52.17 number of genes in chromosome like: 4934 total length of intergenic sequences GC % in intergenic sequences median size of intergenic sequences project id of organism according to NCBI bacteria; elusimicrobia; environmental samples Staphylococcus aureus subsp. aureus str. Newman	<ul style="list-style-type: none"> <input type="checkbox"/> organism_id <input type="checkbox"/> tax_id <input type="checkbox"/> organism <input type="checkbox"/> kingdom <input type="checkbox"/> tax_group <input type="checkbox"/> gram <input type="checkbox"/> shape <input type="checkbox"/> arrangement <input type="checkbox"/> endospore <input type="checkbox"/> motility <input type="checkbox"/> salinity <input type="checkbox"/> oxygen <input type="checkbox"/> habitat <input type="checkbox"/> temp_range <input type="checkbox"/> optimal_temp <input type="checkbox"/> pathogenic_in <input type="checkbox"/> disease 	given Id to Organism like "3" taxonomy id according to NCBI like 224326 name of organism like Borrelia burgdorferi B31 ex:bacteria ex:spirochaetes / gammaproteobacteria/ firmicutes +/- spiral/rod/coccus etc single/pairs/chains no/yes no/yes non-halophilic/halophilic/moderate halophilic anaerobic/aerobic/facultative/microaerophilic aquatic/multiple/host-associated mesophilic/psychrophilic/thermophilic/hyperthermophilic 30-42/25-30/25-40/etc human/animal/plant/insect Q fever
gap3	sequence information for 3-prime-UTR	rna_family	Family of RNA according to Rfam
<ul style="list-style-type: none"> <input type="checkbox"/> start <input type="checkbox"/> end <input type="checkbox"/> strand <input type="checkbox"/> sequence <input type="checkbox"/> size 	start position of 3 prime-UTR end position of 3 prime-UTR strand direction of the corresponding gene sequence of 3 prime-UTR in the same strand as the gene size of 3 prime-UTR	<ul style="list-style-type: none"> <input type="checkbox"/> fam_id <input type="checkbox"/> fam_name <input type="checkbox"/> description <input type="checkbox"/> type <input type="checkbox"/> note 	Rfam accession: RF00001 5S_rRNA 5S ribosomal RNA gene; rRNA some description
		rna_known	Known RNA according to Rfam
		<ul style="list-style-type: none"> <input type="checkbox"/> start <input type="checkbox"/> end <input type="checkbox"/> strand 	start position of RNA end position of RNA strand of RNA

Figure 35—First version of RiboGap tables and attributes (Fig. S1 in the supplementary data)

(A) Write down your own query: (query of MySQL)

(B) Condition:

-	-		-
-	-		-
-	-		-
-	-		-
-	-		-
-	-		-
-	-		-

result number:

(C)

(D) Email:

your email

Figure 36 - The web interface of the database (Fig. S2 in the supplementary data).

(A) Text area where the user can directly write a MySQL query. **(B)** Set conditions to optimize searches by giving a value. **(C)** Set the maximum number of the result hits to display in the browser. **(D)** Set the mail to send a link to retrieve results once the query is done

13.3 Improvement of the database « RiboGap »

13.3.1 Genomes

Two new database versions (RiboGap-complete genomes and RiboGap-incomplete genomes) have been produced knowing that there are two types of genome sequences available: fully assembled genomes (complete genomes) and partially assembled genomes (incomplete genomes)

- Complete genomes

There are 13,908 chromosomes and plasmids corresponding to 9,857 complete genomes in NCBI at the time they were downloaded.

>Genome1

```
CGATTAAAGATAGAAATACACGATGCGAGCAATCAAATTCATAACATCACCATGA
GTTTGGTCCGAAGCATGAGTGTTTACAATGTTTCGAACACCTTATACAGTTCTTATAC
ATACTTTATAAATTATTTCCCAAAGTGTGTTGATACACTCACTAACAGATACTCTATA
GAAGGAAAAGTTATCCACTTATGCACATTTATAGTTTTTCAGAATTGTGGATAATTAG
AAATTACACACAAAGTTATACTATTTTTAGCAACATATTCACAGGTATTTGACATAT
AGAGAAGTAAAAAGTATAATTGTGTGGATAAGTCGTCCAACCTCATGATTTTATAAG
GATTTATTTATTGATTTTTACATAAAAATACTGTGCATAACTAATAAGCAAGATAAAA
GTTATCCACCGATTGTTATTAAGTGTGGATAATTATTAACATGGTGTGTTTAGAAGT
TATCCACGGCTGTTATTTTTGTGTATAACTTAAAAATTTAAGAAAGATGGAGTAAAT
TTATGTCCGAAAAAGAAATTTGGGAAAAAGTGCTTGAAATTGCTCAAGAAAAATTA
TCAGCTGTAAGTTACTCAACTTTCCTAAAAGATACTGAGCTTTACACGATTAAGAT
GGTGAAGCTATCG
```

- Incomplete genomes:

There were 45,122 incomplete prokaryotic genomes in NCBI at the time at which they were downloaded for RiboGap. The N sequences corresponds to nucleotides of unknown sequence.

>Genome2

```
GCATATCCGCGTCAAATAGCTATGTACTTGTCTAGAGAGCTTACAGATTTCTCATT
CCTAAAATTGGTGAAGAATTTGGTGGGCGTGATCATACGACCGTCATTCATGCTCAT
GAAAAAATATCTAAAGATTTAAAAGAAGATCCTATTTTTAAACAAGAAGTAGAGAA
NNNNNNNNNNNNNNNNNNNAATGTATAAGTAGGAACTTTGGGAAATGTAATCTGTTA
TATAACAGCACTAATGATAACAATCATTTTTTACATTTCTATATGCTAATGTGGCAA
GATGAGCAAACTCATTTTGTGGATAATGTTTAAAAGTCATACACACCATACACAAG
TTATCAACATGTGTATAACTTCGCCAAATCTATGTTTTTAAGACATAATTATATATAA
ACGACTGGAAGGAGTTTTAATTAATGATGGAATTCACTATTAANNNNNNNNNNNNNT
TACACAATTAATGACACATTAAGCTATTTACCAAGAACAACATTACCTATATT
AACTGGTATCAAATCGATGCGAAAGAACATGAAGTTATATTAAGTGGTTCAGACTC
TGAAATTTCAATAGAA
```

13.3.1.1 Prokaryotic genomes list download

The exhaustive list of complete and incomplete prokaryotic genomes gathered in the "prokaryotes.txt" file (separate file added as a material supplementary) was downloaded from the NCBI's ftp (File Transfer Protocol) site. The genomic data found in the file are the following:

Tableau 6- Description of a GenBank format (Table S1 in the supplementary data)

Column	Name of the column	Description
1	Organism	Name of the Organisms, ex: <i>Escherichia coli</i> IAI39
2	Taxonomy ID	Ex: 585057
3	Bioproject accession	Ex: PRJNA33411
4	Bioproject ID	Ex: 33411
5	Organism Group	Ex: Proteobacteria
6	Organism Subgroup	Ex: Gammaproteobacteria
7	Size (MP)	Length of genome, Ex: 5.13207
8	GC%	Percent of GC in the genome, Ex: 50.6
9	Replicons	Ex: chromosome:NC_011750.1/CU928164.2
10	WGS	Prefix for whole genome shotgun, Ex: MKVH01
11	Scaffolds	Number of scaffolds, Ex: 1
12	Genes	Number of genes, Ex: 5092
13	Proteins	Number of proteins, Ex: 4725
14	Release data	Ex: 2008/12/16
15	Modify data	Ex: 2016/08/28
16	Statut	The level of assembly, ex: Complete Genome
17	Center	The institute that did the assembly, Ex: Genoscope
18	Biosample Accession	ID of biological Sample in the biosample database, Ex: SAMEA3138234
19	Assembly Accession	Identifier of the genome assembly, Ex: GCA_000026345.1
20	Reference	Ex: REFR
21	Ftp path	Refseq ftp path
22	Pubmed ID	Ex: 19165319
23	Strain	Strain name of sub-species classification, Ex: IAI39

This list was used to fill the data of the genomes table in RiboGap v2 and v2.1. Each column of the table corresponds to a column of the file.

The same “prokaryotes.txt” file was then used to download the different GenBank and fasta files of all the complete and incomplete genomes in order to get other genomic data which will be used in the rest of the project. The download of the GenBank and fasta files was respectively done using the shell language with the following commands.

13.3.1.2 Complete genomes download

- List the FTP path (column 21) from the prokaryotes file, that have "Complete Genome" in the assembly_level (column 16), by using the **awk** command:

```
awk -F "\t" '$16=="Complete Genome" {print $21}' prokaryotes.txt > ftpdirpaths_complete.txt
```

- Collect and modify the FTP URLs to point to the GenBank/fasta files

```
sed -r 's|(ftp://ftp.ncbi.nlm.nih.gov/genomes/all/./+)(GCF_+)|\1\2\2_genomic.gbff.gz|' ftpdirpaths_complete.txt > genomic_GenBank_complete.txt
```

or

```
sed -r 's|(ftp://ftp.ncbi.nlm.nih.gov/genomes/all/./+)(GCF_+)|\1\2\2_genomic.fna.gz|' ftpdirpaths_complete.txt > genomic_fasta_complete.txt
```

- Download the data file for each FTP path in the list

```
wget --input genomic_GenBank_complete.txt
```

or

```
wget --input genomic_fasta_complete.txt
```

13.3.1.3 Incomplete genomes download

The same previous file `prokaryotes.txt` is used to list the incomplete genomes that can be: contigs, scaffolds or chromosomes

- List the FTP path (column 21) from the `prokaryotes` file, that have "Contigs" in the `assembly_level` (column 16), by using the **awk** command:

```
awk -F "\t" $16=="Contigs" {print $21}' prokaryotes.txt > ftpdirpaths_contig.txt
```

- List the FTP path, that have "Scaffolds" in the `assembly_level`:

```
awk -F "\t" $16=="Scaffolds" {print $21}' prokaryotes.txt > ftpdirpaths_scaffold.txt
```

- List the FTP path, that have "Chromosomes" in the `assembly_level`:

```
awk -F "\t" $16=="Chromosomes" {print $21}' prokaryotes.txt > ftpdirpaths_chromosome.txt
```

- Assemble the three files

```
cat ftpdirpaths_* > ftpdirpaths_incomplete.txt
```

- Collect and modify the FTP URLs to point to the GenBank/fastq files

```
sed -r 's|(ftp://ftp.ncbi.nlm.nih.gov/genomes/all/./+)(GCF_./+)\|1\2\2_genomic.gbff.gz|ftpdirpaths_incomplete.txt> genomic_GenBank_incomplete.txt
```

or

```
sed -r 's|(ftp://ftp.ncbi.nlm.nih.gov/genomes/all/./+)(GCF_./+)\|1\2\2_genomic.fna.gz|ftpdirpaths_incomplete.txt> genomic_fasta_incomplete.txt
```

- Download the data file for each FTP path in the list

```
wget --input genomic_GenBank_incomplete.txt
```

or

```
wget --input genomic_GenBank_incomplete.txt
```

13.3.2 Coding sequences

Using the GenBank files of the genome, a perl program was created to extract all the coding sequences annotated CDS in the GenBank file.

```
source      1..2821361
            /organism="Staphylococcus aureus subsp. aureus NCTC 8325"
            /mol_type="genomic DNA"
            /strain="NCTC 8325"
            /sub_species="aureus"
            /db_xref="taxon:83261"
gene        517..1878
            /locus_tag="SAOUHSC_00001"
CDS         517..1878
            /locus_tag="SAOUHSC_00001"
            /codon_start=1
            /transl_table=11
            /product="chromosomal replication initiator protein DnaA"
            /protein_id="ABD29192.1"
            /translation="MSEKEIWEKVL EIAQEKLSAVSYSTFLKDTELYTIKDG EAI VLS
SIPFNANWLNQQYAEIIQAILFDVVGVEVKPHFITTEELANYSNNETATPKETTKPST
ETTEDNHVLGREQFNAHNTFDTFVIGPGNRFPHAASLAVAEAPAKAYNPLFIYGGVGL
GKTHLMHAIGHVLDNNDPAKVIYTSSEKFTNEFIKSIRDNEGEAFREYRNRNIDVLLI
DDIQFIQNKVQTQEEFFYTFNELHQNNKQIVISSDRPPKEIAQLEDRLRSRFEWGLIV
DITPPDYETRMALQKKIEEEKLDIPPEALNYIANQIQSNIRELEGALRLLAYSOLL
GKPITTELTAELKDIIQAPKSKKITIQDIQKIVGQYNNVRIEDFSAKKRKTSIAYPR
QIAMYLSRELTD FSLPKIGEEFPGGRDHTTVIHAHEKISKDLKEDPIFKQEVENLEKEI
ENV"
```

Figure 37 - The coding sequence extracted from a GenBank file (Fig. S3 in the supplementary data)

The extracted data are obtained in the form of the following table.

Tableau 7 - Coding sequences extracted (Table S2 in the supplementary data)

Accession	ABD29192.1
Gene	-
DNA_seq	The sequence gotten from the whole genome using the position start/end of the cds
Locus tag	SAOUHSC_00001
Product	Chromosomal replication initiator protein DnaA
Translation	MSEKEIWEKVVLEIAQEKL SAV...EKEIRNV
Start	517
End	1878
Strand	+1

The following is the Perl program used to extract information about coding sequences from the GenBank files.

```
#!/usr/bin/perl;

##### variables and arrays definition #####

$seqio_obj = Bio::SeqIO->new(-file => $filename, -format => "GenBank" );

while ($seq_obj = $seqio_obj->next_seq){

    $accession=$seq_obj->accession_number;
    $version=$seq_obj->seq_version ;
    $accession=$accession.".".$version;
    $is_circular = $seq_obj->is_circular ;
    if ($is_circular ){
    print OUT_fragment_update "True" ,"\t", $accession, "\n";
    }else {
    print OUT_fragment_update "False", "\t", $accession, "\n";
    }
}
```

```
#####creating the CDS table#####
```

```
for my $feat ($seq_obj->get_SeqFeatures) {
```

```
    if ( $feat->primary_tag eq 'CDS' ) {
```

```
        push @gene_reg,$feat->strand, $feat->start,$feat->end;
```

```
        if (eval{$feat->has_tag("gene")}){
```

```
            # push @gen_inf,$feat->get_tag_values("gene");
```

```
        }else {
```

```
            # push @gen_inf,"null";
```

```
        }
```

```
        if (eval{$feat->has_tag("locus_tag")}){
```

```
            push @gen_inf,$feat->get_tag_values("locus_tag");
```

```
        }else {
```

```
            push @gen_inf,"null";
```

```
        }
```

```
        if (eval{$feat->has_tag("product")}){
```

```
            push @gen_inf,$feat->get_tag_values("product");
```

```
        }else {
```

```
            push @gen_inf,"null";
```

```
        }
```

```
        if (eval{$feat->has_tag("protein_id")}){
```

```
@tmp_protein_id=$feat->get_tag_values("protein_id"); ##### some protien id has tow different id  
so we take just the id of one value
```

```
unless (scalar @tmp_protein_id==1){
```

```
$waste_index=firstidx {$_=~/[A-Z][A-Z][A-Z].*/g}@tmp_protein_id;
```

```
@extra=splice (@tmp_protein_id,$waste_index,1);
```

```
$tmp_prot_id=pop @tmp_protein_id;
```

```
push @gen_inf,$tmp_prot_id;
```

```
    }else{
```

```
push @gen_inf,$feat->get_tag_values("protein_id");
```



```

$coding_seq=coding_sequence( \ $start_cds,\ $end_cds,\ $strd_cds,\ $seq_obj);

}
}catch{
print OUT_error "The subrootin coding_sequence has got this problem $_", "\n";
};
push @gen_inf,$coding_seq;
unshift @protein_id,$accession;
  @data_of_accession_protein_id=@protein_id;
push (@gene_reg,@data_of_accession_protein_id);#### we collect the fragment and
prot_accession
push @gi,$gi,$feat->start,$feat->end,$feat->strand;

push @new_table,@protein_id,@gen_inf,@gi;

if (scalar @new_table>11){
@extra=splice (@new_table,5,1);
}
print OUT_cds_filename join ("\t",@new_table),"\n";
  @gen_inf=();
  @protein_id=();
  @new_table=();
  @gi=();  }
}

```

13.3.3 Intergenic regions

Running a perl program on the GenBank files, the positions of upstream (5') and downstream (3') intergenic regions were defined for each coding sequence, this was done by finding the intergenic region between two coding sequences whose positions are known.

```

#!/usr/bin/perl;

#### variables and arrays definition#####
$seqio_obj = Bio::SeqIO->new(-file => $filename, -format => "GenBank" );
#####starting the processe to IGR#####
my $error=try {

    (@gap5)=IGR_finder (\@gene_reg,\$seq_obj);
}catch{
    print OUT_error $filename, "has got error:", "\n";
    print OUT_error $_, "\n";
};
my $count=1;

foreach my $gap5 (@gap5){
    print OUT_gap5_filename $gap5, "\t";
    unless ($count % 7){
        print OUT_gap5_filename "\n";
    }
    $count++;
}
    print OUT_check_file $filename, " has finished", "\n";
@gene_reg=();
@gap5=();

}
print "End Normal of program", "\n";

```

13.3.4 Noncoding RNA

A noncoding RNA (ncRNA) is transcribed from DNA but not translated into proteins. In general, ncRNA's function is to regulate gene expression at the transcriptional and post-transcriptional level. There are different ncRNA families in Rfam, such as: tRNA, rRNA, glmS, terC, ...etc .

13.3.4.1 Noncoding RNA families

BLAST method presented some false hits because the results were not confirmed using Infernal. The Figure S6 bellow explains how some ncRNA hit gotten by BLAST is not recognized by Rfam and this may pose a database trust problem.

locus_tag	Protein product	cds_start	Cds_end	RNA_family	Intergenic_sequence_gap5
CLOSACC_RS03930	molecular chaperone DnaJ	872492	872932	glmS	TTGTAATTTAGACTTGTGTTTATAAGTAACTAAAATAAGTA ACATTATGTGTTCTGTTTTCTCGGAATTTAGATGGGGATT CTACTACAAGAAGTTGACTAAAATAATTTTTAAAGACAA GCTTTGAAAAATTAATTTAGAAAGTTTAAAGGAAATCGACI CACGTTTCGTTGCTGAGTAAGTTCAATTTACCAGATCATAGA TTTGGAAATTTCACTTAAGAATCACTACAATCAAATTTAATG AAACACTCTACATATAATGTTACTATTTCTACTGTAGATAGT TATCTAACTCTAGTAAATTTCAAAGTAACTGTAATTTGA ATCAAGTGGGGATAAA

(A)

Powered by RNAcentral | Local alignment using nhmmer

```
>CLOSACC_RS03930
UUGUAAUUUAGACUUGUUGUUUAUAAGUAAUAAAUAAGUAAAUUUGUUGUUUCUCUGGAAUUUAGAUG
GGGGAUUCUACUACAAGAAGUUGUACCUAAAUAUUUUUUAAAGACAAGCUUUGAAAAUUUUUAGAAAGUUUA
AAGGAAACUCGACUCACGUUCGUUCGUGAGUAAAGUUCAAUUUACCAGAUCAUAGAUUUGAAUUUCACUUUAGAAU
CACUACAAUCAAAAUUUUAAUGAAACACUCUACAUAUAAUGUUACUUUUUCUACUGUAGAUUUUACUUAACUUCUA
GUAAAUUUCAAACUGUAAACUGUUAAUUGAAUCAAGUGGGGGAUAAA
```

Examples: lysine riboswitch 16S SNORD3A

Rfam classification ?

The query sequence did not match any Rfam families.

(B)

CLOSACC_RS03930
Sequence ID: Query_26927 Length: 356 Number of Matches: 1

Range 1: 151 to 226 [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
97.8 bits(107)	7e-25	67/76(88%)	0/76(0%)	Plus/Plus

```
Query 131 TTTAAAGGAACTCGACTCAGTTCGTTTCGCTGAGTAAATTTCAACTATCCAAATCGAAGA 190
Sbjct 151 TTTAAAGGAACTCGACTCAGTTCGTTTCGCTGAGTAAATTTCAACTATCCAGATCATAGA 210

Query 191 TTTGGAGTTTCATTTA 206
Sbjct 211 TTTGGAAATTTCACTTA 226
```

(C)

(A) –RiboGap’s result shows an association between the intergenic region located upstream of the locus tag "CLOSACC_RS03930" with the ncRNA family *glmS*. (B) - Fetching the

intergenic sequence of A in Rfam shows no hit. (C) – Blast result shows alignment between *glmS* and the intergenic sequence

Figure 38 - An example of a false positive given while using blast. (Fig. S4 in the supplementary data)

An insertion in RNA (like a mobile element) can be the cause of the BLAST hit for which cmsearch from Infernal finds no hit. The only way to avoid this kind of false hit is to use Infernal, a bioinformatic tool based on **covariance models** (CM) that searches the correspondence between nucleotide sequence and structure/sequence of the RNA.

Here are the steps followed to associate intergenic regions and ncRNAs:

Step1: All the covariance models (cm) files of the ncRNA families were downloaded from Rfam.

```
wget ftp://ftp.ebi.ac.uk/pub/databases/Rfam/14.1/Rfam.tar.gz
```

Step2: Filter the list of the ncRNA families and keep only the prokaryotic families. By also deleting the families that correspond to tRNA and rRNA because different tools will be used to generate the hits of these two types of families.

Step3: Execute the function cmsearch of Infernal using fasta files of complete/incomplete genomes and cm files of ncRNA families previously downloaded to find out the positions of ncRNA in genomes. The cmsearch command line is presented as below:

```
cmsearch -E 100 --tblout out/outputs.tab Rfam/files.cm all_complete_genomes.fasta
```

```
cmsearch    search CM(s) against a sequence database
```

```
-E          report the E-value threshold in output
```

```
--tblout    save parseable table of hits to file <s>
```

13.3.5 tRNA

The tRNA family is part of the list of noncoding RNAs that have not been executed with Infernal (in the previous step), because for this case the list of tRNA will be more exhaustive by using another tool "tRNAscan".

```
~/bin/tRNAscan-SE -B inputfile -o outputfile
```

```
-B search for bacterial tRNAs
```

The output of tRNAscan is obtained

Sequence Name	tRNA #	tRNA Begin	Bounds End	tRNA Type	Anti Codon	Intron Begin	Bounds End	Inf Score	Note
NZ_LR699010.1	1	162551	162623	Thr	GGT	0	0	55.9	
NZ_LR699010.1	2	171523	171595	Lys	CTT	0	0	78.3	
NZ_LR699010.1	3	205103	205177	Pro	TGG	0	0	75.9	
NZ_LR699010.1	4	205203	205276	Arg	TCT	0	0	77.2	
NZ_LR699010.1	5	205284	205354	Gly	TCC	0	0	77.4	
NZ_LR699010.1	6	205383	205456	His	GTG	0	0	74.3	
NZ_LR699010.1	7	205464	205535	Gln	TTG	0	0	61.6	
NZ_LR699010.1	8	489120	489192	Gly	GCC	0	0	77.6	
NZ_LR699010.1	9	533546	533617	Arg	ACG	0	0	46.6	
NZ_LR699010.1	10	533664	533747	Tyr	GTA	0	0	58.0	
NZ_LR699010.1	11	757054	757136	Leu	CAA	0	0	71.7	
NZ_LR699010.1	12	1205548	1205731	Leu	CAG	0	0	63.8	

Figure 39 - tRNAscan output (Fig. S5 in the supplementary data)

13.3.6 rRNA

rRNA ribosomal is a ncRNA family, Infernal was not used in this case to get the positions of rRNA in genomes because it is an information that already exist in the GenBank files of genomes previously downloaded.

```
gene complement(2712258..2712377)
/ gene="5S_rRNA"
rRNA complement(2712258..2712377)
/ gene="5S_rRNA"
/ note="hit to 5S_rRNA 1..120 score: 582 percent id: 98.33"
gene complement(2712478..2715483)
/ gene="23S_rRNA"
rRNA complement(2712478..2715483)
/ gene="23S_rRNA"
/ note="hit to 23S_rRNA 487..2904 score: 11314 percent id:
96.69;
hit to 23S rRNA 1..540 score: 2601 percent id: 97.96"
tRNA complement(2715666..2715741)
/ product="tRNA-Ala"
/ note="tRNA Ala anticodon TGC, Cove score 88.77"
tRNA complement(2715853..2715929)
/ product="tRNA-Ile"
/ note="tRNA Ile anticodon GAT, Cove score 88.37"
gene complement(2715999..2717540)
```

Figure 40 - rRNA positions found in GenBank (Fig. S6 in the supplementary data)

13.3.7 Promoters

In bacteria, promoters have a structure where two sequence motifs (boxes -35 and -10) are recognized by sigma factors to allow initiation of transcription with the RNA polymerase. Putative promoters were predicted from all IGRs with bTSSfinder according to the following scheme (Fig. S7).

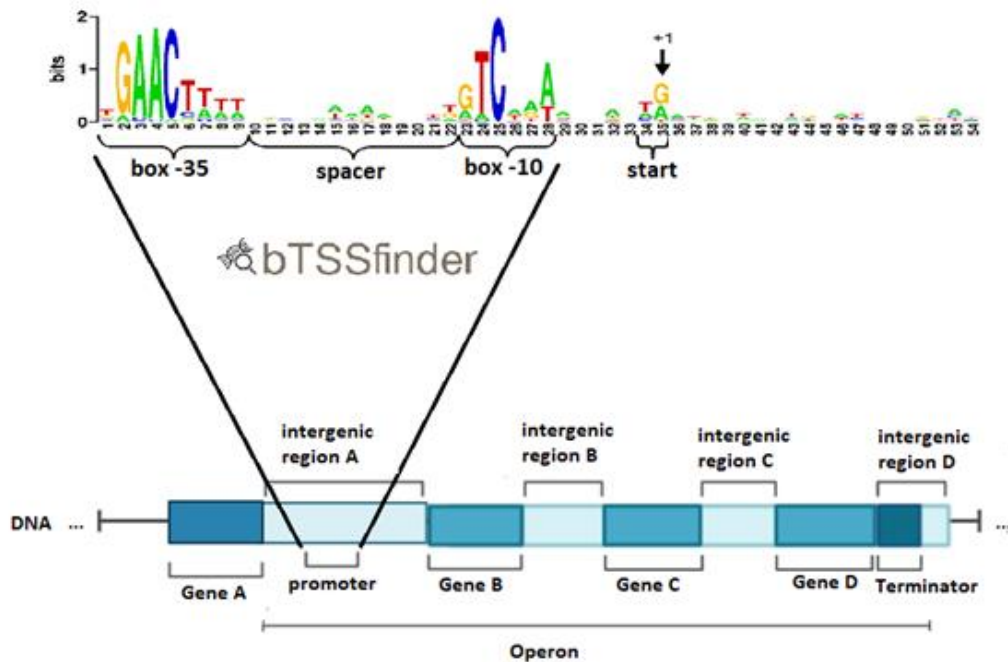


Figure 41 - Finding promoters inside intergenic sequences (Fig. S7 in the supplementary data).

In order to get the promoters of all the genomes (complete genome and incomplete), we put the fasta files of the genomes as an input of the program bTSSfinder, as in the command line below:

```
bTSSfinder -i $filename -o output/$file_name -h 2 -r 0 -t c
```

Options:

-h: 1 to search on the sense strand OR 2 for both strands

The output result contains different information as described in the following table.

Tableau 8 - Description of bTSSfinder result (Table S3 in the supplementary data).

Line	Description
1	The DNA segment identifier usually the fasta ID
2	The start position
3	The end position
4	Score
5	Strand '+' or '-'
6	The number of main promoter blocks (in our case =2, -35 and -10 boxes)
7	Listing of elements of Transcriptional factor binding sites for each predicted promoter

13.3.8 Terminators

A terminator is a sequence element that marks the end of the transcription, typically of a gene or of an operon. There are two types of terminators: rho-independent terminators (RITs) and rho-dependent terminators (RDTs).

13.3.8.1 Rho-independent terminators (RITs)

The RITs have a hairpin structure that destabilizes the RNA polymerase when followed by a poly U sequence of about six nucleotides. This allows easier dissociation of the DNA and RNA hybrid and release of the RNA polymerase.

In this case, the goal is to predict Rho independent terminators. To do so we used a probabilistic approach called “RNIE”. This method is based on the function `cmsearch` of the package `Infernal` that takes as an input the covariance models of Rho independent terminators.

```
rnies.pl <options> -f genome.fasta
```

Options:

-f: Annotate Rho independent terminators on sequences in genome.fasta

RNIE output is given in the example below :

```
# command: cmsearch -T 16 -g --fil-no-qdb --fil-T-hmm 2 --cyk --beta 0.05 --tabfile GCF_002903215.
s/genome.cm input/GCF_002903215.1_ASM290321v1_genomic.fna
# date: Thu Jul 23 14:36:10 2020
# num seqs: 3
# dbsize (Mb): 11.269684
#
# Pre-search info for CM 1: seed-a-14
#
#
#           cutoffs           predictions
# -----
# rnd  mod  alg  cfg  beta  E value  bit sc  surv  run time
# ---  ---  ---  ---  ---  ---
# 1  hmm  fwd  glc  -    21700.717  2.00  0.0957  00:00:40.80
# 2   cm  cyk  glc  0.05  4.171  16.00  1.5e-05  00:00:25.92
# ---  ---  ---  ---  ---  ---
# all  -   -   -   -    -      -    -      -    00:01:06.72
#
#
# CM: seed-a-14
#
#           target coord  query coord
# -----
# model name  target name  start  stop  start  stop  bit sc  E-value  GC%
# -----
# seed-a-14  NZ_CP026235.1  66069  66111  1  44  38.37  8.15e-04  40
# seed-a-14  NZ_CP026235.1  2858395  2858432  1  44  37.04  1.35e-03  34
# seed-a-14  NZ_CP026235.1  1297466  1297507  1  44  36.62  1.59e-03  38
# seed-a-14  NZ_CP026235.1  5156990  5157032  1  44  35.69  2.27e-03  44
# seed-a-14  NZ_CP026235.1  3060157  3060200  1  44  35.25  2.68e-03  52
# seed-a-14  NZ_CP026235.1  1922852  1922895  1  44  34.80  3.18e-03  43
# seed-a-14  NZ_CP026235.1  2027768  2027810  1  44  34.44  3.65e-03  37
# seed-a-14  NZ_CP026235.1  2706091  2706135  1  44  34.01  4.31e-03  47
# seed-a-14  NZ_CP026235.1  5023996  5024037  1  44  33.40  5.44e-03  43
# seed-a-14  NZ_CP026235.1  2237890  2237930  1  44  33.36  5.52e-03  34
# seed-a-14  NZ_CP026235.1  5015209  5015246  1  44  33.09  6.11e-03  42
# seed-a-14  NZ_CP026235.1  659882  659924  1  44  32.91  6.56e-03  47
# seed-a-14  NZ_CP026235.1  1803163  1803207  1  44  32.90  6.58e-03  40
# seed-a-14  NZ_CP026235.1  4822416  4822453  1  44  32.84  6.72e-03  24
# seed-a-14  NZ_CP026235.1  3005517  3005560  1  44  32.66  7.20e-03  43
```

Figure 42 - RNIE output (Fig. S8 in the supplementary data).

13.3.8.2 Rho-dependent terminators

In Rho-dependent terminators, Rho recognizes and binds to a region of the transcript (mRNA), then translocates from 5' to 3' and joins the arrested (or destabilized) RNA polymerase at the pause site and thus dissociates the elongation complex of the mRNA and RNA polymerase (RNAP).

We know numerous RUT sites and RNA polymerase pause sites from data in the literature. RhoTermPredict jointly uses these two data to predict the positions of these terminators in the genome by carrying out two steps:

- The algorithm scans windows of 78 nt on the input sequence to look for C/G ratio and “C” patterns.
- The second step consists in identifying the putative pause sites for RNAPs downstream of the 3'end of the RUT site from a distance of 150 nt.

To get all the Rho dependent terminators of all the genomes (complete genome and incomplete), we put the fasta files of the genomes as an input of the program RhoTermPredict, as in the command line below:

RhoTermPredict_algorithm.py filename.

Input: Genome sequences file

Output: a xlsx file containing Rho-dependent terminators coordinates and a txt file containing information about them

The output is given in a table format with positions of the terminators.

Tableau 9–Example of an output of RhoTermPredict (Table S4 in the supplementary data).

Region	Start RUT	End RUT	Strand	E-value
T1	81	159	+	e^{-04}
T2	338	416	+	e^{-04}
T3	732	810	-	e^{-04}
T4	1039	1117	+	e^{-04}

13.3.9 Conserved proteins

Gene annotation found in GenBank includes a lot of unannotated sequences. To illustrate more this problem here an example is presented:

Example: Several genes assigned to hypothetical protein according to NCBI annotations in RiboGap v1 (see Table s5)

Tableau 10 - Example of genes that code for “hypothetical proteins” (Table S5 in the supplementary data).

gene	product
null	hypothetical protein
DsrE	hypothetical protein
CobT	hypothetical protein
EhaM	hypothetical protein
EhaL	hypothetical protein
EhaH	hypothetical protein
EhaG	hypothetical protein
EhaF	hypothetical protein
EhaE	hypothetical protein
EhaD	hypothetical protein
EhaC	hypothetical protein
EhaB	hypothetical protein
EhaA	hypothetical protein
PflX	hypothetical protein
PmbA	hypothetical protein
CdsA	hypothetical protein

InterProScan composed of several databases is a good tool that allows making a more precise annotation (by mentioning the domain and the family of the protein).

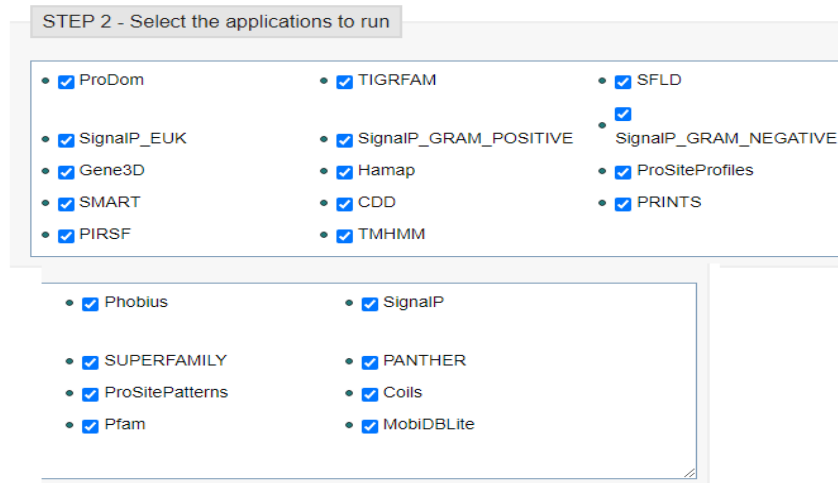


Figure 43 - different databases that compose InterProScan (Fig. S9 in the supplementary data).

InterProScan could be run simply from the following command line.

```
./interproscan.sh -i sequence.fasta -f GFF3
```

-i: The input option

sequence: It represents the gene sequence that is functionally annotated

-f: The format of output

GFF3 A flat tab-delimited file format

The output allows predicting proteins which matches with DNA sequences. Each column of the result is described in the Table S6.

Tableau 11 - Description of InterProScan output (Table S6 in the supplementary data).

Column	Name of the column	Description
1	SeqID	The ID used to coordinate the system for the current feature.
2	Source	It describes the algorithm or the database that generated this feature.
3	Type	The type of the feature, previously called "method"
4	Start	Start position of the feature
5	End	End position of the feature
6	Score	A floating point number that shows the score of the feature
7	Strand	It can be positive or minus strand
8	Phase	It indicates the next codon begins relative to the 5' end of the current CDS feature
9	Attributes	It corresponds to a list of feature attributes in the format tag=value. Where the tags can be: ID, name, alias, target, gap

13.4 Data in RiboGap v2

13.4.1 RiboGap v2's diagram

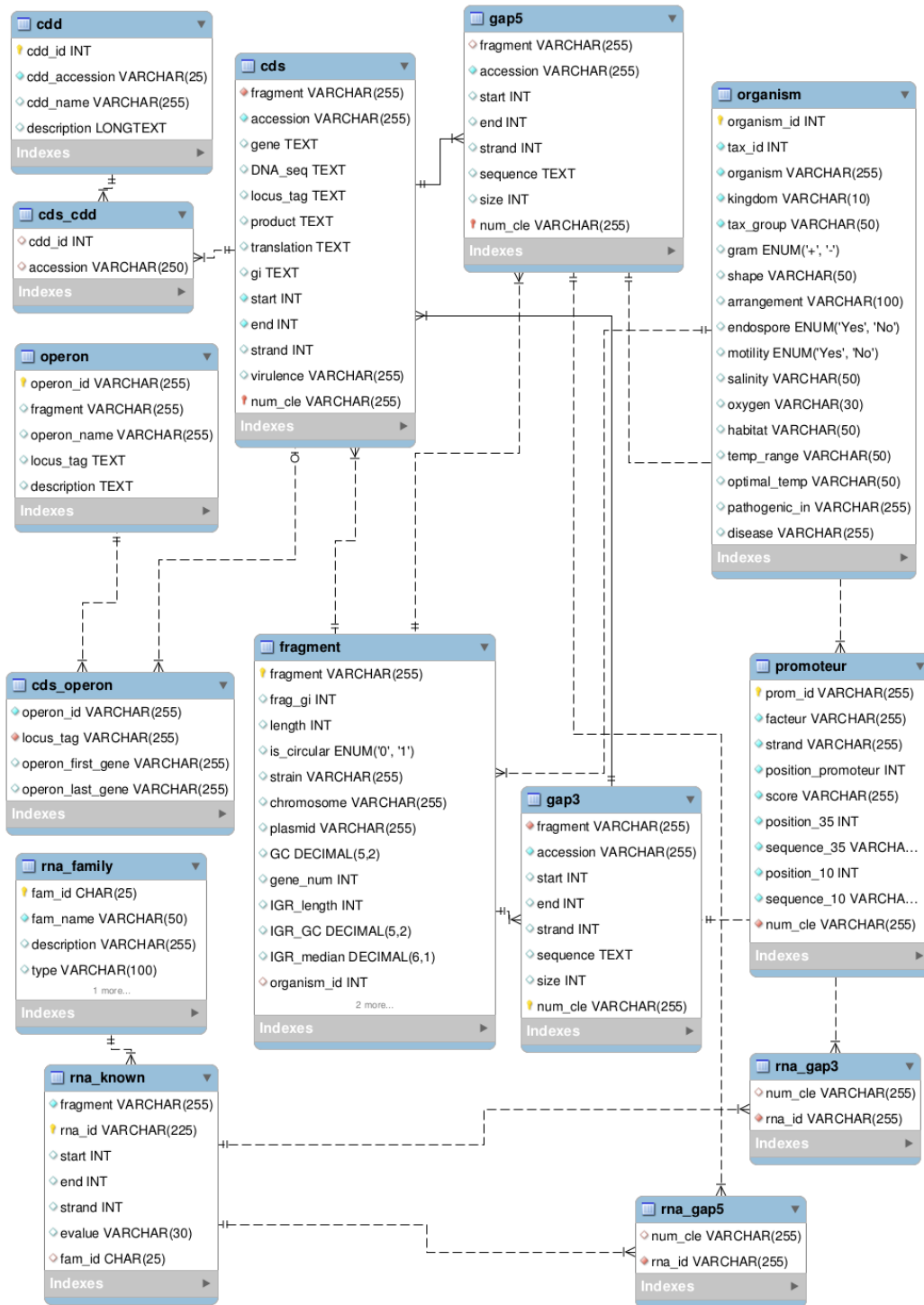


Figure 44 - Detailed diagram of RiboGap v2 (Fig. S10 in the supplementary data).

13.4.2 Compilation of promoter predictions

In order to calculate number of promoters, operons and genes for each genome's taxonomy, we wrote a MySQL query to get them easily in one command line as shown in the Table S7.

Tableau 12 - Different MySQL queries that calculate the number of promoters, genes and operons according to taxonomy (Table S7 in the supplementary data).

Objective of the query	MySQL query
Calculate the number of promoters according to the taxonomy	select fragment.taxonomy, count(promoteur.prom_id) from fragment inner join cds on cds.fragment=fragment.fragment inner join promoteur on cds.num_cle = promoteur.num_cle group by fragment.taxonomy INTO OUTFILE '/var/lib/mysql-files/promoteur_taxonomie.csv';
Calculate the number of genes according to the taxonomy	select fragment.taxonomy, count(cds.num_cle) from fragment inner join cds on cds.fragment=fragment.fragment group by fragment.taxonomy INTO OUTFILE '/var/lib/mysql-files/gene_taxonomie.csv';
Calculate the number of operons according to the taxonomy	select fragment.taxonomy, count(operon.operon_id) from fragment inner join operon on fragment.fragment = operon.fragment group by fragment.taxonomy INTO OUTFILE '/var/lib/mysql-files/operon_taxonomie.csv';

However, this can be done also under the web interface for example if we want to get all the promoters for the taxonomy “**Proteobacteria**” as displayed below:

fragment	Chromosome information	promoter	Predicted putitive promoters																												
<input type="checkbox"/> DNA fragment <input type="checkbox"/> frag_gi <input type="checkbox"/> length <input type="checkbox"/> is_circular <input type="checkbox"/> strain <input type="checkbox"/> chromosome <input type="checkbox"/> plasmid <input type="checkbox"/> GC <input type="checkbox"/> gene_num <input type="checkbox"/> IGR_length <input type="checkbox"/> IGR_GC <input type="checkbox"/> IGR_median <input type="checkbox"/> organism_id <input type="checkbox"/> taxonomy <input type="checkbox"/> description	Refseq accession number like NC_000913 gi of Refseq accession like 158421624 length of chromosome true/false strain information like "Newman" chromosome if exist in database plasmid if exist in database GC % like 52.17 number of genes in chromosome like: 4934 total length of intergenic sequences GC % in intergenic sequences median size of intergenic sequences project id of organism according to NCBI bacteria; elusimicrobia; environmental samples Staphylococcus aureus subsp. aureus str. Newman	<input type="checkbox"/> factor <input type="checkbox"/> strand <input type="checkbox"/> position_promoter <input type="checkbox"/> score <input type="checkbox"/> position_35 <input type="checkbox"/> sequence_35 <input type="checkbox"/> position_10 <input type="checkbox"/> sequence_10	transcription factor strand of promoter position of the promoter score varies from the worst = 0.07 to the best = 2 position of the box 35 motif of the box 35 position of the box 10 motif of the box 10																												
Condition: <table border="1"> <tr> <td>taxonomy</td> <td>find some pattern</td> <td>Proteobacteria</td> <td>-</td> </tr> <tr> <td>-</td> <td>-</td> <td>-</td> <td>-</td> </tr> <tr> <td>-</td> <td>-</td> <td>-</td> <td>-</td> </tr> <tr> <td>-</td> <td>-</td> <td>-</td> <td>-</td> </tr> <tr> <td>-</td> <td>-</td> <td>-</td> <td>-</td> </tr> <tr> <td>-</td> <td>-</td> <td>-</td> <td>-</td> </tr> <tr> <td>-</td> <td>-</td> <td>-</td> <td>-</td> </tr> </table>				taxonomy	find some pattern	Proteobacteria	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
taxonomy	find some pattern	Proteobacteria	-																												
-	-	-	-																												
-	-	-	-																												
-	-	-	-																												
-	-	-	-																												
-	-	-	-																												
-	-	-	-																												

Figure 45 - Selection of promoters for the genome taxonomy "Proteobacteria" (Fig. S11 in the supplementary data).

If we choose this method, the number of promoters can be gotten from the total number of lines of the downloadable CSV file result.

Tableau 13 - Calculation of promoters per genes and operons according to taxonomy (Table S8 in the supplementary data).

	Taxonomy	Nb promo	Nb gene	IGR Size	Nb operon	promo/gene	promo/kb of IGR	promo/operon
Arc	TACK group	187.041	145.525	163	109.112	1.29	10	1.71
	Furvarchaeota	860.547	450.946	210	160.675	1.91	10	5.36
	DPANN group	1.088	971	149	326	1.12	8	3.34
Bacteria	Proteobacteria	24.768.079	17.880.508	176	6.365.58	1.39	9	3.89
	Terrabacteria	15.905.872	10.902.730	182	3.607.89	1.46	9	4.41
	FCB group	1.636.732	1.099.861	167	315.319	1.49	10	5.19
	PVC group	487.096	261.922	208	88.646	1.86	10	5.49
	Spirochaetes	150.190	113.680	169	39.157	1.32	10	3.86
	Acidobacteria	55.140	39.767	180	16.318	1.39	9	3.38
	Calditrichaeota	7.820	3.818	214	730	2.05	11	10.71
	Svnergistetes	11.949	10.768	143	2.976	1.11	10	4.02
	Fusobacteria	105.159	94.534	144	17.803	1.11	9	5.91
	Aquificae	10.663	27.499	94	8.046	0.39	7	1.33
	Thermotogae	54.335	70.177	126	16.339	0.77	9	3.33
	Deferribacteres	10.803	9.957	137	3.034	1.08	11	3.56
	Nitrospirae	32.256	26.655	167	7.098	1.21	9	4.54
	Dictyodromi	2.030	3.702	104	1.051	0.55	7	1.93
	Chrysiogenetes	3.335	2.594	180	955	1.29	10	3.49
	Elusimicrobia	8.769	5.106	195	948	1.72	10	9.25
	Caldiserica	1.308	1.512	136	466	0.87	10	2.81
	Thermodesulfobacteri	7.585	7.374	129	1.899	1.03	10	3.99
	Coprothermobacter	989	1.415	121	410	0.70	9	2.41
	Unclassified Bacteria	309,088	245,287	178	85,569	1.26	9	3.61

Arc: phyla corresponding to Archaea.

Bacteria: phyla corresponding to bacteria.

Nb promo: total number of predicted promoters in IGRs.

Nb gene: total number of genes in RiboGap as annotated in NCBI.

IGR size: average IGR size

Nb operon: total number of operons in RiboGap. It should be noted that some genes (and thus some IGRs) are found in more than one operon.

Promo/gene: average number of promoters predicted per gene.

Promo/Mb: average number of promoters predicted per total size of IGR

Promo/operon: average number of promoters predicted per operon.

These data are gotten from RiboGap v2 for complete genomes

Note that due to some parameters of bTSSfinder and the fact that we used only IGRs for promoter searches, only IGRs larger than 250 bp could have promoter predictions. Moreover, the parameters used for this large-scale prediction of -10 boxes and -35 boxes of promoters, some limitations were noted. First, promoters were only predicted in IGRs of a size > 150 bases. While IGRs that harbor promoters are on average larger than IGRs intervening between CDSs of polycistronic mRNAs, this size limitation prevented predictions of numerous promoters that exist in smaller IGRs. More importantly, empirical survey of predicted promoter positions within IGRs indicate that this limitation appears to be due to some window size used by bTSSfinder, leading to a “dead space” where no promoters are predicted close to the end of the IGR (and thus close to the CDS’ start codon). Second, most of the promoter predictions appear to be false positives. This can in part be inferred from the presence of similar numbers of promoters predicted in Archaea and Bacteria (Supplementary section 3.2, Table S8). Because Archaea have transcription factor B (TFB) instead of sigma factors and a TATA box analogous to that of eukaryotes (Smale & Kadonaga, 2003), promoter predictions in these genomes is expected to be close to background. However, we observe as many predictions in archaeal genomes as in bacteria (on average ~10 promoter predictions per kilobase of IGR, Table S8), indicating that most predicted promoters are false positives. MySQL queries used for this are explained in Table S7. bTSSfinder has been tuned and tested primarily on *Escherichia coli* and three cyanobacteria (Shahmuradov et al., 2017), which are respectively part of proteobacteria (Garrity et al., 2015) and terrabacteria (Superson et al., 2018). While in principle these distantly related models used by bTSSfinder should increase the broadness of promoter prediction, it may also decrease specificity, partly explaining high prediction rates. Nevertheless, with close to 100 million promoter predictions (and almost 1 billion for incomplete genomes), RiboGap v2 and v2.1 provide a vast collection of putative sigma binding sites which can be useful when used appropriately, as a starting point for a more careful analysis. For instance, when combined with comparative genomics and/or experimental data (like RNA-seq), these predictions can be useful to infer the putative corresponding promoter (as exemplified in Fig. S12).

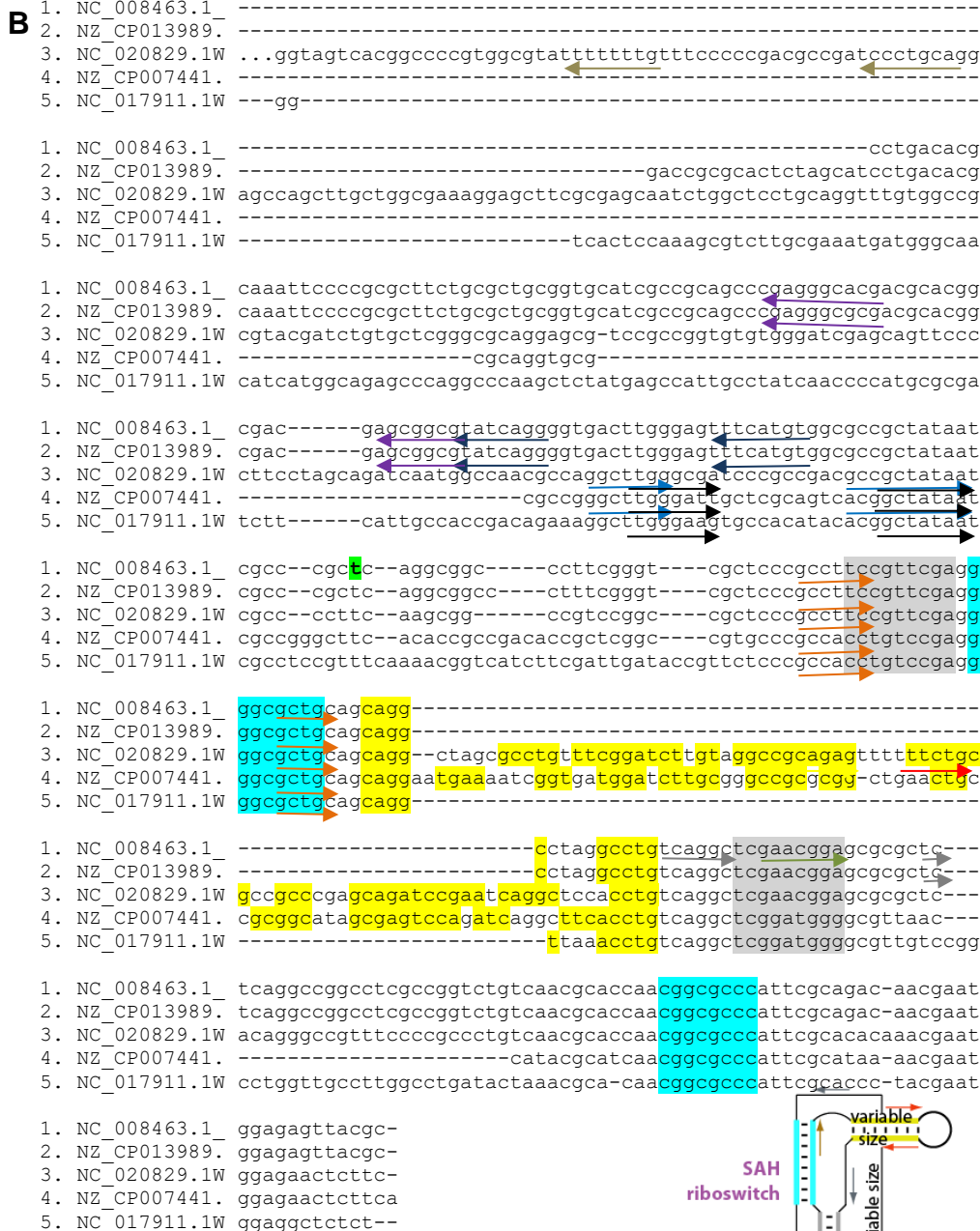
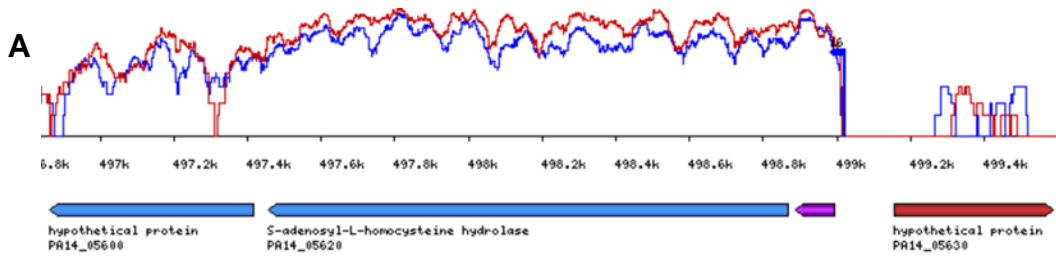


Figure 46 - Evaluation of promoter conservation can help validate predicted -35 and -10 boxes.A (Fig. S12 in the supplementary data)

RNA-seq in *Pseudomonas aeruginosa* UCBPP-PA14 (Wurtzel *et al.*, 2012) (available at http://www.weizmann.ac.il/molgen/Sorek/pseudomonas_browser/) highlights a clear transcription start site downstream (but upstream relative to the gene orientation) of a gene encoding a S-adenosyl-L-homocysteine hydrolase. **B)** Multiple alignment (performed with MAFFT (Katoh & Toh, 2008)) of the IGR region upstream (gap5) of the latter gene from 5 strains of *Pseudomonas* (selected from 1324 such gap5, after filtration for redundancy with CD-hit (Huang *et al.*, 2010), the resulting 55 sequence alignment was reduced to 5 for the sake of this figure). The sequences are presented in the same polarity as the gene encoding the S-adenosyl-L-homocysteine hydrolase (and thus inverse to that pictured in (A)). The “...” represent a truncation of this sequence which was much longer than the others. Arrows beneath each sequence represent promoter boxes that we predicted with bTSSfinder; similar arrow colors represent box pairs. The +1 transcription start site is represented by the “T” highlighted in green in NC_008463.1 (*P. aeruginosa* UCBPP-PA14) in the fifth block of the alignment according to data presented in (A). The other highlighted sequences correspond to the SAH riboswitch found in this IGR, where similar colors highlight sequences from the same stems (as illustrated in (C)). **C)** Schematic of the IGR (in the same polarity as the alignment) with the annotated features of the alignment, but not to scale, arrow and riboswitch sequence colors are the same as those in the alignment in (B). The big black arrows are those that fit best with the experimentally determined transcription start site.

Figure S12 illustrates that each IGR likely has more “false negative” predictions of promoters than accurate predictions. Moreover, the IGR of *P. aeruginosa* UCBPP-PA14 also appears to lack the accurate promoter prediction, even if the other aligned sequences have the correct prediction. This could be due to the presence of a predicted promoter on the other strand (presumably for the transcription of the “hypothetical gene”). Indeed, our own empirical observation of predictions from bTSSfinder results suggest that this software does not predict overlapping promoters on opposite polarities. Moreover, even when looking at multiple sequence alignments to look for conserved -35 and -10 boxes, one should be careful as other conserved features, such as the SAH riboswitch from this example, can erroneously lead to think that the promoter prediction is good due to its conservation, while the conservation is actually due to other functional elements, as is the case for the boxes above the brown arrows at the beginning of the riboswitch. Nevertheless, this example still indicates that our promoter predictions can be useful when used together with comparative genomics and agree with experimental data in this case.

13.5 The use of the database

13.5.1 Find a promoter, a terminator and a riboswitch in the same intergenic region

To achieve this goal, we can easily use RiboGap interface in three steps: select IGRs that have both promoters and a terminator (see Figure S13), then select IGRs that have riboswitches (see Figure S14) and finally get the common intergenic sequences for the two results.

promoter		Predicted putative promoters	
<input checked="" type="checkbox"/> factor	transcription factor	gap5	Sequence information for 5-prime-UTR
<input checked="" type="checkbox"/> strand	strand of promoter	<input checked="" type="checkbox"/> accession	protein accession like NP_038276.1
<input checked="" type="checkbox"/> position_promoter	position of the promoter	<input checked="" type="checkbox"/> start	start position of 5 prime-UTR
<input checked="" type="checkbox"/> score	score varies from the worst = 0.07 to the best = 2	<input checked="" type="checkbox"/> end	end position of 5 prime-UTR
<input checked="" type="checkbox"/> position_35	position of the box 35	<input checked="" type="checkbox"/> strand	strand direction of the corresponding gene
<input checked="" type="checkbox"/> sequence_35	motif of the box 35	<input checked="" type="checkbox"/> sequence	sequence of 5 prime-UTR in same strand as the gene
<input checked="" type="checkbox"/> position_10	position of the box 10	<input checked="" type="checkbox"/> size	size of 5 prime-UTR
<input checked="" type="checkbox"/> sequence_10	motif of the box 10		
rna_family		Family of RNA according to Rfam(+ transcription terminators + pseudo tRNA)	
<input checked="" type="checkbox"/> fam_id	Rfam accession: RF00001		
<input checked="" type="checkbox"/> fam_name	5S_rRNA		
<input checked="" type="checkbox"/> description	5S ribosomal RNA		
<input checked="" type="checkbox"/> type	gene; rRNA		
<input checked="" type="checkbox"/> note	some description		

Condition:

description	find some pattern	Terminator	-
-	-		-
-	-		-
-	-		-
-	-		-
-	-		-
-	-		-

Figure 47 - Step1: select IGRs with terminators and promoters (Fig. S13 in the supplementary data)

gap5		Sequence information for 5-prime-UTR	
<input checked="" type="checkbox"/> accession	protein accession like NP_038276.1		
<input checked="" type="checkbox"/> start	start position of 5 prime-UTR		
<input checked="" type="checkbox"/> end	end position of 5 prime-UTR		
<input checked="" type="checkbox"/> strand	strand direction of the corresponding gene		
<input checked="" type="checkbox"/> sequence	sequence of 5 prime-UTR in same strand as the gene		
<input checked="" type="checkbox"/> size	size of 5 prime-UTR		
rna_family		Family of RNA according to Rfam(+ transcription terminators + pseudo tRNA)	
<input checked="" type="checkbox"/> fam_id	Rfam accession: RF00001		
<input checked="" type="checkbox"/> fam_name	5S_rRNA		
<input checked="" type="checkbox"/> description	5S ribosomal RNA		
<input checked="" type="checkbox"/> type	gene; rRNA		
<input checked="" type="checkbox"/> note	some description		

Condition:

description	find some pattern	Riboswitch	-
-	-		-
-	-		-
-	-		-
-	-		-
-	-		-
-	-		-

Figure 48 - Step2: select IGRs with Riboswitches (Fig. S14 in the supplementary data)

These two steps can be executed in one command by performing the MySQL query below. Alternatively, a similar query can be used, but without constraints for “RNA – description”, which will provide all gaps that have RNAs, including those that have both a riboswitch and a terminator.

MySQL query:

```
SELECT * FROM (select fragment.fragment, fragment.description, gap5.num_cle,
gap5.sequence, gap5.start, gap5.end, gap5.sequence, promoteur.position_35,
promoteur.sequence_35, promoteur.position_10, promoteur.sequence_10 from fragment inner
join gap5 on gap5.fragment = fragment.fragment inner join promoteur on promoteur.num_cle =
gap5.num_cle)A , (select gap5.num_cle, rna_known.start as start_rna, rna_known.end as end_rna,
rna_family.fam_id, rna_family.fam_name, rna_family.description from gap5 inner join rna_gap5
on rna_gap5.num_cle = gap5.num_cle inner join rna_known on rna_gap5.rna_id =
rna_known.rna_id inner join rna_family on rna_family.fam_id = rna_known.fam_id where
rna_family.description like '%riboswitch%') B, (select gap5.num_cle, rna_known.start as
start_rna, rna_known.end as end_rna, rna_family.fam_id, rna_family.fam_name,
rna_family.description from gap5 inner join rna_gap5 on rna_gap5.num_cle = gap5.num_cle inner
join rna_known on rna_gap5.rna_id = rna_known.rna_id inner join rna_family on
rna_family.fam_id = rna_known.fam_id where rna_known.fam_id like '%DP%') C WHERE
A.num_cle = B.num_cle and B.num_cle = C.num_cle and A.position_35 < A.position_10 and
A.position_10 < B.start_rna and B.start_rna < B.end_rna and B.end_rna < C.start_rna and
C.start_rna < C.end_rna INTO OUTFILE '/var/lib/mysql-
files/promoteur_riboswitch_terminateur.csv';
```

The result of this query shows three different types of genetic elements (promoter, riboswitch and terminator) within the same IGRs. Complete results can be found in:

Tableau 14 - Analysis of the existence of a riboswitch, promoter and a terminator in an intergenic region (Table S10 in the supplementary data)

Intergenic sequence accession	Intergenic sequence	Color code
<p>Example 1 : Sequence16737528</p>	<p>AAGTGT TAGGAGCTAGAAACTAATTAAAAATAGTTAA TTAACGATAATAAAAAATAAAAAATTTAAATAAATTTA TTGACAAATACTATTAATGAATATATCATATATAACAT AGCAAC AAATGATAATAGAATTAAATTAATAAC GTC TTATCAAGAGTGGTGGAGGGACTGGCCCTTTGAAACC CGGCAACCAGTATATTTTATATACATTGGTGCTAAAT CCTGCAACTAATATAGTTGATAGATGAGTATAAATAG CTTTCT AAGCCTGAGTTTATACTCGGGCTTTTATTTTG TTGTAATACATAAGAGAGGATAAACATAAAAACTGAGC AGAATATATGGGAGGAAATTTT</p>	<p>Riboswitch</p> <p>Promoter</p> <p>box -35</p> <p>box -10</p>
<p>Example 2 : Sequence11638308</p>	<p>ATTACCATCCTTTTCATTTAATGAAATATTTCTCGTAA TCCCCAACAATAGTATAACACAATTATATAAATAATGT ATACTATTTATTTCTTTTTGTTTGTGTGAATACATTAT TTCCTTATTAATATTTTCATTTTAGGCGCATTGAGAAAA ACTTTCACGAAAAT GTGAATATTAATTGACAAATGAAT ATCATTTCTTTAGTATAGGTTGGGACGGATACT CTCT TATCCCGAGCTGGCGGAGGGACAGGCCCGATGAAGCC CAGCAACCTCACTTGTAGTGGTAAATACAGGTGAATA GGTGCTAAAACCTGTGCGAGGCTACAGGTCTCGAACG ATAAGAGCGAAGGGCAAAAAGCAGTATGCAAGTAGCA AATTAAACCTTCTCTAT ATAAAGTAGGAAAGGTTTT TCTGTATGCTTGTGTGGGAGAATAAATGTATGTGCGCAA TCTGTGGCAAATTAAGGATGAGTTCCGTACAATATATA CAATTACTGTAGGGAGGTTTACCAC</p>	<p>RTI</p> <p>RDT</p>

13.5.2 Search the expression platforms of all riboswitches

In this part, we present three cases, IGRs that contain Riboswitches + Terminators, Riboswitches only (no ncRNA nor Terminator) or Riboswitches + ncRNAs. Using the web interface, each may be done in two steps: select all the IGRs that have Riboswitches and then select regions which contain either terminators or ncRNAs or neither (according to the case). Alternatively, a single query looking for all IGRs with RNAs (riboswitches or RITs or RDTs) would retrieve many relevant IGRs, which would then need to be sorted for those containing riboswitches and terminators.

The formulation of the request of the third case “Riboswitch + ncRNA” from the interface is given in Figures S15 and S16 as an example.

gap5	Sequence information for 5-prime-UTR
<input checked="" type="checkbox"/> accession	protein accession like NP_038276.1
<input checked="" type="checkbox"/> start	start position of 5 prime-UTR
<input checked="" type="checkbox"/> end	end position of 5 prime-UTR
<input checked="" type="checkbox"/> strand	strand direction of the corresponding gene
<input checked="" type="checkbox"/> sequence	sequence of 5 prime-UTR in same strand as the gene
<input checked="" type="checkbox"/> size	size of 5 prime-UTR

rna_family	Family of RNA according to Rfam(+ transcription terminators + pseudo tRNA)
<input checked="" type="checkbox"/> fam_id	Rfam accession: RF00001
<input checked="" type="checkbox"/> fam_name	5S_rRNA
<input checked="" type="checkbox"/> description	5S ribosomal RNA
<input checked="" type="checkbox"/> type	gene; rRNA
<input checked="" type="checkbox"/> note	some description

Condition:

description	▼	find some pattern	▼	Riboswitch	-	▼
-	▼	-	▼		-	▼
-	▼	-	▼		-	▼
-	▼	-	▼		-	▼
-	▼	-	▼		-	▼
-	▼	-	▼		-	▼
-	▼	-	▼		-	▼

Figure 49 - Step1: select IGRs with Riboswitches (Fig. S15 in the supplementary data)

gap5	Sequence information for 5-prime-UTR
<input checked="" type="checkbox"/> accession <input checked="" type="checkbox"/> start <input checked="" type="checkbox"/> end <input checked="" type="checkbox"/> strand <input checked="" type="checkbox"/> sequence <input checked="" type="checkbox"/> size	protein accession like NP_038276.1 start position of 5 prime-UTR end position of 5 prime-UTR strand direction of the corresponding gene sequence of 5 prime-UTR in same strand as the gene size of 5 prime-UTR

rna_family	Family of RNA according to Rfam(+ transcription terminators + pseudo tRNA)
<input checked="" type="checkbox"/> fam_id <input checked="" type="checkbox"/> fam_name <input checked="" type="checkbox"/> description <input checked="" type="checkbox"/> type <input checked="" type="checkbox"/> note	Rfam accession: RF00001 5S_rRNA 5S ribosomal RNA gene; rRNA some description

Condition:

description	▼	find some pattern	▼	Terminator	-	▼
-	▼	-	▼		-	▼
-	▼	-	▼		-	▼
-	▼	-	▼		-	▼
-	▼	-	▼		-	▼
-	▼	-	▼		-	▼
-	▼	-	▼		-	▼

Figure 50 - Step2: select all the ncRNA (no terminators) (Fig. S16 in the supplementary data)

Once, we get two different lists, we just need to cross these lists in order to get the common IGR to find out the intergenic sequences that have riboswitches and ncRNAs, but no terminators, to get occurrences of the third case mentioned above.

For those familiar with MySQL, this can be done with an online command as shown in Table S10 carried out for the three cases.

Tableau 15 - MySQL queries and programs that calculate different distances (Table S11 in the supplementary data)

Objective of the query	MySQL query	Executed perl program on RiboGap's result
List positions of Rho dependent terminators (RDTs) that are downstream of riboswitches	<pre>SELECT * FROM (select gap5.num_cle, gap5.start, gap5.end, gap5.size, rna_known.start as start_rna, rna_known.end as end_rna, rna_family.fam_id, rna_family.fam_name, rna_family.description from gap5 inner join rna_gap5 on rna_gap5.num_cle= gap5.num_cle inner join rna_known on rna_gap5.rna_id = rna_known.rna_id inner join rna_family on rna_family.fam_id = rna_known.fam_id where rna_family.fam_id like '%RhoDP%')A , (select gap5.num_cle, rna_known.start as start_rna, rna_known.end as end_rna, rna_family.fam_id, rna_family.fam_name, rna_family.description from gap5 inner join rna_gap5 on rna_gap5.num_cle = gap5.num_cle inner join rna_known on rna_gap5.rna_id = rna_known.rna_id inner join rna_family on rna_family.fam_id = rna_known.fam_id where rna_family.description like '%riboswitch%') B WHERE A.num_cle = B.num_cle and B.end_rna < A.start_rna INTO OUTFILE '/var/lib/mysql-files/riboswitch_terminateur_RhoDP.csv';</pre>	<p><u>Perl program 1.1:</u> Calculate the distance between the riboswitch and the terminator</p>
List positions of Rho independent terminators (RITs) that are downstream of riboswitches	<pre>SELECT * FROM (select gap5.num_cle, gap5.start, gap5.end, gap5.size, rna_known.start as start_rna, rna_known.end as end_rna, rna_family.fam_id, rna_family.fam_name, rna_family.description from gap5 inner join rna_gap5 on rna_gap5.num_cle= gap5.num_cle inner join rna_known on rna_gap5.rna_id = rna_known.rna_id inner join rna_family on rna_family.fam_id = rna_known.fam_id where rna_family.fam_id like '%RhoINDP%')A , (select gap5.num_cle, rna_known.start as start_rna, rna_known.end as end_rna, rna_family.fam_id, rna_family.fam_name, rna_family.description from gap5 inner join rna_gap5 on rna_gap5.num_cle = gap5.num_cle inner join rna_known on rna_gap5.rna_id = rna_known.rna_id inner join rna_family on rna_family.fam_id = rna_known.fam_id where rna_family.description like '%riboswitch%') B WHERE A.num_cle = B.num_cle and B.end_rna < A.start_rna INTO OUTFILE '/var/lib/mysql-files/riboswitch_terminateur_RhoINDP.csv';</pre>	<p><u>Perl program 1.2:</u> Calculate the distance between the terminator and the start codon</p> <p><u>Perl program 1.3:</u> Calculate the distance between the riboswitch and the start codon</p>
List all positions of riboswitches that do not have transcription terminators	<pre>SELECT * FROM (select gap5.num_cle from gap5 inner join rna_gap5 on rna_gap5.num_cle = gap5.num_cle inner join rna_known on rna_gap5.rna_id = rna_known.rna_id inner join rna_family on rna_family.fam_id = rna_known.fam_id group by gap5.num_cle having max(rna_family.description not like '%riboswitch%') = 0) A , (select gap5.num_cle, gap5.start, gap5.end, rna_known.start as start_rna, rna_known.end as end_rna, rna_family.fam_id, rna_family.fam_name, rna_family.description from gap5 inner join rna_gap5 on rna_gap5.num_cle = gap5.num_cle inner join rna_known on rna_gap5.rna_id = rna_known.rna_id inner join rna_family on rna_family.fam_id = rna_known.fam_id) B WHERE A.num_cle = B.num_cle into OUTFILE '/var/lib/mysql-files/riboswitch_codon_start.csv';</pre>	<p><u>Perl program 2:</u> Calculate the distance between the start codon and the riboswitch</p>
List positions of riboswitches and	<pre>SELECT * FROM (select gap5.num_cle, gap5.sequence, gap5.start, gap5.end, gap5.size, rna_known.start as start_rna, rna_known.end as end_rna, rna_family.fam_id, rna_family.fam_name, rna_family.description from gap5 inner join rna_gap5 on rna_gap5.num_cle= gap5.num_cle inner join rna_known on rna_gap5.rna_id = rna_known.rna_id inner join rna_family on rna_family.fam_id = rna_known.fam_id where rna_family.description not like</pre>	<p><u>Perl program 3:</u> Calculate the distance between the riboswitch and the ncRNA</p>

ncRNAs found in the same IGRs	'%riboswitch%' and rna_family.fam_id not like '%DP%')A , (select gap5.num_cle, gap5.sequence, gap5.start, gap5.end, rna_known.start as start_rna, rna_known.end as end_rna, rna_family.fam_id, rna_family.fam_name, rna_family.description from gap5 inner join rna_gap5 on rna_gap5.num_cle = gap5.num_cle inner join rna_known on rna_gap5.rna_id = rna_known.rna_id inner join rna_family on rna_family.fam_id = rna_known.fam_id where rna_family.description like '%riboswitch%') B WHERE A.num_cle = B.num_cle and B.end_rna < A.start_rna INTO OUTFILE '/var/lib/mysql-files/riboswitch_rna.csv';	
-------------------------------	--	--

Once we get the list of intergenic sequences for the three cases choosing one of the mentioned methods, three programs have been executed on the lists in order to get distances for the cases.

Case 1: Riboswitch + Terminator (RIT or RDT)

For this case, we calculated the distance between the riboswitch and terminator, then between the terminator and start codon and between the riboswitch and the start codon.

Case 2: Riboswitch (no terminators and no ncRNA)

We calculated the distance between the riboswitch and the start codon.

Case 3: Riboswitch + ncRNA

We calculated the distance between the riboswitch and the ncRNA.

The programs calculated the distances as described below:

Program 1.1:

```
#!/usr/bin/perl
use strict;
use warnings;

open(IN, "riboswitch_terminateur_RhoDP.csv");
open(OUT, ">Case_riboswitch+RDT_distance_riboswitch_terminateur_RhoDP_distance.c
```

```

sv");

my $ligne;

print OUT "IGR accession",";", "Distance",";", "Size","\n";

while ($ligne = <IN>){

chomp($ligne);

my @tmp=split("\t",$ligne);

my $num_cle=$tmp[0];

my $start_terminateur=$tmp[4];

my $end_riboswitch=$tmp[11];

my $distance=0;

my $size=$tmp[3];

if ($start_terminateur>$end_riboswitch){

    $distance=$start_terminateur-$end_riboswitch;

    print OUT $num_cle,";", $distance,";", $size,"\n";

}}

```

Program 1.2:

```

#!/usr/bin/perl

use strict;

use warnings;

open(IN, "riboswitch_terminateur_RhoDP.csv");

```

```

open(OUT, ">Case_riboswitch+RDT_distance_terminateur_RhoDP_startCodon_distance.c
sv");
my $ligne;
print OUT "IGR accession",";","Distance",";","Size","\n";
while ($ligne = <IN>){
chomp($ligne);
my @tmp=split("\t",$ligne);
my $num_cle=$tmp[0];
my $end_terminateur=$tmp[5];
my $start=$tmp[2];
my $distance=0;
my $size=$tmp[3];

if ($start>$end_terminateur){
    $distance=$start-$end_terminateur;
    print OUT $num_cle,";",$distance,";",$size,"\n";
}}

```

Program 1.3:

```

#!/usr/bin/perl
use strict;
use warnings;

```

```

open(IN, "riboswitch_terminateur_RhoINDP.csv");
open(OUT, ">Case_riboswitch+RIT_distance_riboswitch_startCodon_distance.csv");
my $ligne;
print OUT "IGR accession",";","Distance",";","Size","\n";
while ($ligne = <IN>){
chomp($ligne);
my @tmp=split("\t",$ligne);
my $num_cle=$tmp[0];
my $end_riboswitch=$tmp[11];
my $start=$tmp[2];
my $distance=0;
my $size = $tmp[3];
if ($start>$end_riboswitch)
{
    $distance=$start-$end_riboswitch;
    print OUT $num_cle,";","$distance,",";$size","\n";
}}

```

Program 2:

```

#!/usr/bin/perl

use strict;

use warnings;

```

```

open(IN, "riboswitch_codon_start.csv");
open(OUT, ">Case_riboswitch_distance_riboswitch_codon_start_distance");

my $ligne;

print OUT "IGR accession",";","Distance",";","Size","\n";
while ($ligne = <IN>)
{
chomp($ligne);
my @tmp=split("\t",$ligne);
my $num_cle=$tmp[0];
my $start_codon=$tmp[2];
my $end_riboswitch=$tmp[5];
my $distance=0;
my $size=$tmp[3];
if ($start_codon>$end_riboswitch){
    $distance=$start_codon-$end_riboswitch;
    print OUT $num_cle,";",$distance,";",$size,"\n";
}}

```

Program 3:

```
#!/usr/bin/perl
```

```

use strict;

use warnings;

open(IN, "riboswitch_rna.csv");

open(OUT, ">Case_riboswitch+RNAnc_distance_riboswitch_rna_moyenne");

my $ligne;

print OUT "IGR accession",";","Distance",";","Size","\n";

while ($ligne = <IN>)
{
chomp($ligne);
my @tmp=split("\t",$ligne);
my $num_cle=$tmp[0];
my $start_rna=$tmp[5];
my $end_riboswitch=$tmp[15];
my $distance=0;
my $size=$tmp[4];
if ($start_rna>$end_riboswitch)
{
    $distance=$start_rna-$end_riboswitch;
    print OUT $num_cle,";",$distance,";",$size,"\n";
}
}}

```

The distances are then displayed on a box plot to show the difference between the different cases.

13.5.3 Discover new small RNAs

Many approaches have been used to discover sRNAs. Here, a simple query to look for a putative terminator between two predicted promoters is used as an example of a rapid way to look at sequences with a higher probability of being a sRNA. As for previous examples, we used a MySQL query tailored to this question, but also used the web interface to get intergenic sequences that contain promoters and terminators.

Web interface :

promoter		Predicted putitive promoters	
<input checked="" type="checkbox"/> factor	transcription factor	<input checked="" type="checkbox"/> gap5	Sequence information for 5-prime-UTR
<input checked="" type="checkbox"/> strand	strand of promoter	<input checked="" type="checkbox"/> accession	protein accession like NP_038276.1
<input checked="" type="checkbox"/> position_promoter	position of the promoter	<input checked="" type="checkbox"/> start	start position of 5 prime-UTR
<input checked="" type="checkbox"/> score	score varies from the worst = 0.07 to the best = 2	<input checked="" type="checkbox"/> end	end position of 5 prime-UTR
<input checked="" type="checkbox"/> position_35	position of the box 35	<input checked="" type="checkbox"/> strand	strand direction of the corresponding gene
<input checked="" type="checkbox"/> sequence_35	motif of the box 35	<input checked="" type="checkbox"/> sequence	sequence of 5 prime-UTR in same strand as the gene
<input checked="" type="checkbox"/> position_10	position of the box 10	<input checked="" type="checkbox"/> size	size of 5 prime-UTR
<input checked="" type="checkbox"/> sequence_10	motif of the box 10		
rna_family		Family of RNA according to Rfam(+ transcription terminators + pseudo tRNA)	
<input checked="" type="checkbox"/> fam_id	Rfam accession: RF00001		
<input checked="" type="checkbox"/> fam_name	5S_rRNA		
<input checked="" type="checkbox"/> description	5S ribosomal RNA		
<input checked="" type="checkbox"/> type	gene; rRNA		
<input checked="" type="checkbox"/> note	some description		

Condition:

description	find some pattern	Terminator	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-

Figure 51 - select IGRs with terminator and promoter (Fig. S17 in the supplementary data)

MySQL query:

```
SELECT * FROM (select fragment.fragment, fragment.description, gap5.num_cle,
gap5.sequence, gap5.start, gap5.end, promoteur.position_35 as prom_35_a,
promoteur.sequence_35 as prom_35_seq_a, promoteur.position_10 as prom_10_a,
promoteur.sequence_10 as prom_10_seq_a from fragment inner join gap5 on gap5.fragment =
```

```

fragment.fragment inner join promoteur on promoteur.num_cle = gap5.num_cle)A , (select
gap5.num_cle, gap5.sequence, gap5.start, gap5.end, rna_known.start as rna_start, rna_known.end
as rna_end, rna_family.fam_name from gap5 inner join rna_gap5 on rna_gap5.num_cle =
gap5.num_cle inner join rna_known on rna_gap5.rna_id = rna_known.rna_id inner join
rna_family on rna_family.fam_id=rna_known.fam_id where rna_known.fam_id like
'%RhoINDP%') B, (select gap5.num_cle, gap5.sequence, gap5.start, gap5.end,
promoteur.position_35 as prom_35_b, promoteur.sequence_35 as prom_35_seq_b,
promoteur.position_10 as prom_10_b, promoteur.sequence_10 as prom_10_seq_b from gap5
inner join promoteur on promoteur.num_cle = gap5.num_cle) C WHERE A.num_cle = B.num_cle
and C.num_cle = B.num_cle and A.prom_35_a <A.prom_10_a and A.prom_10_a < B.rna_start
and B.rna_end < C.prom_35_b and C.prom_35_b <C.prom_10_b INTO OUTFILE
'/var/lib/mysql-files/promoteur_terminateur_promoteur.csv';

```

However, in this case a MySQL command is more recommended because we want to select IGRs that have a terminator between two promoters and not only a terminator and a promoter. Unlike the web interface, MySQL allows us to specify the positioning of the terminator and promoters.

The result of the MySQL is displayed in the following: **Table S11 - MySQL result of the existence of a rho-independent terminator between two promoters (separate file)**. Based on these signatures of RITs between two promoters, we can find 227,810 IGRs with a putative sRNA in Table S11, comparison with RiboGap’s known sRNAs reveals 3,357 IGRs from our list of terminators between two promoters. Two examples are shown below in the table S12.

Tableau 16 - occurrence of terminators between two promoters (Table S12 in the supplementary data)

Intergenic sequence accession	Intergenic sequence	Color code	RiboGa results
Sequence2344 7847	GCCTCGGCAGCACGCGTCCGGCGCTGGCCGGGCC TGCATCCCGCCAAGCCGCCCTCGGGCGGCTTGGC TTTTTCACCCGGCCGATTTTGACACGGGGCTGGA AATCCGTCATAATTCCTGACTTCGGTTCGGGTCGT TAGCTCAGTTGGTAGAGCAGCGACTTTTAATCC	Promoter	no known sRNA found

	<p>GTTGGTCGC GCGTTCGAGTCGCGCACGACCCACC AACCAGATTCAGCAAGCAAGGCCAGCAACAAGGC CAGCAAGA AAAAAGCCGCTCATGGAAGCGGCTTT TTTCTTGCTGGTTTTTTTCTTGCTGGCTATTTTC ATGCGGTGCGCCTGGTTGTTGCGCGCGGCCATTA TCCCGCCTTCGATATCCGTTTGCCGTGACGCGGT TTCCCGACAATTGGGAAACCGCCGCGCCTGCCTG CCTGCGCG GATCCCGCAGCCTGCTAAGTGTGAAGA TTCAATAGGTTGTATGCATGGTTCATCCGAACCG GATTTGAGAAACTGGAAATCGCCACCCCCCAGT TCACTCAAGGAGCCCGGCCGG</p>	<p>Box -35</p> <p>Box -10</p> <p>RIT</p>	
<p>Sequence2347 5143</p>	<p>GCCTCGGCAGCACGCGTCCGGCGCTGGCCGGGCC TGCATCCCGCCAAGCCGCCCTCGGGCGGCTTGGC TTTTTCACCCGGCCGATTTTGACACGGGGCTGGA AATCCGTCATAATCCTGACTTCGGTCGGGTCGT TAGCTCAGTTGGTAGAGCAGCGGACTTTTAATCC GTTGGTCGC GCGTTCGAGTCGCGCACGACCCACC AACCAGATTCAGCAAGCAAGGCCAGCAACAAG GC CAGCAAGAAAAAGCCGCTCATGGAAGCGGCTTT TTTCTTGCTGGTTTTTTTCTTGCTGGCTATTTTC ATGCGGTGCGCCTGGTTGTTGCGCGCGGCCATTA TCCCGCCTTCGATATCCGTTTGCCGTGACGCGGT TTCCCGACAATTGGGAAACCGCCGCGCCTGCCTG CCTGCGCG GATCCCGCAGCCTGCTAAGTGTGAAGA TTCAATAGGTTGTATGCATGGTTCATCCGAACCG GATTTGAGAAACTGGAAATCGCCACCCCCCAGT TCACTCAAGGAGCCCGGCCGG</p>	<p>Promoter</p> <p>Box -35</p> <p>Box -10</p>	<p>no known sRNAfound</p>

13.5.4 Motif finder : G-quadruplex example

Pattern matching through the simple “find some pattern” (MySQL “LIKE”) or through the more complex REGEX options offers powerful ways to look for keywords and sequence motifs. As an example, we looked for G-quadruplexes. While G-quadruplexes with only two sets of four Gs stacked together are known, and numerous examples of three Gs stacked within a quartet are also

known, we used a pattern requiring tracks of at least four consecutive Gs as an example here to increase their likelihood of being real.

Formulating the query from the web interface:

cds	Coding sequence	fragment	Chromosome information
<input type="checkbox"/> accession <input checked="" type="checkbox"/> gene <input checked="" type="checkbox"/> DNA_seq <input checked="" type="checkbox"/> locus_tag <input checked="" type="checkbox"/> product <input type="checkbox"/> translation <input type="checkbox"/> gi <input checked="" type="checkbox"/> start <input checked="" type="checkbox"/> end <input checked="" type="checkbox"/> strand <input type="checkbox"/> virulence	protein accession like NP_038276.1 gene name like rhlA DNA sequence of gene ex:Marme_0002 ex:Mg transporter amino acid sequence protein gi like: 326793324 start position end position strand direction of the gene relative to chromosome Gene virulence is known=VF, or not=NULL	<input checked="" type="checkbox"/> DNA fragment <input type="checkbox"/> frag_gi <input type="checkbox"/> length <input type="checkbox"/> is_circular <input type="checkbox"/> strain <input type="checkbox"/> chromosome <input type="checkbox"/> plasmid <input type="checkbox"/> GC <input type="checkbox"/> gene_num <input type="checkbox"/> IGR_length <input type="checkbox"/> IGR_GC <input type="checkbox"/> IGR_median <input type="checkbox"/> organism_id <input type="checkbox"/> taxonomy <input checked="" type="checkbox"/> description	Refseq accession number like NC_000913 gi of Refseq accession like 158421624 length of chromosome true/false strain information like "Newman" chromosome if exist in database plasmid if exist in database GC % like 52.17 number of genes in chromosome like: 4934 total length of intergenic sequences GC % in intergenic sequences median size of intergenic sequences project id of organism according to NCBI bacteria; elusimicrobia; environmental samples Staphylococcus aureus subsp. aureus str. Newman

Condition:

DNA_seq	find some pattern	'G{4,}[A C G T]{1,7}G{4,}[A C G T]{1,7}G'	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-

Figure 52 - Using the web interface to select coding sequences that have G4 motif with REGEXP (Fig. S18 in the supplementary data)

Generated MySQL query:

```
select * from cds inner join fragment on cds.fragment=fragment.fragment where cds.DNA_seq
REGEXP 'G{4,}[A|C|G|T]{1,7}G{4,}[A|C|G|T]{1,7}G{4,}[A|C|G|T]{1,7}G{4,}'
INTOOUTFILE '/var/lib/mysql-files/g_quadruplex.csv';
```

This led to over 6,000 coding sequences that had such patterns:

Table S13 - MySQL result of G-quadruplex motif search in RiboGap (separate file).

You can find some examples below:

Tableau 17 - Comparison of G-quadruplex motif search in RiboGap with G4RNA analysis. The motif box is highlighted in yellow (Table S14 in the supplementary data)

Sequence accession	Sequence	G4RNA's result
Sequence3206718	GTGGACGCCTTCGCCGACTGGTTCTTCGTGGCCTT TTGGGTCCTCGGGGTCCTCCTCACCCCTCCTTCCCT TCGTCCCCGCCACCTTGGTGATCCTCTTCGGGGCC CTGGTGCACGAGCTTCTCGTGGGCTTCCGGGAGCT TTCCCTGGGGACGTGGCTTGGGCTTGGGGCCCTGG CCCTCCTCGCCATGCTTTTGGACAACGTGGCCGCC CTGGTGGGGGCCAGGCGCTACGGGGCGGGACGGGC GGGGCTTTGGGGGGCCTTTTGGGGGGCGTTTTGG GG CCTCTTCTTCGGGGTGGTGGGGGTCTTGGTCCTC CCTTTCCTCCTCGCTTGGCTCTTTGAGTACCTCTC GGAAGGCGGCCGGAGGAGGCGCTGCGGGCGGCCT GGGGACCCCTGGTGGGGCTTATGGGCGGGGTGGTG GCCAAGGTCTTCGTCCACCTGGCCATGGGGGTTCT GGTCTCAGGGCCATCTTTTGA	found as putative G- quadruplex
Sequence2341518	GTGATCGTGAGAAGGACGTGGCTGCTCGGTGCGCT CGTCGTGGCCCTGGCGGCGTGCAAGGGATCCTCGG GCGAGGACGGGACGGC GGGGCAGGCGGGGCTGCAG GGGCCACAGGGG CCGGAGGGCCCCGCGGGTCCGGC CGGGCCGCCGGGGCCCGCGGGGCCCGCGGGCGAGG CCGGGCCGATCGGTCCCGCCGGCGCGCCCGGGCCC GCTGGCGCGACCGGCCCGGCGGGCGCGACCGGCC GGCGGGCGCGACCGGTCCTGCGGGCCCCGGTCGGCG CGCCGGGACCGATGGGGCCGGCCGGCTTGCAGGGG CCGCAGGGGCCCGAGGGACCTCCGGGGCCGGAATC ACGCTTCGGGGAGCCCGGCTACGCCGTGACGGGC GCGGCCGCGAGTGCACCATCGGTGAGGTGTGGCTC GTCGCCGGCTCCGTGCGGGCGCGACCGGGCGCG CGGGCAGCTGCTGTCGATCAGCTCGAACACCGCGC TGTTCTCGCTGCTCGGGACGCTGTACGGCGGGCAG GGCAGGACCAGTTTCGCGCTGCCCGACCTCAGCCA GCGGGCCCCGAACGGCCTCACCTACGTATCTGCA CCCAGGGCATCTTCCCGGCGCGCCTCTGA	Found as putative G- quadruplex

13.5.5 Intergenic sequences with cis regulatory RNAs

To calculate the number of hits from gap5 regions which regulate translation in cis and which have ncRNAs such as riboswitch, T-box, thermoregulator or leader types, we basically used the MySQL function count () as show below for different regulatory elements.

Tableau 18 - Different MySQL queries that calculate the number of intergenic sequences (Table S15 in the supplementary data)

Objective of the query	MySQL query	Result	
Number of sequences that have riboswitch, leader, T-box or thermoregulator and at the same time being in front of start codon “ATG” and close to AGGAGG	<pre> select rna_family.type, count(gap5.sequence) from cds inner join gap5 on gap5.num_cle=cds.num_cle inner join rna_gap5 on gap5.num_cle = rna_gap5.num_cle inner join rna_known on rna_gap5.rna_id = rna_known.rna_id inner join rna_family on rna_family.fam_id = rna_known.fam_id where sequence REGEXP 'AGGAGG[A C G T]{5,12}\$' and DNA_seq REGEXP '^ATG' and (rna_family.type Like '%thermoregulator%' OR rna_family.type Like '%riboswitch%' OR rna_family.type Like '%leader%' OR rna_family.fam_name like '%T-box%') group by rna_family.type </pre>	leader	18 668
		riboswitch	12 875
		thermoregulator	2 138
		T-box	13 549
Number of sequences that have riboswitch, leader, T-box or thermoregulator and at the same time being in	<pre> select rna_family.type, count(gap5.sequence) from cds inner join gap5 on gap5.num_cle=cds.num_cle inner join rna_gap5 on gap5.num_cle = rna_gap5.num_cle inner join rna_known on rna_gap5.rna_id </pre>	leader	1
		riboswitch	2

front of start codon “CTG” and close to AGGAGG	= rna_known.rna_id inner join rna_family on rna_family.fam_id = rna_known.fam_id where sequence REGEXP 'AGGAGG[A C G T]{5,12}\$' and DNA_seq REGEXP '^CTG' and (rna_family.type Like '%thermoregulator%' OR rna_family.type Like '%riboswitch%' OR rna_family.type Like '%leader%' OR rna_family.fam_name like '%T-box%') group by rna_family.type	thermoregulator	1
		T-box	1
Number of sequences that have riboswitch, leader, T-box or thermoregulator and at the same time being in front of start codon “TTG” and close to AGGAGG		leader	949
		riboswitch	886
		thermoregulator	119
		T-box	787
Number of sequences that have riboswitch,	select rna_family.type, count(gap5.sequence) from cds inner join gap5 on gap5.num_cle=cds.num_cle	leader	847

leader, T-box or thermoregulator and at the same time being in front of start codon “GTG” and close to AGGAGG	inner join rna_gap5 on gap5.num_cle = rna_gap5.num_cle inner join rna_known on rna_gap5.rna_id = rna_known.rna_id inner join rna_family on rna_family.fam_id = rna_known.fam_id where sequence REGEXP 'AGGAGG[A C G T]{5,12}\$' and DNA_seq REGEXP '^GTG' and (rna_family.type Like '%thermoregulator%' OR rna_family.type Like '%riboswitch%' OR rna_family.type Like '%leader%' OR rna_family.fam_name like '%T-box%') group by rna_family.type	riboswitch	1506
		thermoregulator	21
		T-box	532
Number of sequences that have riboswitch, leader, T-box or thermoregulator and at the same time being in front of start codon “ATG”	select rna_family.type, count(gap5.sequence) from cds inner join gap5 on gap5.num_cle=cds.num_cle inner join rna_gap5 on gap5.num_cle = rna_gap5.num_cle inner join rna_known on rna_gap5.rna_id = rna_known.rna_id inner join rna_family on rna_family.fam_id = rna_known.fam_id where DNA_seq REGEXP '^ATG' and (rna_family.type Like '%thermoregulator%' OR rna_family.type Like '%riboswitch%' OR rna_family.type Like '%leader%' OR rna_family.fam_name like '%T-box%') group by rna_family.type	leader	94 195
		riboswitch	116 889
		thermoregulator	26 812
		T-box	41 554
Number of sequences that have riboswitch,	select rna_family.type, count(gap5.sequence) from cds inner join gap5 on gap5.num_cle=cds.num_cle	leader	138

leader, T-box or thermoregulator and at the same time being in front of start codon “CTG”	<pre>inner join rna_gap5 on gap5.num_cle = rna_gap5.num_cle inner join rna_known on rna_gap5.rna_id = rna_known.rna_id inner join rna_family on rna_family.fam_id = rna_known.fam_id where DNA_seq REGEXP '^CTG' and (rna_family.type Like '%thermoregulator%' OR rna_family.type Like '%riboswitch%' OR rna_family.type Like '%leader%' OR rna_family.fam_ name like '%T-box%') group by rna_family.type</pre>	riboswitch	469
		thermoregulator	49
		T-box	31
Number of sequences that have riboswitch, leader, T-box or thermoregulator and at the same time being in front of start codon “TTG”	<pre>select rna_family.type, count(gap5.sequence) from cds inner join gap5 on gap5.num_cle=cds.num_cle inner join rna_gap5 on gap5.num_cle = rna_gap5.num_cle inner join rna_known on rna_gap5.rna_id = rna_known.rna_id inner join rna_family on rna_family.fam_id = rna_known.fam_id where DNA_seq REGEXP '^TTG' and (rna_family.type Like '%thermoregulator%' OR rna_family.type Like '%riboswitch%' OR rna_family.type Like '%leader%' OR rna_family.fam_ name like '%T-box%') group by rna_family.type</pre>	leader	5 953
		riboswitch	6 291
		thermoregulator	1 223
		T-box	1 859
Number of sequences that have riboswitch, leader, T-box or thermoregulator and at the same time being in	<pre>select rna_family.type, count(gap5.sequence) from cds inner join gap5 on gap5.num_cle=cds.num_cle inner join rna_gap5 on gap5.num_cle = rna_gap5.num_cle inner join rna_known on rna_gap5.rna_id</pre>	leader	5 834
		riboswitch	13 506

front of start codon “GTG”	= rna_known.rna_id inner join rna_family on rna_family.fam_id = rna_known.fam_id where DNA_seq REGEXP '^GTG' and (rna_family.type Like '%thermoregulator%' OR rna_family.type Like '%riboswitch%' OR rna_family.type Like '%leader%' OR rna_family.fam_name like '%T-box%') group by rna_family.type	thermoregulator	1 956
		T-box	2 603

These different queries can be also formulated from the interface. Let’s take the example of number of IGRs that have a thermoregulator with the Shine-Dalgarno at a distance of 5-12 nt from the start codon “GTG”.

cds	Coding sequence	rna_family	Family of RNA according to Rfam(+ transcription terminators + pseudo tRNA)																												
<input type="checkbox"/> accession <input checked="" type="checkbox"/> gene <input checked="" type="checkbox"/> DNA_seq <input checked="" type="checkbox"/> locus_tag <input checked="" type="checkbox"/> product	protein accession like NP_038276.1 gene name like rhlA DNA sequence of gene ex:Marme_0002 ex:Mg transporter	<input checked="" type="checkbox"/> fam_id <input checked="" type="checkbox"/> fam_name <input checked="" type="checkbox"/> description <input checked="" type="checkbox"/> type <input checked="" type="checkbox"/> note	Rfam accession: RF00001 5S_rRNA 5S ribosomal RNA gene; rRNA some description																												
gap5	Sequence information for 5-prime-UTR	Condition:																													
<input checked="" type="checkbox"/> accession <input checked="" type="checkbox"/> start <input checked="" type="checkbox"/> end <input checked="" type="checkbox"/> strand <input checked="" type="checkbox"/> sequence <input checked="" type="checkbox"/> size	protein accession like NP_038276.1 start position of 5 prime-UTR end position of 5 prime-UTR strand direction of the corresponding gene sequence of 5 prime-UTR in same strand as the gene size of 5 prime-UTR	<table border="1"> <tr> <td>type</td> <td>find some pattern</td> <td>Thermoregulator</td> <td>AND</td> </tr> <tr> <td>DNA_seq</td> <td>find some pattern</td> <td>^GTG</td> <td>AND</td> </tr> <tr> <td>sequence</td> <td>find some pattern</td> <td>^AGGAGG[A C G T]{5,12}\$</td> <td>-</td> </tr> <tr> <td>-</td> <td>-</td> <td>-</td> <td>-</td> </tr> <tr> <td>-</td> <td>-</td> <td>-</td> <td>-</td> </tr> <tr> <td>-</td> <td>-</td> <td>-</td> <td>-</td> </tr> <tr> <td>-</td> <td>-</td> <td>-</td> <td>-</td> </tr> </table>		type	find some pattern	Thermoregulator	AND	DNA_seq	find some pattern	^GTG	AND	sequence	find some pattern	^AGGAGG[A C G T]{5,12}\$	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
type	find some pattern	Thermoregulator	AND																												
DNA_seq	find some pattern	^GTG	AND																												
sequence	find some pattern	^AGGAGG[A C G T]{5,12}\$	-																												
-	-	-	-																												
-	-	-	-																												
-	-	-	-																												
-	-	-	-																												

Figure 53 - Select Thermoregulator and “perfect Shine-Dalgarno” from the web interface (Fig. S19 in the supplementary data)

14 ANNEXE V: L' ARTICLE « A SURVEY OF CIS REGULATORY NON-CODING RNA INVOLVED IN BACTERIAL VIRULENCE »

Sachant que la partie Bio-informatique de cet article est faite en exécutant la requête dans la base de données RiboGap v2, qui est une partie majeure dans ce travail. Nous avons décidé d'inclure cet article publié par bioRxiv comme annexe dans ce mémoire pour démontrer les différentes utilisations de RiboGap v2. À noter que les fichiers Excel du matériel supplémentaire associés à cet article ne sont pas inclus.

Ci-joint le lien de l'article complet ainsi que le matériel supplémentaire : <https://www.biorxiv.org/content/10.1101/2021.11.03.467129v1>



14.1 Résumé

L'étude de la pathogénèse chez les bactéries est importante pour trouver de nouvelles cibles médicamenteuses pour traiter les infections bactériennes. Les bactéries pathogènes, y compris les opportunistes, expriment de nombreux gènes dits de virulence pour échapper aux défenses naturelles et au système immunitaire de l'hôte. La régulation des gènes de virulence est souvent nécessaire pour que les bactéries infectent leur hôte. Une telle régulation peut être réalisée par des ARN cis-régulateurs, comme les thermorégulateurs ou les *riboswitchs* se liant aux métabolites. Malgré les centaines de familles d'ARN annotées comme cis-régulatrices, il existe relativement peu d'exemples de tels ARN non codants (ARNnc) dans les régions 5'-non traduites

(UTR) de bactéries décrites pour réguler les gènes de virulence en aval. Pour réévaluer les rôles potentiels de ces éléments régulateurs dans la pathogenèse bactérienne, nous avons collecté des gènes importants pour la virulence à partir de différentes bases de données et évalué la présence d'ARNnc dans leurs UTR pour mettre en évidence le rôle potentiel de ce type de régulation génique pour la virulence et, en même temps, obtenir un aperçu de certains des déclencheurs physiques et chimiques de la virulence.