Thèse présentée devant

l'Université Paul Sabatier

Détection des ARNnc dans les séquences génomiques. Application au génome de *Ralstonia solanacearum*

par

KOZOMARA Ana

Sous la direction de : **GASPIN Christine, BOUCHER Christian et CIERCO Christine**

Année 2009 Ecole Doctorale SEVAB

Table des matières

Nomenclature

ACUR Les régions dans le génome de *Ralstonia solanacearum* à l'usage de codon alternatif (Alternatif Codon Usage Regions)

DLPR Distribution des longueurs des plages reconstruites dans les algorithmes de reconstruction du chemin caché

IGR Région intergénique (InterGenic Region)

IS Séquences d'insertion

NCBI (National Center for Biotechnology Information) est l'institut national américain pour l'information biologique et moléculaire. Le NCBI héberge des bases de données génomiques de référence telles que GenBank, dbSNP ou RefSeq

nr La banque des séquences nucléiques non-redondantes de NCBI (non-redundant)

Introduction

Les molécules d'ARN ont été longtemps confinées au rôle du porteur du message génétique *via* les ARN messagers, l'étude de la régulation des processus cellulaires étant dirigée essentiellement vers la régulation protéique. Or les dernières décennies ont vu apparaître la découverte d'un grand nombre d'ARN non-traduits ayant un rôle régulateur dans la cellule au même titre que les protéines.

Si les preuves de l'importance et de l'omniprésence des ARN non-codant ne cessent de s'accumuler, les signaux caractérisant leur séquence sont encore mal compris. De ce fait et contrairement aux gènes codant pour les protéines, dont la détection est aujourd'hui bien maîtrisée, la détection bioinformatique des ARN non-codant reste un problème ouvert, malgré les différentes méthodologies de prédiction qui ont été employées afin de le résoudre. Toutefois, dans certains génomes A+T riches, les séquences des ARN non-codant possèdent les propriétés intrinséques pouvant être exploitées pour leur détection. Au contraire, les propriétés des séquences des ARN non-codant des génomes G+C riches n'ont pas été étudiées à ce jour.

C'est dans ce contexte d'exploration des propriétés des séquences des ARN non-codant dans les génomes G+C riches que se situe cette thèse. A l'intersection des mathématiques, de l'informatique et de la biologie, elle s'inscrit dans une démarche multidisciplinaire visant à contribuer aux connaissances sur la détection des ARN non-codant dans de tels organismes.

Trois thématiquesont été abordées au cours de la thèse. La première, présentant la partie la plus théorique de travail, est l'exploration d'une approche statistique pour établir une différence de composition entre les ARN non-codant et le reste du génome. Dans les cas favorables, il a été possible d'utiliser une approche de segmentation de génome afin de détecter des ARN non-codant.

Deuxième thématique est l'approche comparative pour détecter des ARN non-

codant. Cette partie de la thèse est plus appliquée et vise notamment à améliorer l'outil RNAsim, implémentant une approche dont l'objectf est de permettre d'exploiter la conservation entre les régions intergéniques de plusieurs génomes.

Troisième thématique est la bio-analyse des résultats de RNAsim qui a été conduite afin de proposer les candidats pour une validation biologique.

Les développements réalisés au cours de la thèse ont été appliqués au génome de *Ralstonia solanacearum*, une bactérie phytopathogène modèle au génome G+C riche dont la souche GMI1000 est séquencée et annotée. Si les protéines régulatrices de la pathogénie sont à ce jour bien connues, aucune recherche d'ARN non-codant n'a été entreprise dans cette bactérie. Cet organisme présentait donc pour nous plusieurs aspects intéressants : un génome G+C riche, l'absence d'études sur les ARN non-codant dans cet organisme et enfin proximité des équipes d'accueil. De plus, la disponibilité récente des séquences de deux autres souches de cette bactérie nous a permis d'utiliser une approche comparative en plus de l'approche par biais de composition. Les résultats issus de ce travail de thèse seront soumis à une validation expérimentale.

En accord avec la nature multidisciplinaire de ce travail, la thèse a été effectuée en collaboration étroite entre deux équipes de l'INRA Toulouse, l'équipe SaAB du BIA (Biométrie et Intelligence artificielle), impliquée dans le développement et la mise à disposition des outils d'analyse des génomes et l'équipe de C. Boucher et S. Génin du LIPM (Laboratoire Interaction Plantes-Microorganismes), étudiant les mécanismes moléculaires de pathogénie de *Ralstonia solanacearum*.

Structure du manuscript

Le manuscript est composé de cinq parties. La première partie introduit la thématique des ARN non-codant du point de vue biologique et bio-informatique. Les parties II, III et IV correspondent aux trois grandes parties de la thèse et dans la cinquième partie nous discutons les résultats et présentons les perspectives de ce travail de thèse.

Etant donnée la matière hétérogène présentée, allant de la méthodologie statistique aux considérations biologiques en passant par les éléments d'algorithmique, un état de l'art unifié des questions abordées a été difficile à concevoir. A la place, j'ai fait le choix d'une présentation découpée en fonction de la question traitée. Ainsi, la présentation de chacune des trois parties de la thèse sera précédée par le point sur les connaissances

actuelles sur la question traitée.

Notons encore qu'un index des notions employées tout au long du manuscript est donné à la fin afin de faciliter la lecture.

Première partie

Introduction générale aux ARN non-codant : contexte biologique et bioinformatique

Chapitre 1

Contexte biologique

Ce premier chapitre présente tout d'abord les bases biologiques des ARN non-codant (section 1.1). Nous présentons leurs principales propriétés, avec un accent particulier sur celles que nous avons exploitées lors du travail de thèse. Néanmoins, les notions de base concernant la définition et la structure des ARNnc ont été omises, en considérant qu'elles sont aujourd'hui bien connues des deux communautés, bioinformatique et biologique.

Nous présentons ensuite les principales fonctions et processus pour lesquels l'implication des ARN non-codant est aujourd'hui connue et les différents classes d'ARN en fonction de leur mode d'action. Nous nous focaliserons surtout sur la présentation des ARN non-codant bactériens en donnant parfois les exemples liés à la pathogénie, car les dévéloppements de la thèse ont été appliqués à un génome bactérien étudié pour son pouvoir pathogène.

Dans un deuxième temps nous présenterons les connaissances actuelles sur la distribution taxonomique des ARN non-codant au sein des organismes bactériens (section 1.2).

Enfin, nous présenterons le génome de *Ralstonia solanacearum*, du point de vue de la régulation de la pathogénie, particulièrement bien étudiée dans cet organisme, ainsi que du point de vue de l'organisation génomique (section 1.3).

1.1 ARN non-codant

1.1.1 Propriétés des ARN non-codant

Les connaissances sur les ARNnc ont été longtemps confinées aux ARN de transfert (ARNt) et aux ARN ribosomiques (ARNr). Leur rôle régulateur était alors considéré comme étant mineur par rapport au rôle des protéines.

Ces dernières années ont vu une grande progression des connaissances sur les ARNnc. De nouveaux ARNnc ont été détectés, d'abord par hasard, dans les études des protéines et du fait de leur abondance dans le milieu cellulaire, et ensuite dans le cadre de recherches systématiques, soit par des méthodes expérimentales, soit en utilisant des stratégies bioinformatiques appliquées sur les séquences génomiques disponibles en nombre toujours croissant. La présence des ARNnc est aujourd'hui connue dans les organismes des trois règnes de la vie, les bactéries, les archées et les eucaryotes. Leur implication dans une vaste gamme de processus cellulaires, tels la transcription, la traduction, l'épissage, la virulence et l'adaptation, a été établie. En 2005, on considérait que pour environ la moitié des ARNnc identifiés, une fonction avait pu être attribuée (Hüttenhofer *et al.* (2005)).

Les séquences dans le génome correspondant aux ARNnc sont appelés *les gènes ARN* et leur nom s'écrit avec la première lettre en miniscule en italic, par analogie aux gènes codant pour les protéines. Leur transcript, qui est l'ARNnc fonctionnel est noté par une majuscule.

Une molécule d'ARN est constituée d'une succession de 4 différents types de nucléotides représentés par les caractères A, C, G et U, qui définit sa *structure primaire*. Certaines régions de ces molécules simple-brin peuvent se replier par le biais d'appariements entre nucléotides (principalement les appariements Watson-Crick, C-G, G-C, A-U, U-A et les appariements Wobble, G-U et U-G), formant ainsi la structure de la molécule. La représentation planaire (sans pseudonoeud) de ces appariements est appelée *structure secondaire*. La conformation dans l'espace tri-dimensionnel ainsi que les liaison complémentaires entre les différents éléments de la structure secondaire forment la structure tertiaire, conférant à l'ARN sa fonction dans la cellule. Ces trois aspects structuraux des ARNnc sont représentés sur la figure 1.1. La structure tertiaire étant compliquée à représenter et à analyser, la structure secondaire est souvent utilisée lors des études portant sur les ARNnc.

Propriétés de la séquence primaire des ARNnc Les ARNnc qui sont transcrits indépendament des gènes adjacents conteniennent des signaux de début et de fin de transcription dans leurs régions 5' et 3'. De tels signaux identifiés dans les régions intergéniques peuvent donc indiquer la présence d'ARNnc. Cependant, les séquences promotrices, indiquant l'endroit du début de la transcription, restent peu connues pour la plupart des organismes bactériens.

Deux mécanismes de terminaison de la transcription existent. Le premier dépend de la protéine rho et le deuxième, dit rho-indépendant, repose sur la présence d'une structure en tige-boucle se trouvant en aval de la séquence transcrite. L'utilisation des terminateurs rho-indépendants pour la détection des ARNnc est exposée dans la partie *Analyse des résultats* (section 10.1).

Aucune propriété caractéristique commune à l'ensemble des séquences d'ARNnc n'est connue à ce jour. Néanmoins, les ARNnc des génomes A+T riches, et notamment ceux des génomes d'archées A+T riches hypethermophiles, présentent une composition plus élevée en G+C que le reste du génome. Cette propriété a permis l'identification de nouveaux ARNnc dans ces organismes (Klein *et al.* (2002), Schattner (2002)).

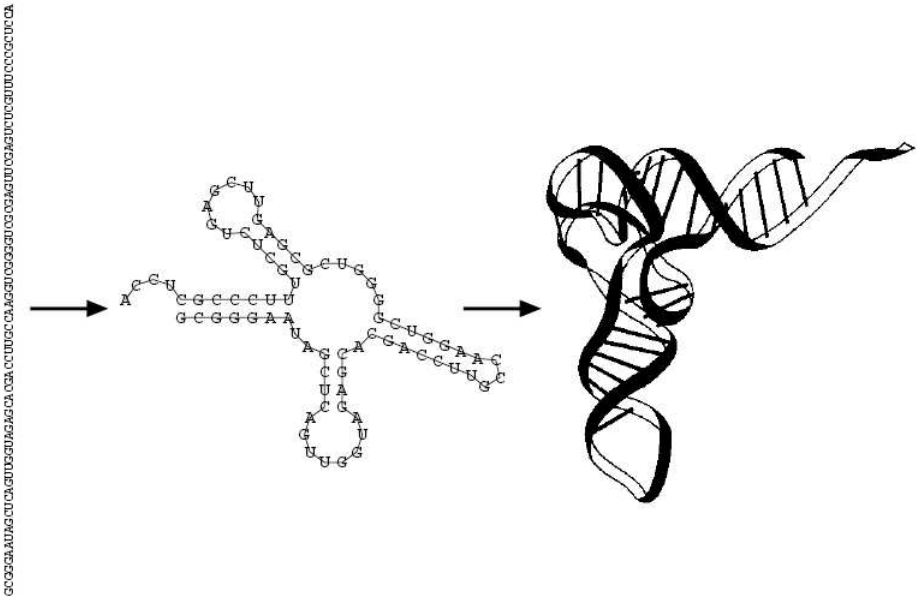


FIG. 1.1 – Les différentes structures d'un ARN : primaire, secondaire et tertiaire (Repris de Washietl (2005)).

En revanche, à ce jour, il n'existe pas d'étude mettant en évidence ce type de propriété chez les organismes ayant une composition en G+C élevée.

Les travaux portant sur ce sujet sont plus amplement présentés dans la partie *Approche ab initio* (section 3).

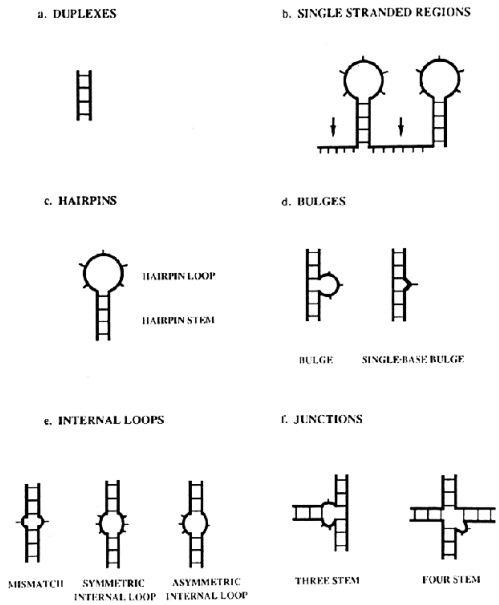


FIG. 1.2 – Les motifs élémentaires d'une structure secondaire d'ARN (repris de http ://darwin.nmsu.edu/ molb470/fall2004/projects/chinna/).

Modèle thermodynamique Nous présentons ici le modèle thermodynamique le plus souvent utilisé pour évaluer la structure secondaire des ARNnc (Cech et Atkins (2005)).

Le modèle thermodynamique repose sur l'hypothèse que l'ARN adoptera la structure secondaire la plus stable thermodynamiquement, c'est-à-dire celle dont l'énergie libre est la plus basse. Il s'appuie sur le modèle dit *du plus proche voisin* (Borer *et al.* (1974)) qui permet de calculer l'énergie d'une structure secondaire comme la somme des motifs élémentaires qui la composent (empilements entre paires de base, différents types de boucles...). Les motifs élémentaires sont décrits dans la figure 1.2).

Familles des ARNnc Comme dans le cas des protéines, les ARNnc peuvent être classés en familles, les ARNnc de la même famille descendant probablement du même ancêtre (Griffiths-Jones *et al.* (2003)). Les membres d'une même famille d'ARNnc peuvent

partager plusieurs propriétés communes, dont la structure secondaire consensus. L'ensemble des propriétés communes d'une famille sont appelés *sa signature*. Les éléments de la signature sont considérés comme essentiels à la fonction de l'ARNnc dans la cellule.

Un exemple classique d'une famille d'ARNnc est donné par la famille des ARNt (figure 1.3) présentant une structure caractéristique en feuille de trèfle. 607 différentes

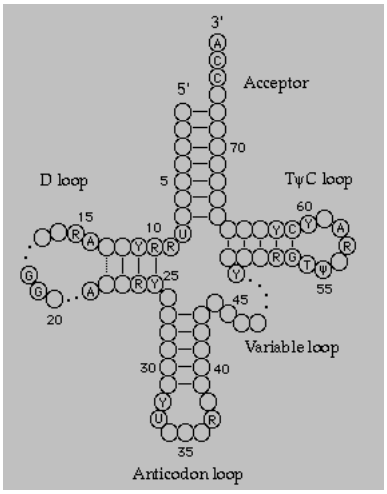


FIG. 1.3 – La structure secondaire consensus d'un ARN de transfert.

familles d'ARNnc, tous règnes confondus, sont ainsi répertoriées dans Rfam version 8.1, voir section 1.2).

1.1.2 Eléments de régulation par les ARN non-codant : fonctions et mécanismes d'action

Dans la présente section nous décrirons quelques principes de régulation par ARNnc, à travers les fonctions et les principaux processus cellulaires pour lesquels une implication des ARNnc a été mise en évidence. Ces fonctions sont exercées par les différents mécanismes d'action sur lesquels nous nous appuierons pour présenter les principales classes d'ARNnc connues aujourd'hui, en focalisant sur les ARNnc bactériens. Enfin, nous présenterons un exemple de régulation de la virulence par un ARNnc.

1.1.2.1 Fonctions des ARNnc

Outre les ARN ribosomiques et les ARN de transfert, indispensables au processus de traduction, on sait aujourd'hui que les ARNnc sont impliqués dans un grand nombre

d'autres processus cellulaires où ils sont souvent responsable des "réglages subtils" (Shimoni *et al.* (2007)). Par exemple, comme la revue de Storz *et al.* (2005) et la figure 1.4 le montrent, les ARNnc sont impliqués dans toutes les étapes de l'expression génique, allant de l'activation de la transcription d'ARNm ou de remodelage de la chromatine (chez les eucaryotes) jusqu'à la traduction en protéines.

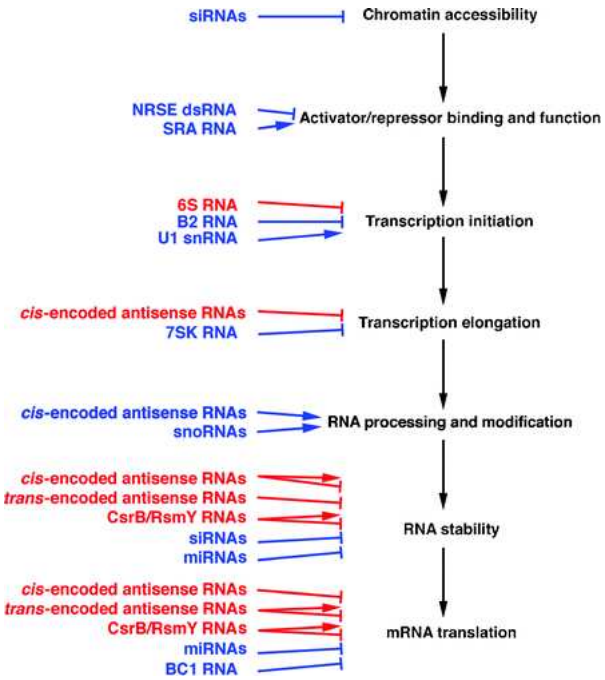


FIG. 1.4 – L'implication des ARNnc dans les différentes étapes de l'expression génique. Les ARNnc bactériens sont indiqués en rouge et les ARNnc eucaryotes en bleu. La régulation positive est indiquée par les flèches et la régulation négative par les barres verticales. Repris de Storz *et al.* (2005).

Les ARNnc sont également impliqués dans bon nombre de cascades de régulation, et font souvent partie des voies de régulation de réponses aux différentes conditions de stress. Des exemples mis en évidence chez *E. coli*, sont présentés sur la figure 1.5. Le rôle des ARNnc impliqués est détaillé dans la légende de la figure.

Certains ARNnc, comme l'ARN de la RNase P, impliquée dans la maturation des précurseurs des ARNt, possèdent une activité enzymatique intrinsèque (Frank et Pace (1998)). La conservation universelle (voir la section 1.2) de cet ARN témoigne de son importance et de l'apparition précoce au cours de l'évolution, d'une forme de régulation par ARNnc.

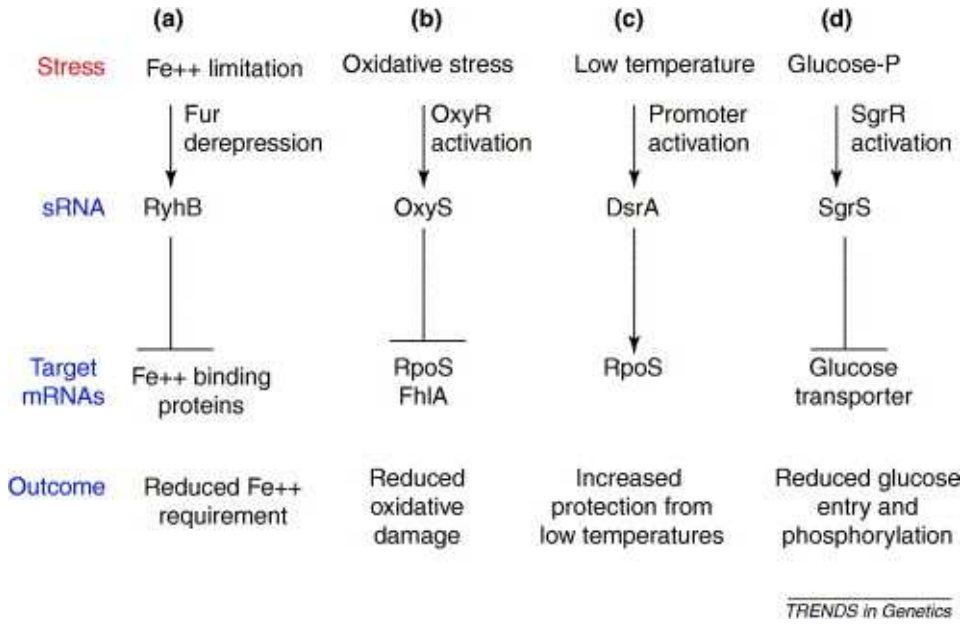


FIG. 1.5 – Les ARNnc bactériens impliqués dans les cascades de régulation. Différentes conditions déclenchent généralement une réponse à travers l'activation ou la synthèse d'un régulateur de transription. La figure présente plusieurs voies de régulation, étudiées chez *E. coli* et impliquant des ARNnc qui sont transcrits à travers cette même voie. (a) L'ARNnc RyhB est produit dans la réponse aux conditions de limitation de l'apport de fer ; RyhB entraîne une dégradation rapide de plusieurs ARNm cibles codant pour les protéines non essentielles dont le fer est le substrat, provoquant ainsi une réduction de demande de fer dans la cellule. (b) L'ARNnc OxyS est synthétisé comme réponse au stress oxydatif. OxyS réprime l'expression de deux facteurs de transcription, RpoS et FhlA et ainsi contribue à la protection de la cellule. (c) L'ARNnc DsrA est produit dans des conditions de basse temperature. Il permet d'augmenter le niveau de traduction du facteur de transcription RpoS qui, à son tour, permet de produire un ensemble de gènes impliqués dans la survie de la bactérie dans des conditions de basse temperature.(d)L'ARNnc SgrS est produit dans les conditions d'accumulation de glucose-phosphate dans la cellule. SgrS entraîne la dégradation des ARNm codant pour la protéine responsable du transport de glucose, réduisant ainsi l'accumulation de phosphate de glucose. Repris de Gottesman (2005).

De façon plus générale, l'accumulation des connaissances sur la régulation par ARNnc laisse supposer qu'au moins un ARNnc serait impliqué dans toute voie de régulation de réponse au stress (Gottesman (2005)).

1.1.2.2 Classification des ARNnc par leur mécanismes d'action

Les ARNnc sont impliqués dans les processus cellulaires *via* deux mécanismes d'action principaux : appariement de bases avec une autre séquence nucléique (ADN ou ARN, encore appelé mécanisme *antisens* ; fixation à une protéine ou à un complexe protéique (figure 1.6).

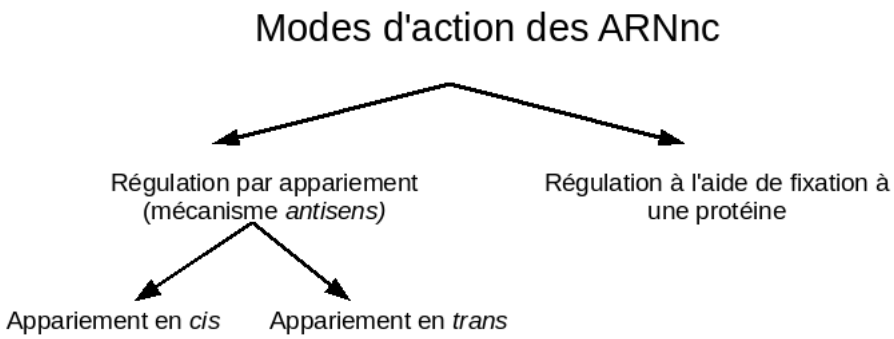


FIG. 1.6 – Les mécanismes d'action d'ARNnc.

Mécanisme de régulation par appariements

Les mécanismes de régulation par appariements peuvent être regroupés en deux grandes classes. La *régulation en cis*, où les ARNnc régulateurs, co-localisés avec les gènes régulés, se trouvent sur le brin opposé et sont parfaitement complémentaires avec une région du gène cible. Le deuxième mécanisme est la *régulation en trans* où l'ARNnc régulateur et le(s) gène(s) cible(s) sont codés dans des régions différentes. La complémentarité entre l'ARNnc et le gène n'est alors généralement pas parfaite.

Des exemples des ARNnc agissant en *cis* sont donnés par l'ARN *sok*, impliqué dans le système *toxin-antitoxin* présent dans de nombreux génomes bactériens (pour une revue voir Gerdes *et al.* (1997)) ou encore, chez les eucaryotes, les *siRNA* (small interfering RNA) permettant de "mettre en sourdine" (silencing, en anglais) l'expression de l'ARN dont ils proviennent, par un mécanisme antisens (pour une revue voir, par exemple,

Meister et Tuschl (2004)).

Chez les bactéries, ces ARNnc ont une longueur d'environ 100nt (Storz *et al.* (2005)).

La régulation par appariements en *trans* a été particulièrement bien étudiée chez *E. coli*. Dans cet organisme, tous les ARNnc connus ayant ce mode d'action (ils sont possiblement au nombre de 36, Zhang *et al.* (2003)), interagissent avec la protéine Hfq (Storz *et al.* (2005)). Cette protéine hexamérique, très conservée et homologue avec une protéine présente aussi chez les eucaryotes et les archées impliquée dans l'épissage, se fixe à des régions simple-brin A+U riches avec une affinité particulière pour des régions flanquant des tige-boucles (Gottesman (2004)). En association avec les ARNnc, cette protéine est impliquée dans leur stabilisation. Les ARNnc agissant en *trans*, en plus d'être associés à la protéine Hfq, constituent tous des unités transcriptionnelles indépendantes et sont induits sous des conditions spécifiques (Storz *et al.* (2005)).

Des exemples d'ARNnc agissant en *trans* chez *E. coli* sont RyhB (Massé *et al.* (2007)) mais aussi DsrA et RprA (Lease et Belfort (2000)). Des ARNnc agissant en *trans* en association avec la protéine Hfq ont été identifiés dans d'autres organismes bactériens tels que *Pseudomonas aeruginosa* (PrrF1 et PrrF2, voir Wilderman *et al.* (2004)), *Vibrio harveyi* et *Vibrio cholerae* (Qrr1, Qrr2, Qrr3 et Qrr4, voir Lenz *et al.* (2004)).

Un exemple, maintenant classique, d'ARNnc agissant en *trans* chez les eucaryotes est donné par la classe des miARN (micro ARN), dont la longueur est d'environ 22nt et qui sont impliqués dans la régulation négative des ARNm cibles (pour une revue voir, par exemple, Voinnet (2009)).

Mécanisme de fixation à la protéine cible

Les protéines cibles des ARNnc, dont l'action se fait *via* la modification de leur activité, appartiennent, selon les connaissances actuelles, soit aux régulateurs transcriptionnels soit aux protéines impliquées dans la régulation de la stabilité et la traduction des ARNm (Storz *et al.* (2005)).

Chez les bactéries, un exemple d'ARNnc modifiant l'activité d'un régulateur transcriptionnel est l'ARN 6S. Il a été démontré que cet ARNnc abondant, universellement conservé parmi les bactéries, se fixe sur l'ARN polymérase σ^{70} en mimant l'ADN ce qui a pour conséquence l'inhibition compétitive de la transcription (pour revue voir Wassarman (2007)).

Dans le groupe des ARNnc inter-agissant avec les protéines impliquées dans la régulation de la stabilité et de la traduction des ARNm, les plus connus sont les membres de la famille CsrB/RsmY, découverts chez *E. coli* et dans différentes espèces de *Pseudomonas*. Ces ARNnc se fixent aux protéines de la famille CsrA/RsmA, impliquées dans la régulation de la traduction et dans la dégradation des ARNm (Babitzke et Romeo (2007)).

A la différence des ARNnc agissant par appariements à des régions de longueur généralement courte, les ARNnc interagissant avec les protéines sont généralement plus longs (>200nt)(Storz *et al.* (2005)).

Riboswitch

Les riboswitch sont une classe particulière de ribo-régulateurs identifiés relativement récemment dans les génomes bactériens. Ils ne sont pas à proprement parler des ARNnc. Ces éléments génétiques structurés sont généralement localisés dans les régions 5' UTR de certains gènes. Ils modulent leur expression en *cis*. Ayant la capacité de changer leur conformation en fonction des changements de concentration de certains métabolites (petites molécules) dans le milieu sans qu'un facteur protéique ne soit impliqué, ils régulent ainsi l'expression du gène se trouvant en aval.

Les riboswitch les mieux caractérisés à ce jour sont composés de deux domaines : une région dite *aptamer* responsable de la fixation du métabolite et une *plateforme d'expression* responsable de la modulation de l'expression du gène en aval (figure 1.7). Chez les bactéries, la platforme d'expression contrôle typiquement la transcription par formation d'un système terminateur-antiterminateur ou bien la traduction, *via* la séquestration du site de fixation du ribosome (Kim et Breaker (2008)).

La régulation peut être effectuée au niveau du contrôle de la transcription *via* un système terminaison-antiterminaison de la plateforme d'expression et au niveau du contrôle de la traduction *via* la séquestration du site de fixation du ribosome.

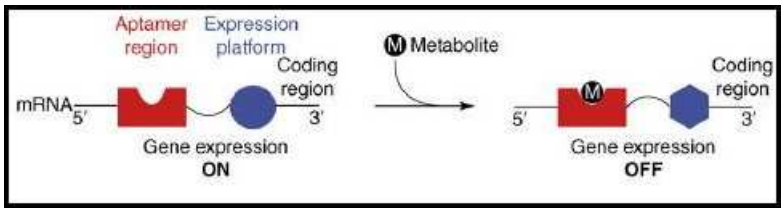


FIG. 1.7 – Structure schématique d'un riboswitch. Repris de Coppins *et al.* (2007)).

Cochrane et Strobel (2008) répertorient 11 différentes familles de riboswitch, correspondant à 11 différents métabolites capables d'être capturés : Guanine, Adenine, Lysine, les co-enzymes FMN (flavin mononucléotide), SAM (S-adenosyl methionine), TPP (thiamine pyrophosphate), Cobalamine, GlcN6P et la glycine, preQ1 (preQuosine1) et deoxyguanosine.

Il est difficile d'estimer le nombre exact de gènes régulés par ce mécanisme. Néanmoins, dans *Bacillus subtilis*, l'organisme le mieux étudié de ce point de vue là, au moins 2% des gènes sont régulés par des riboswitch (Kim et Breaker (2008)). D'autres candidats riboswitch attendent d'être validés, notamment dans les alpha-protéobactéries (Corbino *et al.* (2005)).

1.1.2.3 L'exemple de régulation par ARNnc : régulation de la pathogénie

Nous présentons ici un exemple de régulation par ARNnc. Etant donné que l'organisme étudié au cours de cette thèse est une bactérie phytopathogène Gram négative, notre choix s'est porté sur un exemple bien étudié d'ARNnc, impliqué dans la régulation de la pathogénie d'une autre bactérie phytopathogène Gram négative, *Pectobacterium carotovorum subsp. carotovorum*. Un système similaire a été décrit chez *Pseudomonas syringae*, une autre bactérie phytopathogène (Mole *et al.* (2007)).

Dans *P. carotovoruma* les ARNnc sont impliqués dans la pathogénie à travers le système de régulation post-transcriptionnelle Rsm. Ce système est constitué de deux protéines, RsmA et RsmC et d'un ARNnc régulateur RsmB (Mole *et al.* (2007)), appartenant à la famille des ARNnc CsrB (voir section 1.1.2.2 et figure 1.8).

Dans ce système, RsmA est le régulateur principal et il fixe les ARNm codant pour des facteurs de virulence tels que les enzymes de la dégradation de la paroi végétale, entraînant leur déstabilisation et leur dégradation. L'ARN RsmB est l'antagoniste de la protéine RsmA inhibant par titration de RsmA. RsmB contient une vingtaine de motifs GGA se trouvant dans les boucles des tiges-boucles multiples présentes dans sa structure secondaire (figure 1.8). Ces motifs présentent probablement les sites de fixaton de la protéine RsmA et une molécule d'ARN RsmB peut séquestrer jusuqu'à 18 molécules RsmA, inhibant ainsi, de façon efficace, la dégradation par RsmA des messagers cibles (Liu *et al.* (1998)).

La troisième protéine, RsmC, impliquée dans le système Rsm agit comme répresseur de *rsmB* et comme activateur de *rsmA* même si son mécanisme d'action n'est pas connu

à ce jour.

Plus généralement, les bactéries pathogènes alternent entre différentes niches écologiques (extérieur et intérieur de l'hôte) et doivent être capables de s'adapter aux changements rapides de conditions d'environnement (Romby *et al.* (2006)). De ce fait, ces bactéries ont dévéloppé des réseaux de régulations subtils permettant une réaction rapide aux signaux environnementaux. La régulation de la virulence se fait majoritairement par le biais des protéines mais dernièrement une implication directe ou indirecte dans la pathogénie a été établie pour un nombre important d'ARNnc (voir l'annexe A pour la liste de tels ARNnc et Romby *et al.* (2006) pour la revue).

1.2 Distribution taxonomique des familles d'ARNnc

Les ARNnc connus ne sont pas distribués uniformement à travers les différents organismes. Dans cette section, nous nous intéressons à leur distribution taxonomique, dans les génomes bactériens en particulier. Cette étude est réalisée à partir des données présentes dans Rfam, une banque de données recensant les familles d'ARNnc connues et leur annotation dans un grand nombre de génomes séquencés.

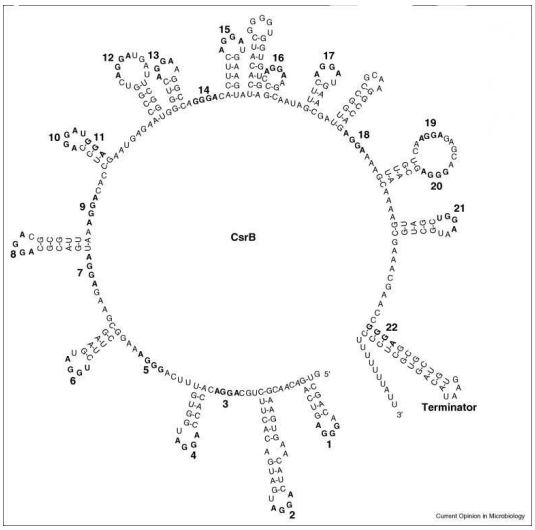


FIG. 1.8 – La structure secondaire prédite de l'ARNnc CsrB de *E. coli* (de la même famille que RsmB de *P. corotovoruma*). Les motifs GGA, les site de fixation de la protéine CsrA (RsmA dans le cas *P. corotovoruma*) sont indiqués en gras et numérotés de 1 à 22. Repris de Babitzke et Romeo (2007).

Rfam : les données utilisées Il existe aujourd'hui plusieurs banques de données dédiées aux ARNnc, pour la plupart spécifiques d'une famille particulière Dans cette étude, et dans le travail de thèse de manière générale, nous avons utilisé les données de Rfam, version 8.1. Par conséquent, lorsque la banque Rfam sera évoquée dans le manuscript de thèse on se référéra, par défaut, à cette version.

Nous nous sommes appuyés sur cette base de données car elle est généraliste et répond aux exigences suivantes :

- Intégre les alignements structuraux pour les familles d'ARN
- Propose des outils bioinformatiques adaptés (notamment le modèle de covariance, Klein et Eddy (2003)) pour rechercher dans les bases de données de séquences tous les homologues détectables des ARNnc déjà annotés

Un inconvénient que la base Rfam présente est l'utilisation d'un filtre qui réalise une première sélection sur la base d'une homologie en séquence. De ce fait, les membres d'une famille peuvent être omis en raison d'une divergence en séquence avec les ARN de cette famille, déjà identifiés par ailleurs et la recherche d'une famille d'ARN peut difficilement dépasser les barrières d'un phylum ou d'une espèce bactérienne.

Enfin, nous avons pu constater que cette base de données n'est pas stable entre versions successives :les nouvelles versions ne recensent pas tous les ARNnc prédits dans les versions antérieures.

Familles reprértoriées dans Rfam A ce jour 607 différentes familles d'ARNnc sont répertoriées dans Rfam, dont 107 pour des organismes bactériens. La figure 1.9, déjà obsolète mais donnée à titre d'illustration, représente la répartition des familles connues en 2005 dans les trois règnes (les bactéries, les archées et les eucaryotes). Elle montre notamment que la plus grande diversité des familles d'ARNnc se situe au sein des organismes bactériens. Cette figure montre aussi que seule une petite proportion de familles est commune entre les différents règnes.

La thèse portant sur la détection des ARNnc bactériens, nous nous focaliserons, dans la suite, sur la distribution taxonomique au sein de ce règne. Nous nous limiterons également aux ARNnc autres que les ARNt et les ARNr, étant donné que ces ARNnc sont universellement présents.

1.2.1 Familles Rfam bactériennes selon les groupements taxonomiques

A ce jour, 685 génomes bactériens, entièrement séquencés, sont recensés au NCBI (version Août 2008) et y sont sont distribués en 18 phyla taxonomiques (figure 1.10).

Pour présenter la distribution taxonomique des familles d'ARN répertoriées dans Rfam nous nous appuyerons sur la classification des organismes bactériens en phyla.

Les familles présentes de façon ubiquitaire chez tous les organismes bactériens dans Rfam sont peu nombreuses : l'ARN de la RNaseP et l'ARN SRP universellement conservés dans les trois règnes de la vie (voir figure 1.9) et l'ARN 6S et le riboswitch de Cobalamine identifiés majoritairement dans les génomes bactériens.

La distribution taxonomique est présentée dans le tableau 1.1 et sur l'histogramme de la figure 1.11. La figure 1.11 montre que le nombre de familles d'ARNnc recensées dans chacun des phyla est corrélé positivement avec le nombre d'organismes séquencés que ce phylum contient. Ce biais est probablement dû au fait que les groupes d'organismes les plus séquencés sont les plus étudiés, en général, et ce aussi pour ce qui concerne les ARNnc. Par exemple, le groupe des gamma-protéobactéries, recensant le plus grand nombre de familles d'ARNnc, contient *Escherichia coli*, l'organisme modèle

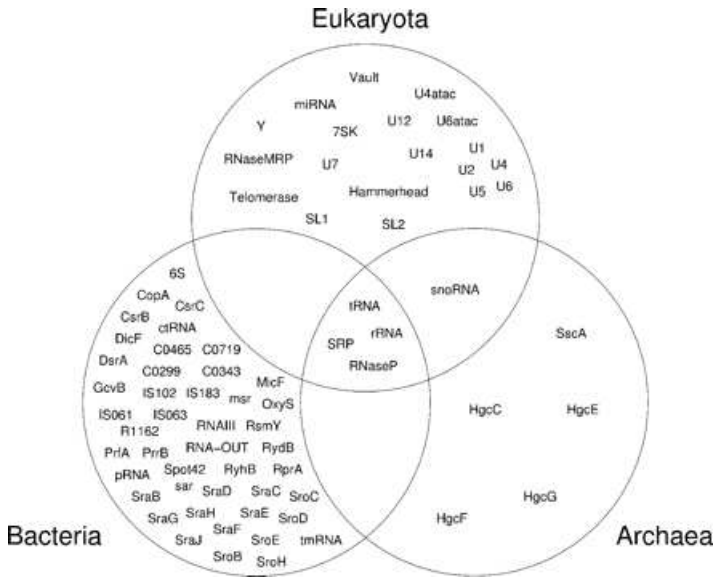


FIG. 1.9 – Distribution des familles d'ARNnc répertoriés dans Rfam (en 2005) dans les trois règnes (repris de Griffiths-Jones *et al.* (2005)).

qui est le mieux étudié d'un point de vue des ARNnc. Dans ce groupe se trouve également l'espèce *Salmonella*, qui a fait l'objet de plusieurs études de recherche et de caractérisation des ARNnc (Vogel (2008)). De plus, l'analyse plus détaillée de la distribution des ARNnc dans le groupe des gamma-protéobactéries montre que tous les organismes contenant plus de 25 ARNnc appartiennent aux enterobactéries, un sous-groupe des gamma-protéobactéries. Les organismes ayant plus de

40 familles d'ARNnc répertoriées appartiennent à la même sous-branche des entérobactéries qu'*E. coli*, ce qui suggère qu'il s'agit des ARNnc conservés entre *E. coli* et ces autres organismes.

Le deuxième groupe le plus peuplé est celui des firmicutes, caractérisés par un nombre important d'ARNnc spécifiques à certains organismes de ce groupe. Le groupe des firmicutes contient un grand nombre de bactéries pathogènes de l'homme (telles que *Staphylococcus aureus*), d'un intérêt économique (telles que *Lactobacillus casei*) ou encore l'organisme modèle *Bacillus subtilis*, ce qui explique le grand nombre de génomes séquencés appartenant à ce groupe, le génome de *S. aureus* étant particulièrement bien exploré d'un point de vue des ARNnc liés à la virulence (Romby *et al.* (2006)).

Ainsi, il est probable que la distribution actuelle des ARNnc reflète nos conaissances

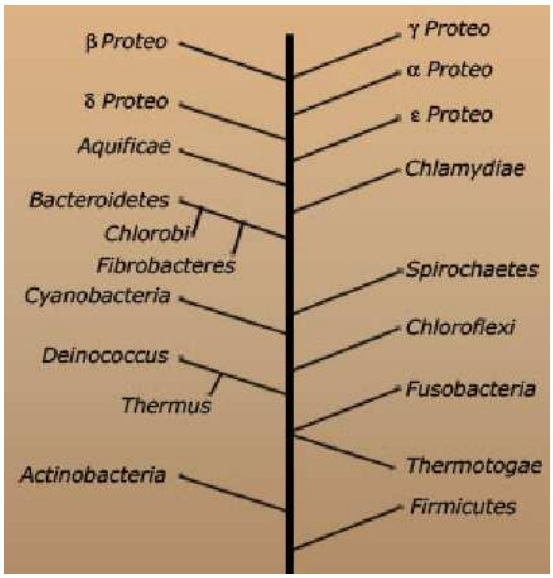


FIG. 1.10 – L'arbre phylogénétique des différents phyla bactériens (repris de www.bacterialphylogeny.info/overview.html). La majorité des phyla de NCBI y est représentée, à l'exception des acidobactéries et plantomycètes (voir tableau 1.1).

Groupe	Nombre d'organismes dans NCBI	Nombre d'organismes dans Rfam	Nombre de familles présentes dans Rfam
Gamma protéo-bactéries	174	119	82
Firmicutes	151	103	33
Alpha protéo-bactéries	89	60	19
Beta protéo-bactéries	60	39	17
Actino bactéries	54	37	15
Delta protéo-bactéries	19	15	14
Cyanobactéries	33	23	12
Chloroflexi	7	4	11
Fusobactéries	1	1	11
Bacteroidetes/Chlorobi	25	16	8
Thermotogae	7	4	8
Deinococcus-Thermus	4	4	8
Acidobactéries	2	2	6
Spirochaetes	14	9	5
Chlamydiae/Verrucomicrobia	13	10	4
Epsilon protéo-bactéries	20	15	4
Aquificae	3	1	2
Planctomycetes	1	0	0
Autres bactéries	8	3	14

TAB. 1.1 – La répartition des familles d'ARNnc dans Rfam selon la classification taxonomique des bactéries trouvée au NCBI.

actuelles et non pas leur réelle distribition taxonomique et l'état actuel ne permet de conslure si certains organismes bactériens sont plus riches en ARNnc que d'autres. Par exemple, il est probable que des ARNnc nouveaux existent dans les nombreuses bactéries possèdant la protéine Hfq, partenaire de nombreux ARNnc déjà connus (voir section 1.1.2.2 et Valentin-Hansen *et al.* (2004)).

A la différence des gamma-protéobactéries ou firmicutes, dans les phyla chlamydiae et epsilon-protéobactéries, il n'existe que 4 familles d'ARNnc identifiés, correspondant aux familles présentes dans toutes les bactéries. En dehors de ces phyla, où peu de génomes ont été séquencés (tableau 1.1), certains groupes bactériens, comme les béta-protéobactéries, sont caractérisés par le nombre élevé de génomes séquencés, indiquant qu'ils sont très étudiés, et par le nombre de familles d'ARNnc relativement petit, indiquant que la régulation par les ARNnc n'y a pas été étudiée de façon spécifique. De

plus ce groupe bactérien contient un grand nombre de génomes G+C riches dont la composition des ARNnc est inexplorée à ce jour et présente un défi méthodologique d'un point de vue de l'approche par composition en séquence.

Un tel génome, *Ralstonia solanacearum*, une béta-protéobactérie au génome G+C riche, est étudié au LIPM, laboratoire partenaire de ce projet. La régulation protéique de cette bactérie est bien étudiée, notamment d'un point de vue de la virulence, mais l'ensemble des travaux existants fait l'impasse sur le rôle potentiel des ARN régulateurs. Plus largement, une recherche systématique des ARNnc n'a été entreprise chez aucune béta-protéobactérie au moment où cette thèse a démarré.

Dans la suite, nous présentons plus amplement *R. solanacearum*, l'organisme sur lequel seront appliqués les dévéloppements réalisés au cours de cette thèse.

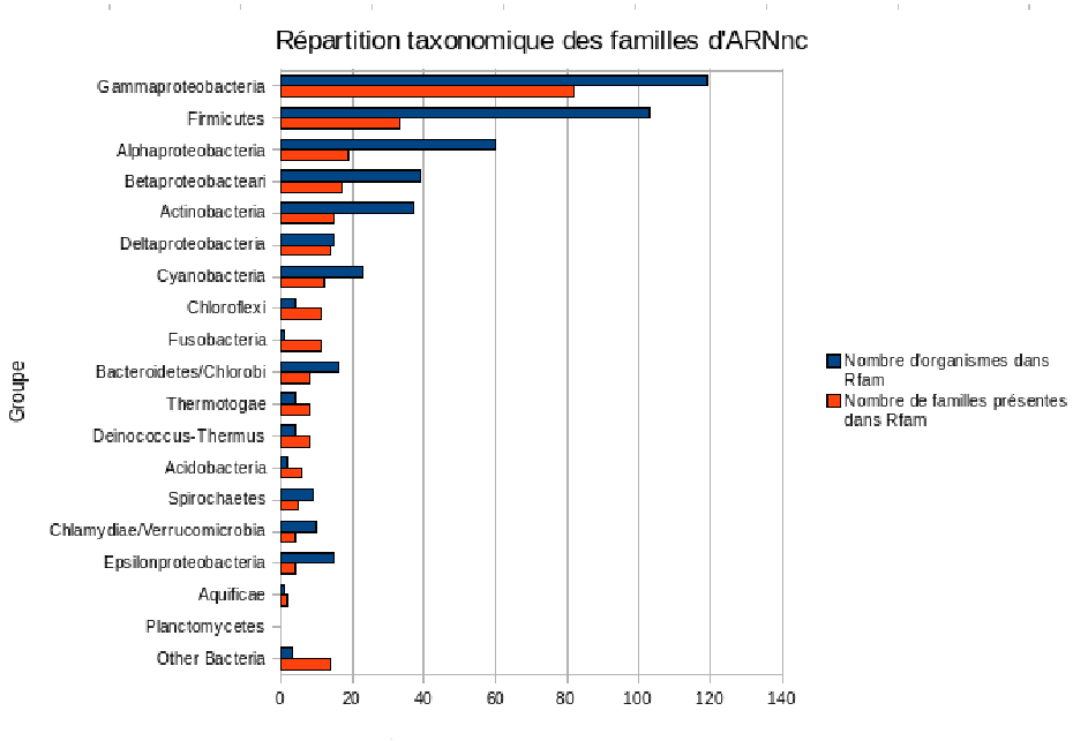


FIG. 1.11 – Le nombre de génomes présents dans Rfam et le nombre de familles d'ARNnc annotées, en fonction du groupe bactérien.

1.3 *Ralstonia solanacearum*, un organisme phytopathogène modèle

Ralstonia solanacearum est une bactérie Gram-négative, phytopathogène vivant dans le sol, et appartenant au groupe des béta-protéobactéries. Elle est responsable du flétrissement de plus de 200 espèces végétales réparties dans 50 familles botaniques. Cette bactérie a été choisie comme système modèle pour élucider les mécanismes moléculaires de la pathogénie mis en place par les bactéries au cours des interactions qu'elles entretiennnent avec les plantes. A plus long terme, son étude a pour objectif de contribuer à la la conception des nouvelles stratégies innovantes de lutte contre les agents pathogènes des plantes.

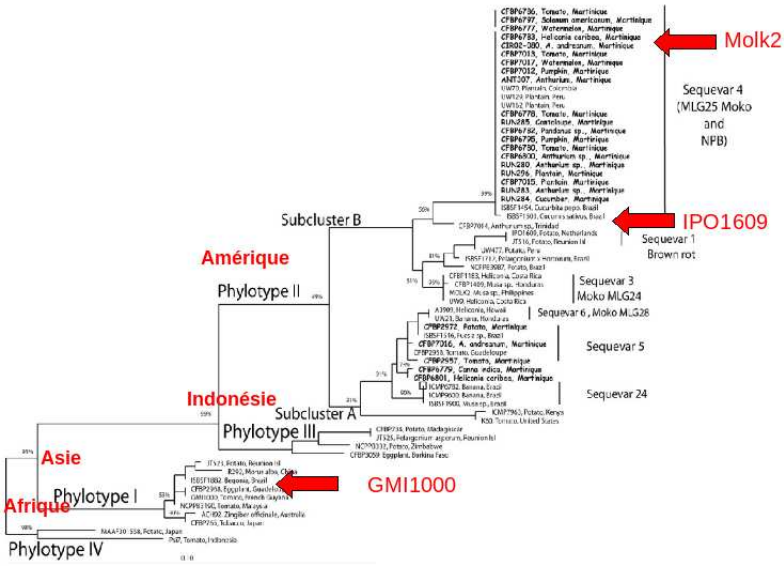


FIG. 1.12 – L'arbre phylogénétique de l'espèce *R. solanacearum*. Les emplacements des souches GMI1000, Molk2 et IPO1609 sont indiqués par les flêches rouges. La correspondence entre les clades et la localisation géographique est indiquée en rouge. L'arbre phylogénétique a été repris de Wicker *et al.* (2007). La souche Molk2 a été positionnée d'après Guidot *et al.* (2007).

L'analyse de la diversité au sein de l'espèce a permis d'établir l'existence de 4 grands clades. Ces clades sont très fortement corrélés avec la provenance géographique des souches de la bactérie, comme indiqué dans l'arbre phylogénétique de la figure 1.12. La souche GMI1000, appartenant au clade correspondant au phylotype I sur la figure,

a été entièrement séquencée par l'équipe de Christian Boucher et Stéphane Génin au LIPM, en collaboration avec le Génoscope. Cette souche est capable d'infecter une très large gamme de plantes hôtes, mais toutefois elle n'est pas pathogène sur le bananier. Deux autres souches, Molk2 et IPO1609 séquencées plus récemment appartiennent au phylotype II et se caractérisent par un spectre d'hôte étroit et spécialisé réciproquement sur le bananier et sur la pomme de terre.

1.3.1 Déterminants de la virulence connus chez *R. solanacearum*

De nombreuses études ont permis d'identifier différents facteurs impliqués dans la virulence de la bactérie. Les connaissances ainsi acquises sont compilées dans plusieurs revues dont la plus récente (Genin et Boucher (2004)) renvoie à l'ensemble de la littérature traitant cette question.

Plusieurs fonctions biologiques conférant la virulence à la bactérie ont été identifiées. Dans le processus précoce de l'invasion de la plante, la mobilité et l'attachement aux cellules végétales ont été identifiées comme une disposition importante pour la virulence. L'étape suivante, celle du franchissement de la barrière de la paroi végétale, fait intervenir des enzymes hydrolytiques de la dégradation de la paroi végétale, telles que des pectinases et des cellulases. La bactérie pénètre alors dans le parenchyme cortical, le tissu de la racine permettant le transport des matières absorbées de la périphérie au centre de la racine, où elle présente un fort tropisme vers les vaisseaux du xylème, le tissu conducteur de la sève brute. La bactérie envahit ces vaisseaux pour s'y multiplier abondamment et produire massivement des exoplysaccharides, qui produisent l'occlusion des vaisseaux du xylème et sont responsables du développement des symptômes caractéristiques du flétrissement qui conduisent à la mort de la plante.

Il a été montré qu'un déterminant absolument essentiel à la pathogénie est le système de sécrétion dit de type III codé par les gènes *hrp* ainsi que les effecteurs afférents. Cet appareil de sécrétion de protéines agit comme une seringue moléculaire qui permet l'injection des effecteurs bactériens dans la cellule végétale. Il a été démontré chez d'autres bactéries phytopathogènes que ces effecteurs agissent pour empêcher la mise en place des réactions de défense de la plante et permettre ainsi la multiplication bactérienne au sein des tissus végétaux (pour une revue voir Grant *et al.* (2006)). Une

analyse détaillée de la séquence de la souche GMI1000 a permis de montrer que cette souche produit probablement 74 effecteurs (Poueymiro et Genin (2009)).

1.3.2 Régulation de la virulence

Le contrôle de la virulence est effectué à travers un réseau de régulation sophistiqué (Schell (2000)). Nous presenterons ici brièvement le role du régulateur maître PhcA et la cascade de régulation spécifique du système de sécrétion de type III.

PhcA, un régulateur central

Au centre de ce réseau se trouve le régulateur central PhcA (figure 1.13) qui agit par mesure de la densité cellulaire ou *quorum sensing*, par l'intermediaire de la concentration en 3-hydroxypalmitate methyl ester (3-OH-PAME). Ce système agit à faible densité cellulaire pour induire la sécrétion de pectinase, la mobilité cellulaire et pour permettre l'expression du système de sécrétion de type III et des effecteurs afférents, facilitant ainsi la colonisation de la plante. A forte densité cellulaire ce système permet la production de cellulase et d'exopolysaccharides bactériens responsables de la production de symptômes de flétrissement (Flavier *et al.* (1997)).

Régulation du système de sécrétion de type III

Il a été établi que l'expression du système de sécrétion de type III et des effecteurs afférents est contrôlée par deux types de signaux : la reconnaissance du contact de la bactérie avec une cellule végétale et le statut métabolique de la bactérie (Aldon *et al.* (2000)). Cette régulation est sous la dépendance d'une cascade présentée en figure 3 qui fait intervenir les 4 activateurs transcriptionnels PrhI, PrhJ, HrpG et HrpB (figure 1.14).

La protéine PrhI agit en concertation avec PrhA et PrhR pour former un système de régulation à trois composants où PrhA agit comme senseur du contact avec la cellule végétale. L'intégration du signal issu du statut métabolique de la bactérie est intégré au niveau de l'activateur HrpG. Des analyses transcriptomiques ont permis de montrer que le gène *hrpB* a pour cibles principales les gènes codant pour les composants structuraux du système de sécrétion de type III ainsi que les effecteurs afférents (Occhialini *et al.* (2005)), tandis que l'activateur HrpG régule, en plus des fonctions HrpB-dépendant, un ensemble de fonctions d'adaptation à la vie dans la plante (Valls *et al.* (2006)).

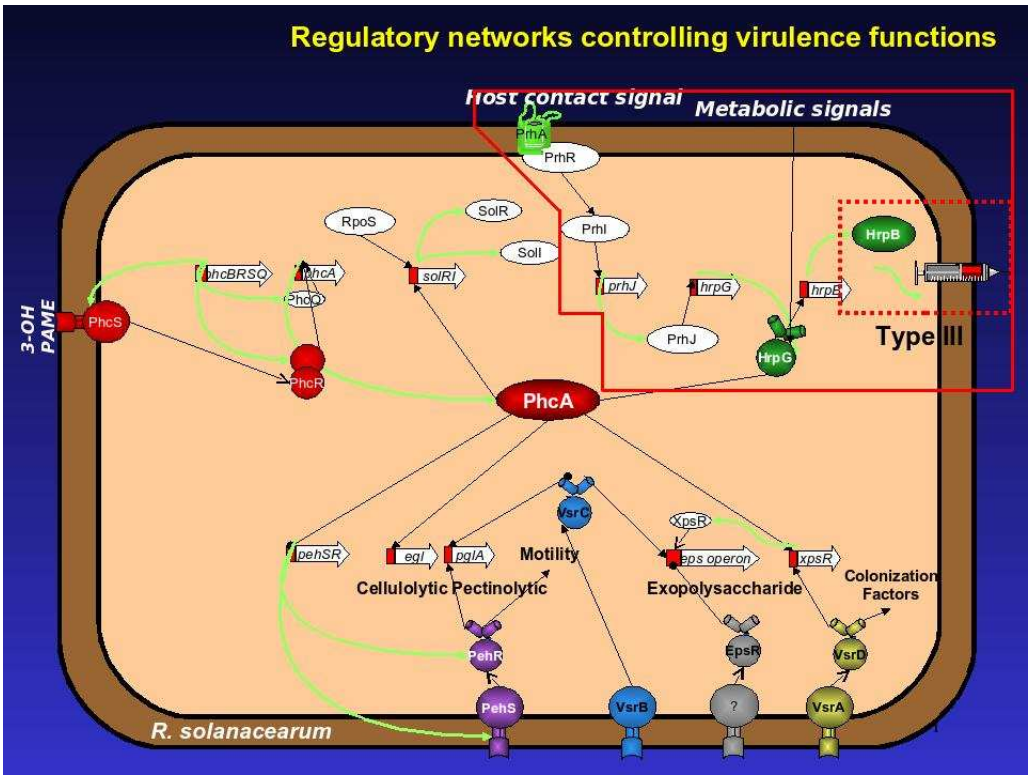


FIG. 1.13 – Modèle de la topologie de réseau de régulation des facteurs de virulence de *R. solanacearum*. Les effets régulateurs entre composants, directs ou non, et tels que spécifiés par les connecteurs en noir, correspondent à une situation où les sinaux environnementaux d'entrée du système sont présents en périphérie de la cellule bactérienne. Pour la plupart, on ignore leur nature. Les protéines représentées sous forme de "montgolfière à deux nacelles" sont des membres canoniques de la famille des régulateurs de réponse des systèmes de régulation bactériens à deux composants. Leur récepteur kinase correspondant est dessiné sur la membrane cytoplasmique avec une couleur identique. Repris de Cunnac (2004).

1.3.3 Absence d'études sur la régulation par ARN non-codant

Si la régulation de l'expression génique médiée par des protéines est bien documentée chez *R. solanacearum*, notamment en ce qui concerne la virulence, au contraire, le rôle des ARNnc dans la régulation chez cet organisme n'a pas été établie à ce jour, si ce n'est par un inventaire des ARNnc conservés en séquence dans la plupart des génomes bactériens et repertoriés dans Rfam (pour plus de détails voir la section 1.2). Le tableau 1.2 présente les ARNnc de *R. solanacearum* qui étaient répertoriés dans Rfam au début de cette thèse.

1.3.4 Structure et organisation du génome de *R. solanacearum* souche GMI1000

Le génome de *R. solanacearum* GMI1000 est organisé en deux réplicons de grande taille : un chromosome de 3.7Mb et le mégaplasmide de 2.1Mb (figure 1.15). Cette organisation est conservée dans l'ensemble des souches de *R. solanacearum* même si des

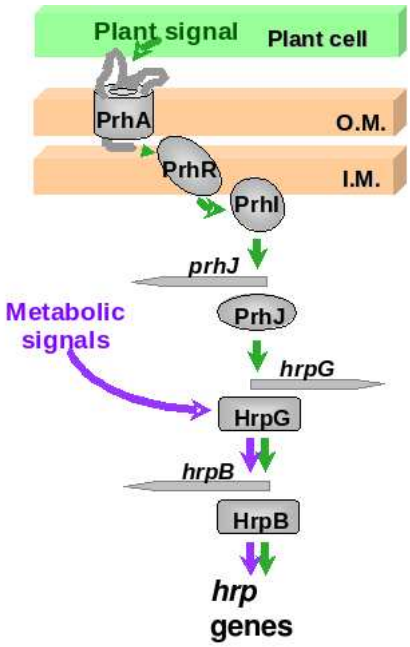


FIG. 1.14 – Modèle de régulation des gènes *hrp* dans *R. solanacearum*, pour plus de détails voir le texte. OM : membrane externe ; EM : membrane interne. Repris de Cunnac (2004)

Famille d'ARNnc	Nombre d'occurences
Chromosome	
ARNt	54
ARNr 16S	3
ARNr 23S	3
Glycine riboswitch	2
6S	1
Cobalamine riboswitch	1
yybP-ykoY riboswitch	2
TPP riboswitch	1
SRP	1
RnaseP	1
Nombre total	69
Nombre de différentes familles	10
Mégalasmide	
ARNt	3
ARNr 16S	1
ARNr 23S	1
FMN riboswitch	1
Cobalamine riboswitch	1
Nombre total	7
Nombre de différentes familles	5

TAB. 1.2 – Les familles d'ARNnc de *R. solanacearum* répertoriées dans Rfam.

plasmides supplémentaires de faible poids moléculaire sont parfois présents dans certaines souches (Genin et Boucher (2004) et C. Boucher, communication personnelle). Les deux réplicons ont un G+C% comparable et proche de 67% mais sont organisés en mosaïque, avec des fragments de taille variable, nommés ACUR (pour Alternative Codon Usage Regions), probablement acquis à la suite de transferts génétiques horizontaux et présentant une composition en bases très différente, alternant avec des régions issues d'un génome ancestral (Salanoubat *et al.* (2002)). Il a été montré plus récemment que ces dernières régions n'étaient pas uniformément distribuées au sein de l'espèce, renforçant ainsi l'hypothèse d'acquisition par transferts horizontaux.

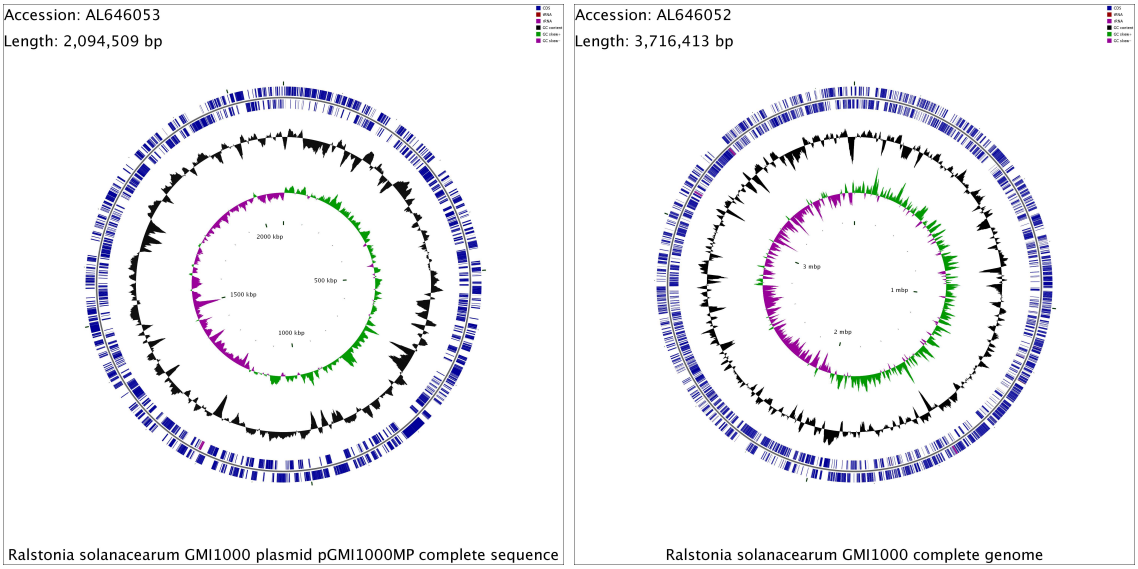


FIG. 1.15 – La présentation circulaire des deux réplicons de *R. solanacearum*. Code des couleurs : CDS (bleu), tRNA (bordeaux), rRNA (violet), G+C% (noir), GC skew + (vert), GC ckew- (rose). Repris de : http ://iant.toulouse.inra.fr/bacteria/annotation/cgi/ralso.cgi

Chapitre 2

Contexte bioinformatique de la détection des ARN non-codant

Ce travail de thèse porte sur le problème de la détection des ARNnc dans les génomes et dans ce chapitre nous présenterons trois grandes approches existantes pour aborder cette question. Certaines de ces approches seront rediscutées plus amplement dans les parties correspondantes de la thèse.

En préambule, nous décrirons succinctement les approches utilisées pour la prédiction des structures secondaires des ARNnc. Celles-ci sont d'une part integrées dans les différentes approches de prédiction des ARNnc et d'autre part, nous nous en sommes servis dans une partie de la thèse afin d'évaluer la qualité de nos prédictions.

2.1 Prédiction de la structure secondaire

Actuellement, aucune méthode théorique ne permet de prédire, de façon fiable, la structure secondaire d'une séquence d'ARNnc. En effet, les modèles sous-jacents à ces méthodes imposent un cadre simplifié au sein duquel un calcul efficace de la structure secondaire prédite devient possible, mais sacrifient ainsi la prise en compte de la totalité des aspects rentrant en jeu dans le repliement d'une séquence d'ARN (par exemple, la prise en compte de la formation des pseudo-noeuds n'est pas possible dans certaines de ces méthodes).

De nombreuses stratégies ont été développées pour essayer d'y apporter une solution. Ces stratégies peuvent être divisées en trois groupes : les approches thermodynamiques,

les approches comparatives et les approches hybrides, ces dernières combinant les deux premières approches. Les approches thermodynamiques recherchent la structure minimisant l'énergie libre selon un modèle d'énergie (le plus souvent le modèle thermodynamique, présenté dans la section 1.1.1). Quant aux approches comparatives, elles recherchent les mutations compensatoires dans les alignements des séquences d'ARNnc d'une famille. Un état de l'art sur les différentes méthodes de prédiction de structures secondaires d'ARN peut être trouvé dans Bindewald et Shapiro (2006).

Au cours de la thèse, nous avons utilisé RNAfold (Hofacker (2003)) qui prédit la strucure secondaire d'énergie minimale d'une séquence calculée à l'aide de l'algorithme de programmation dynamique proposée par Zuker et Stiegler (1981). RNAfold prédit la structure secondaire à partir d'une seule séquence.

Afin de prendre en compte l'information de plusieurs séquences dans la prédiction de la structure secondaire, nous avons aussi utilisé RNAz (Gruber *et al.* (2007), Washietl *et al.* (2005)), un outil hybride, tenant compte de la minimisation de l'énergie libre et de l'existence de mutations compensatoires entre les séquences étudiées. RNAz est présenté plus amplement dans la partie décrivant l'approche comparative, section 6.

2.2 Approches bioinformatiques pour la détection des ARNnc

La détection des ARNnc a pour but de localiser, dans un génome donné, ou plus généralement dans une séquence donnée, l'emplacement des séquences correspondant aux ARNnc. Contrairement aux séquences codantes, pour lesquelles la tâche analogue est aujourd'hui relativement bien maîtrisée, l'absence de cadres ouverts de lecture, de biais de codon ou d'un autre signal statistique commun à tous les ARNnc, fait que cette tâche demeure difficile dans le cas des ARNnc (Moulton (2005)).

Différentes stratégies bioinformatiques ont malgré tout été proposées. Elles peuvent être réparties dans trois grandes catégories, selon les informations utilisées (figure 2.1).

Ici nous présenterons brièvement le principe sur lequel chacune de ces approches est fondée, ainsi que les principaux outils permettant leur mise en oeuvre. Des détails complémentaires sur l'approche comparative pour la recherche des nouvelles familles et

l'approche *ab initio* peuvent être trouvés dans les introductions des chapitres correspondants (respectivement chapitre 6 et chapitre 3).

2.2.1 Recherche des membres d'une famille connue

Un premier objectif consiste à rechercher les ARNnc membres d'une famille déjà connue et caractérisée par sa structure secondaire.

Nous présenterons d'abord les méthodes mettant en place une recherche de motif correspondant à une structure secondaire à l'aide des algorithmes de recherche des motifs dans les textes et ensuite nous présenterons les méthodes de recherche des motifs implémentant des outils probabilistes. Dans cette présentation je m'appuyerai sur l'exposé à ce sujet dans la thèse de M. Zytnicki (Zytnicki (2007)).

Méthodes non-probabilistes Dans ces méthodes, la structure secondaire est décrite sous la forme d'un motif et celui-ci est recherché dans le texte génomique. Les principaux outils implémentant ce type d'approche sont RnaMot (Gautheret *et al.* (1993)), RNABOB (Eddy (1996)), Patscan (Ray *et al.* (1987)), RnaMotif (Macke *et al.* (2001)) et plus récemment MilPat (Thebault *et al.* (2006)), DARN ! (Zytnicki *et al.* (2008)) et Locomotif (Reeder *et al.* (2007)).

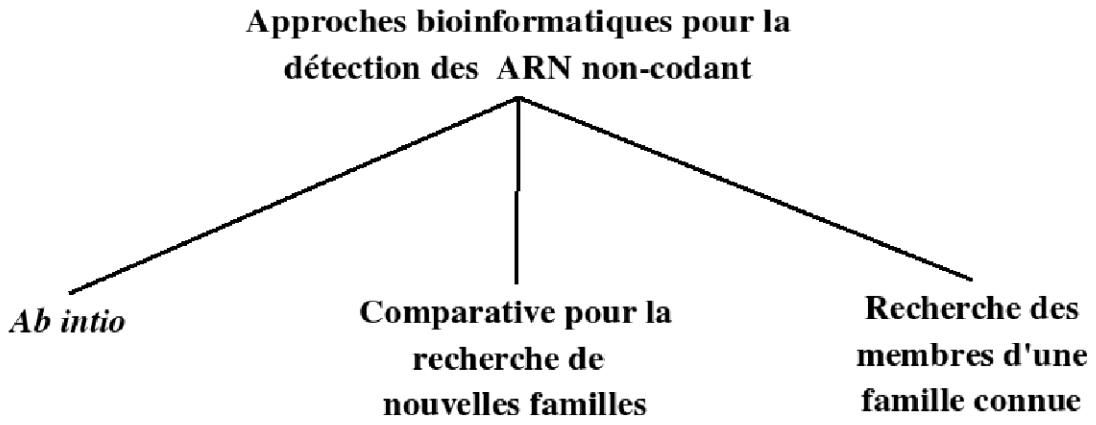


FIG. 2.1 – Trois grandes approches pour la détection bioinformatique des ARNnc.

Méthodes probabilistes Ces approches mettent généralement en oeuvre une modélisation par modèles de Markov cachés qui leur confère un aspect probabiliste et infèrent la présence d'un ARNnc appartenant à une famille donnée à partir des alignements multiples des séquences.

ERPIN (Lambert *et al.* (2004)), combine une méthode non-probabiliste décrivant les éléments de la structure secondaire et une méthode probabiliste analysant le contenu des ces éléments à l'aide de matrices position-spécifiques et de chaînes de Markov cachées.

RSEARCH (Klein et Eddy (2003)) et FastR (Zhang *et al.* (2005)) modélisent la structure secondaire d'une famille d'ARNnc par les SCFG (Stochastic Context-Free-Grammars) (Durbin *et al.* (1998b)), un formalisme permettant de définir un motif en terme de la séquence et d'appariements entre les différents nucléotides dans cette séquence. Une variante de ces modèles est également implementée dans le logiciel tRNAscan-SE (Lowe et Eddy (1997), largement employé pour la recherche des ARNt.

Notons que la base de données de référence recensant les membres des différentes familles d'ARNnc connues, Rfam (Griffiths-Jones *et al.* (2003)), est basée sur INFERNAL ("INFERence of RNA ALignment"), le modèle sous-jacent à RSEARCH.

2.2.2 Approches pour la recherche de nouvelles familles

Les méthodes décrites précedemment sont impuissantes lorsqu'il s'agit de rechercher des ARNnc appartenant à des familles nouvelles. Or, les familles d'ARNnc sont très diverses dans les organismes différents et seulement peu de familles sont universelement conservées (voir section 1.2). Ainsi, dans les groupes d'organismes où peu de recherches expérimentales d'ARNnc ont été menées l'utilisation des méthodes de recherche des familles connues est très limitée.

Etant donné que *R. solanacearum*, notre organisme d'intérêt appartient au phylum des béta-protéobactéries, peu étudié du point de vue des ARNn (voir section1.2), nous nous intéresserons plus particulierement aux méthodes permettant la découverte des nouvelles familles d'ARNnc. De telles méthodes peuvent être divisées en deux grandes catégories, approche *comparative* et approche *ab initio*.

Approche comparative L'approche comparative pour la recherche des ARNnc sans connaissance de leur famille utilise le fait que les ARNnc sont des régions fonctionnelles conservées, du moins entre génomes proches.

Les stratégies mettant en oeuvre la recherche des conservations en séquences dans les régions intergéniques combinées avec la recherche des différents signaux de transcription tels que les terminateurs ont été, à ce jour, les plus fructueuses et ont permis de détecter un grand nombre de nouveaux ARNnc dans les organismes tels que *E. coli* (pour revue voir Hershberg *et al.* (2003)), *Staphyococcus aureus* ou *Pseudomonas aeruginosa*. Une synthèse de ces travaux est présentée dans la section 9.

Des stratégies alternatives, basées sur le formalisme des grammaires formelles hors-contexte, prennent en compte, en plus de la conservation en séquence, la conservation en structure, par le biais de l'existence des mutations compensatoires. Ces méthodes ont été appliquées avec succès, notamment dans la détection des ARNnc chez *E. coli* (Rivas *et al.* (2001)). Nous pouvons citer le principal outil représentant cette catégorie, le logiciel QRNA (Rivas et Eddy (2001)) ainsi que RNAz, qui apporte une notion de stabilité en structure (Washietl *et al.* (2005)) et permet de travailler sur des alignements multiples. Une présentation plus ample de ces approches se trouve dans la section 6.

Approche ab initio Lorsque nous ne disposons pas de séquences génomiques apparentées au génome dans lequel nous voulons effectuer la recherche des ARNnc ou lorsque nous nous intéressons aux ARNnc spécifiques à un seul génome, une approche comparative ne peut pas être appliquée et seule la recherche *ab initio*, cherchant les signaux de présence des ARNnc contenus dans la séquence seule, est alors envisageable.

Une première approche consiste à rechercher des signaux de transcription orphelins, à savoir les signaux qui se trouvent dans les régions intergénique et qui ne sont pas associés aux gènes adjacents. De tels signaux sont essentiellement les promoteurs et terminateurs de transcription. Cependant, une telle méthode est alors limitée aux organismes dont les séquences promotrices sont bien connues (essentiellement, le génome d'*E. coli*). Une telle étude a été effectuée par Chen *et al.* (2002).

Une autre approche se base sur l'hypothèse de l'existence d'un biais de composition dans les séquences d'ARNnc et recherche, dans le génome, les zones présentant ce biais. Cependant, à ce jour, l'existence d'un tel biais à été clairement démontrée seulement dans les génomes des archées A+T riches hyperthermophiles où la stabilité de la structure des ARNnc est probablement maintenue par un contenu en nucléotides G et C plus élevé (Klein *et al.* (2002), Schattner (2002)). Quant aux ARNnc des génomes bactériens A+T riches, certains travaux indiquent que leur composition en G+C est plus élevée et

qu'ils peuvent être recherchés sur cette base (Upadhyay *et al.* (2005), Pichon et Felden (2005)). En revanche, l'ensemble des travaux existants font l'impasse sur la composition des ARNnc dans les génomes G+C riches.

Les ARNnc sont, pour la plupart, des molécules structurées ce qui indique que leur énergie libre est plus basse que celle d'une séquence non-structurée. Même si ce fait semble être confirmé *in vivo* (Massé *et al.* (2003), Vogel *et al.* (2003)), la pertinence de ce critère seul, calculé à partir du modèle thermodynamique a été remise en question (d'abord partiellement par Workman et Krogh (1999) et ensuite par Rivas et Eddy (2000)). Enfin, Clote *et al.* (2005) montrent, sur un ensemble constitué essentiellement des ARNt, que les séquences structurés ont une énergie libre plus basse que les séquences aléatoires ayant la même composition en di-nucléotides. Partant de l'hypothèse que la stabilité en structure est un élément à prendre en cosidération dans la recherche des ARNnc, un troisième type d'approches, combinant les approches comparatives avec la recherche des zones du génome à l'énergie libre basse a été proposé. Cette approche est à la base de des outils comme RNAz (Washietl *et al.* (2005)) et MSARi (Coventry *et al.* (2004)).

Etant donné que cette thèse traite, entre autres, le biais de composition des ARNnc, les travaux correspondants sont décrits plus amplement dans le chapitre 3.

Deuxième partie

Approche *ab initio* pour la caractérisation des ARNnc

Introduction

Dans ce chapitre nous examinons l'existence d'un biais de composition au sein des ARNnc de *Ralstonia solanacearum* et la pertinence de son utilisation pour leur détection.

Nous présentons d'abord un état de l'art sur l'approche *ab initio* pour la détection des ARNnc à travers des travaux existants.

Nous proposons une méthodologie statistique pour mettre en évidence la différence en composition entre les ARNnc et le reste du génome sur la base de leur contenu en G+C. Nous exposons son application à deux génomes différents du point de vue du contenu en G+C, un génome G+C pauvre, *Staphylococcus aureus* et le génome G+C riche de *Ralstonia solanacearum*. Ensuite, nous présentons la modélisation markovienne des séquences et nous l'appliquons à la modélisation de la composition en nucléotides et en di-nucléotides du génome de *R. solanacearum* afin de mettre en évidence la différence en composition entre les ARNnc et le reste du génome.

Une fois l'existence d'un biais en composition au sein des ARNnc établie, nous l'utilisons pour segmenter le génome à l'aide des modèles de Markov cachés. Ces modèles ainsi que la modélisation des séquences génomiques par ces modèles sont présentés. Une étude de deux algorithmes de segmentation à l'aide de ces modèles (Viterbi et Forward-Backward) est aussi proposée et elle nous amenera à conclure que la pertinence de l'algorithme de Viterbi est hautement des paramètres des modèles utilisés et que de ce fait, il doit être utilisé avec précaution.

Enfin, la segmentation effectuée, nous présentons les résultats obtenus et nous évaluons leur spécificité et sensibilité.

Chapitre 3

Utilisation de biais en composition pour la détection des ARNnc : état de l'art

La notion de biais en composition se réfère à la différence de composition en nucléotides entre différents éléments génomiques existant dans un ou plusieurs génomes. L'existence d'un biais en composition a pu être mis en évidence dans un grand nombre d'éléments génomiques, comme par exemple, le biais présent sur les différents brins d'ADN (Lobry (1996)), le biais au niveau de l'origine de réplication dans les génomes bactériens (Rocha *et al.* (1999)) ou encore le biais de codons (McInerney (1998)). L'existence d'un biais de codon est à l'origine de nombreux programmes de prédiction des gènes (pour une revue, dans les génomes bactériens, voir Azad et Borodovsky (2004)).

Ici, nous nous intéressons au biais de composition lié aux séquences des ARNnc.

3.1 Biais de composition dans les ARNnc

L'existence d'un biais de composition dans les séquences d'ARNnc a été observée pour la première fois par Rivas et Eddy (2000). Ils constatent que dans le génome de *Methanocaldococcus jannaschii* (ancien *Methanococcus janachii*), une archébactérie A+T riche hyperthermophile, les ARNnc sont caratérisés par le G+C% très élevé par rapport au G+C% du reste du génome (31.4% pour les ARNnc et 63.1% pour le génome).

Ce constat a été exploité dans les études de Klein *et al.* (2002) et Schattner (2002) pour prédire de nouveaux ARNnc chez *M. jannachii*.

Dans Schattner (2002), les auteurs examinent également la possibilité d'utiliser le biais de composition dans un autre organisme A+T riche, la bactérie *Plasmodium falciparum* (le G+C% génomique moyen autour de 20%), avec la conclusion est que le G+C% des ARNnc de ce génome n'est pas assez discriminant.

D'autre part, Klein *et al.* (2002) utilisent le biais de composition pour détecter des ARNnc avec succès dans le génome de *Pyrococcus furiosus*, une autre archébactérie A+T riche hyperthermophile (G+C% génomique moyen 40.8%) .

A ce jour, les seuls organismes où la différence en composition entre les ARNnc et le reste du génome a été utilisée avec succès comme seul critère de prédiction sont les archébactéries hyperthermophiles *M. jannaschii* et *P. furiosus*. Les appariements G-C étant plus stables que les appariements A-T, le fort biais en G+C au sein des ARNnc de ces organismes est supposé être lié à la stabilisation des structures secondaires dans les conditions de haute température (Klein *et al.* (2002), Galtier et Lobry (1997)). Des indications existent (Das *et al.* (2006)) qu'un biais en compostion des ARNnc existerait dans d'autres génomes archéens A+T riches hyperthermophiles, comme le montre le tableau 3.1. Néanmoins, des études spécifiques, permettant la découverte de nouveaux ARNnc en exploitant le biais de composition des ARNnc n'ont pas été menées.

	G+C%			
Organisme	génomique	ORF	ARNr	ARNt
Aquifex aeolicus	43.3	43.7	64.8	68.4
Methanocaldococcus jannaschii	31.3	32.0	63.8	66.5
Nanoarchaeum equitans	31.5	31.2	66.2	72.5
Pyrococcus abyssi	44.6	45.2	65.9	70.5
Pyrococcus furiosus	40.7	41.2	66.3	70.5
Pyrococcus horikoshii	41.8	42.3	63.2	70.7
Sulfolobus solfataricus	35.7	36.5	62.1	67.3
Sulfolobus tokodaii	32.7	33.6	63.8	67.4

TAB. 3.1 – La composition génomique, des ORF, des ARNr et des ARNt dans les génomes archéens A+T riches hyperthermophiles. Tableau partiel repris de Das *et al.* (2006).

Dans certains génomes bactériens A+T riches, l'utilisation du biais de composition s'est révélée comme un critère utile en combinaison avec d'autres approches. Ainsi, Upadhyay *et al.* (2005) effectuent une étude de recherche d'ARNnc chez *P. falciparium*

à l'aide de biais en composition en renforçant les prédictions par une approche comparative. Cette étude a permis de retrouver de nouveaux ARNnc chez *P.falciparium*. L'approche par biais de composition complétée par une approche comparative a également été employée par Pichon et Felden (2005) permettant la mise en évidence de nouveaux ARNnc dans *Staphylococcus aureus* (G+C% génomique moyen de 33%).

Quant aux génomes qui ne sont pas A+T riches, l'analyse du biais de composition combinée avec le calcul de l'énergie libre ont été utilisés dans une recherche systématique des ARNnc d'*Escherichia coli* (de G+C% génomique moyen de 50.8%) (Carter *et al.* (2001)). L'existence d'un biais de composition au sein des ARNnc n'a pas été mis en évidence, d'autant plus que les prédictions issues de cette étude n'ont pas été validées à ce jour.

Ainsi, l'ensemble des travaux utilisant le biais en composition pour la recherche des ARNnc portent sur les organismes A+T riches, à l'exception de Carter *et al.* (2001) dont les résultats n'ont pas été validés. En revanche, il n'existe pas d'étude portant sur la composition des ARNnc dans les organismes G+C riches, l'autre extrême du point de vue de le composition en nucléotides. Dans les développements de la présente partie, nous nous proposons d'étudier la composition des ARNnc d'un tel organisme, *R. solanacearum* (G+C% génomique moyen 67%) et nous explorons la possibilité d'utiliser une approche par biais en composition pour la recherche des ARNnc.

3.2 Utilisation du biais de composition pour détecter des ARNnc : méthodologies employées

Lorsque nous nous intéressons à l'utilisation du biais de composition pour détecter des ARNnc deux problèmes disctincts peuvent être identifiés :

1. Comment mettre en évidence le biais de composition entre les ARNnc et le reste du génome ?

2. Comment localiser dans le génome les segments présentant un biais donné ?

Les approches utilisées pour répondre à ces deux questions, dans le cadre de la détection des ARNnc, seront présentées dans la suite.

3.2.1 Mettre en évidence le biais en composition

Dans les études évoquées précedamment l'existence d'un biais de composition au sein des ARNnc par arpport au reste du génome était sous-entendu la plupart du temps.

La seule étude proposant une telle démarche est Schattner (2002) qui utilise le test de Student pour décider si la différence entre le G+C% dans les différents groupes est significative. Le premier groupe est constitué des ARNnc connus et le deuxième groupe de séquences de 100nt échantillonnées dans le génome de façon aléatoire.

3.2.2 Utiliser le biais en composition pour segmenter le génome

Pour sélectionner les régions d'un génome présentant un biais de composition correspondant aux ARNnc trois plusieurs approches ont été utilisées.

Fenêtres glissantes Pour discriminer les segments dans le génome présentant un biais en composition, Schattner (2002) et Upadhyay *et al.* (2005) utilisent des fenêtres glissantes.

Cette approche est peu compliquée et rapide. Néanmoins, il s'agit d'une méthode manquant de fondement théorique permettant d'évaluer la significativité statistique des résultats obtenus. Un autre inconvénient de cette approche est que les frontières entre les segments de composition différente ne sont pas déterminées avec précision.

Modèles de Markov cachés Klein *et al.* (2002) considèrent le G+C% comme critère de biais de composition et proposent un modèle de Makov caché (HMM, pour la définition voir la section 5.1) à deux états : un état G+C riche modélisant la séquences des ARNnc et un état A+T riche modélisant le reste de génome. Cette démarche s'inscrit dans un cadre théorique de modélisation des séquences génomiques et permet une éventuelle complexification du modèle dont le but serait une modélisation plus « réaliste » de la séquence génomique.

Au cours de ce travail de thèse, une généralisation des modèles de Markov cachés a été utilisée dans Tjaden (2007). Ce travail présente un modèle d'intégration de données hétérogènes tenant compte du biais de composition, des données d'expression et de l'information sur la conservation en structure déduite de l'analyse comparative. La composante du modèle servant à la détection des régions présentant un biais de compo-

sition caractéristique des ARNnc est un modèle de Markov caché ayant 9 états cachés correspondants, entre autres, aux promoteurs et aux terminateurs adjacents à l'ARN putatif. La performance de ce modèle a été testé sur l'ensemble des ARNnc connus d'*E. coli* et l'efficacité du modèle de biais de composition seul, sans tenir compte des données d'expression et de al conservation, n'a pas pu être démontrée pour ce génome.

Critère visuel Dans Pichon et Felden (2005) les régions G+C riches ont été déterminées à l'aide d'un outil de visualisation, ce qui constitue une démarcje peu fiable et peu efficace.

Chapitre 4

Tester la difference en compostion

La composition des séquences génomiques peut être étudiée à plusieurs niveaux différents :

1. en G+C%, où deux paramètres sont considérés : le taux des nucléotides G ou C (noté G+C%) et le taux des nucléotides A ou T (noté A+T%)
2. en nucléotides A, C, G et T où quatre paramètres sont considérés : les taux de A, C, G et T
3. en di-nucleotides AA, AC, ..., TT où 16 paramètres sont considérés
4. les n-nucléotides où la fréquence d'occurence de différents mots de longueur n à partir de l'alphabet $\{A, C, G, T\}$ est considérée

La façon la plus concise de décrire la composition d'une séquence génomique est son G+C% : il résume par un paramètre la composition de la séquence, le deuxième paramètre A+T% se déduisant comme A+T%= 100−G+C%. Par ailleurs, la composition des génomes est traditionnellement décrite par la valeur de son G+C% ; en effet, avant les séquençages massifs où les séquences des génomes n'étaient pas disponibles, le G+C% pouvait être calculé à l'aide de la température nécessaire pour dénaturer les deux brins d'ADN (Vinogradov (1994)).

Dans le présent chapitre nous abordons la question de la mise en évidence de l'existence d'un biais de composition au sein des séquences d'ARNnc avec application au génome de *Ralstonia solanacearum*. Dans un premier temps, nous étudions le G+C%, à l'aide de la théorie des modèles linéaires généralisés (GLM). Dans un deuxième temps, une modélisation markovienne des séquences est proposée pour modéliser la composition

en nucléotides et en di-nucléotides.

4.1 Utilisation du G+C% pour caractériser les ARNnc

Lorsque nous voulons conclure si le G+C% diffère entre plusieurs groupes de séquences (par exemple : les séquences d'ARN, le codant et le reste du génome) il ne suffit pas de comparer les valeurs moyennes des G+C% dans ces groupes. En effet, la question est de savoir à partir de quel seuil de différence de G+C% entre deux groupes cette différence est significative et non pas le résultat d'une fluctuation aléatoire due au choix de l'échantillon.

Les travaux cités, portant sur la composition des ARNnc (section 3.2), n'effectuent pas, pour la plupart, de test statistique démontrant la différence de composition entre les séquences des ARNnc et le reste du génome, et sont essentiellement basés sur l'observation empirique. Dans Schattner (2002) le test de Student des moyennes est utilisé. Plusieurs critiques peuvent être adressées à cette approche. Tout d'abord, pour utiliser le test de Student, les observations des populations dont les moyennes sont testées doivent être tirées d'une seule distribution qui doit être normale. Or, les valeurs de G+C% sont comprises entre 0 et 1 et ne sont donc pas distribuées normalement. De plus, dans la population d'ARNnc les longueurs des séquences sont variables, ce qui induit des variances différentes des G+C% à l'intérieur d'un même groupe. Ce point n'est pas pris en compte dans l'approche proposée. Un dernier point concernant cette solution est qu'elle néglige l'éventuelle composition non-homogène au sein des différents groupes.

Dans le contexte plus général d'étude de composition des génomes, l'homogénéité a été étudiée par Karlin *et al.* (1994), Karlin *et al.* (1998) ou Li *et al.* (1998). Toutefois, ces travaux ne donnent pas une méthode directe pour répondre à notre question d'intérêt.

Une solution consiste à utiliser la théorie des modèles linéaires généralisés (GLM, McCullagh et Nelder (1989a)). Elle nous a été suggérée par Sophie Schbath de l'unité MIG et par Stéphane Robin d'AgroParisTech. Afin d'étudier le G+C%, on modélise chaque nucléotide dans la séquence par une variable binaire, valant 1 si le nucléotide est un G ou un C et 0 sinon. Selon cette modélisation le G+C% est identique à la probabilité d'observer un 1 dans la séquence. Il s'agit d'un cadre formel bien étudié permettant de répondre à la question posée d'une manière satisfaisante : nous ne faisons

pas l'hypothèse de la distribution normale des observations et la différence de longueur de séquences est prise en compte par le modèle. De plus, ce cadre permet une analyse détaillée de la structure de chacun des groupes et l'identification d'éventuelles séquences atypiques.

Dans la suite nous présenterons d'abord le modèle de regresssion logistique, un modèle de la classe des GLM, ensemble avec les outils d'analyse de données qui lui sont associés. Enuite nous présenterons les résultats de son application sur les données génomiques.

4.1.1 Modèle de regression logistique

4.1.1.1 Hypothèses et définition du modèle

Supposons que nous disposons de N observations $Y_1, \ldots, Y_N$ suivant une loi binomiale de paramètres (n_i, p_i), $Y_i \sim \mathcal{B}(n_i, p_i)$. Rappelons que la loi binomiale de paramètres (n_i, p_i) correspond à l'expérience où une épreuve de Bernoulli de paramètre p_i est renouvelée n_i fois de manière indépendante et dont les deux issues possibles sont dénommées *succès* et *échec*. La variable aléatoire Y_i correspond au nombre de succès à l'issu de n_i épreuves. Nous avons que l'espérance et la variance de Y_i s'écrivent comme :

$$E(Y_i) = n_i p_i \tag{4.1}$$

$$Var(Y_i) = n_i p_i (1 - p_i) \tag{4.2}$$

Le modèle de régression logistique est un modèle de la classe des modèles linéaires généralisés (décrits dans l'ouvrage de référence McCullagh et Nelder (1989a)). Il décrit la relation linéaire supposée entre la variable *réponse* Y_i et K variables $x_1, \ldots, x_K$, dites *explicatives*, $x_j, 1 \leq j \leq K$, pouvant prendre des valeurs quantitatives ou qualitatives, servant à expliquer la variable réponse Y_i.

Les hypothèses du modèle de régression logistique sont :

1. $Y_1, \ldots, Y_N$ sont les variables aléatoires indépendantes
2. Pour tout i, $1 \leq i \leq N$, la variable Y_i suit une loi binomiale de paramètres (n_i, p_i)

Notons x_j^i est la valeur de la variable x_j pour l'individu i, $1 \leq i \leq N$ et $1 \leq j \leq K$. Le modèle est fondé sur l'existence d'une fonction f, que l'on appelle la *fonction de lien*,

$f :]0,1[\rightarrow \mathbf{R}$ telle que :

$$f(p_i) = \mu + \sum_{j=1}^{K} \beta_j x_j^i \tag{4.3}$$

μ est appelé *l'intercept* et le β_j sont *les coefficients de la régression*. Nous noterons β le vecteur des paramètres du modèle, $\beta = (\mu, \beta_1, \ldots, \beta_K)$.

Dans le cadre général des modèles linéaires généralisés, différentes fonctions de la famille exponentielle peuvent être utilisées comme fonction de lien. Dans le cas de la régression logistique, que nous exposons ici, la fonction *logit* est utilisée. Pour $p \in]0,1[$, nous avons :

$$\eta = logit(p) = \log(\frac{p}{1-p}) \tag{4.4}$$

Ainsi, dans le modèle de régression logistique, le lien suivant existe entre les paramètres p_i, $1 \leq i \leq N$ et les coefficients de la régression :

$$p_i(\beta) = \frac{\exp(\mu + \sum_{j=1}^{K} \beta_j x_j^i)}{1 + \exp(\mu + \sum_{j=1}^{K} \beta_j x_j^i)} \tag{4.5}$$

Sous le modèle logistique, sous l'hypothèse de l'indépendance des tirages et d'une distribution binomiale, la vraisemblance sachant les données s'écrit comme :

$$L(\beta) = \prod_{i=1}^{N} \binom{n_i}{y_i} p_i(\beta)^{y_i} (1 - p_i(\beta)^{n_i - y_i}) \tag{4.6}$$

La relation ci-dessus peut être réécrite sous la forme de log-vraisemblance :

$$\begin{aligned} \log L(\beta) &= \sum_{i=1}^{N} \big(\log \binom{n_i}{y_i} + y_i \log p_i(\beta) + (n_i - y_i) \log(1 - p_i(\beta)) \big) \\ &= \sum_{i=1}^{N} \big(\log \binom{n_i}{y_i} + y_i \log \frac{p_i(\beta)}{1 - p_i(\beta)} + n_i \log(1 - p_i(\beta)) \big) \\ &= \sum_{i=1}^{N} \big(\log \binom{n_i}{y_i} + y_i \eta_i(\beta) - n_i \log(1 + \exp(\eta_i(\beta))) \end{aligned} \tag{4.7}$$

Régression logistique dans le cas spécial des variables nominales

Une variable nominale décrit un nom ou une catégorie (par exemple, l' appartenance à un groupe) et entre ses valeurs possibles, appelées *les modalités*, n'existe pas un ordre naturel. Exprimer de telles variables par une valeur numérique n'a pas de sens et il est d'usage de les transformer en variables indicatrices d'appartenance à un groupe à l'aide

Groupe	x_1	x_2	x_3	...	x_{M-1}
0	0	0	0	...	0
1	1	0	0	...	0
2	0	1	0	...	0
...					
$M-1$	0	0	0	...	1

TAB. 4.1 – Codage disjonctif complet pour une variable nominale à M modalités.

de de la transformation appelé *codage disjonctif complet* (Bouyer *et al.* (1995)). Dans ce codage, une variable X à M modalités, numérotées de 0 à $M-1$, est transformée en $M-1$ variables, la m-ième variable, $1 \leq m \leq M-1$, prenant la valeur 1 ou 0 selon que la valeur de X correspond ou pas à la m-ième modalité (tableau 4.1). Nous appelerons *la modalité de référence* la modalité pour laquelle toutes les variables explicatives valent 0 (le groupe 0 dans le tableau 4.1). Dans le cas d'une variable explicative nominale, le modèle de régression logistique s'écrit comme :

$$logit(p_i) = \mu + \sum_{j=1}^{K} \beta_j x_j^i \tag{4.8}$$

, où $K = M-1$ et x_j^i est la valeur de la variable explicative x_j obtenue par le codage disjonctif complet pour l'individu i. La même expression peut s'écrire aussi sous la forme :

$$logit(p_i) = \begin{cases} \mu & j = 0 \\ \mu + \beta_j & j \in \{1, \ldots, K\} \end{cases} \tag{4.9}$$

faisant apparaître le lien entre les paramètres de la régression logistique et $logit(p_i)$ associés à chacune des modalités de la variable explicative.

4.1.1.2 Estimation des paramètres

Les paramètres d'un modèle linéaire généralisé peuvent être estimés par le maximum de vraisemblance. Le zéro de la dérivée de la fonction de la log-vraisemblance (équation 4.7) par rapport au j-ième paramètre donne β_j maximisant la vraisemblance :

$$\frac{\mathrm{d} \log L(\beta)}{\mathrm{d}\beta_j} = \sum_{i=1}^{N} y_i x_j^i - \sum_{i=1}^{N} n_i x_j^i \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \tag{4.10}$$

Ainsi, les β_j annulant les équations (4.10) sont la solution d'un système de K équations non-linéaires. Ce système n'ayant pas une solution analytique, une solution numérique, implementée par *la méthode des moindres carrés pondérés itérés*, est utilisée.

Le prédicteur linéaire du modèle pour la i-ième observation peut être écrit à partir des paramètres $\hat{\beta}$ estimés comme :

$$\hat{\eta}_i = \hat{\mu} + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_K x_{Ki} \tag{4.11}$$

4.1.1.3 Analyse des résidus

Une fois le modèle défini et les paramètres associés estimés, il est important de vérifier la validité de ce modèle au regard des données. A condition que le modèle utilisé soit adapté au phénomène étudié, l'inadéquation entre le modèle estimé et les données peut exister, pour différentes raisons. Une raison peut être l'existence d'observations isolées qui ne sont pas en adéquation avec le modèle proposé. L'identification et puis l'analyse de ces observations peut nous amener à identifier d'éventuelles erreurs dans le plan d'expérience et dans les données. Dans ce cas, ces erreurs peuvent être corrigées ou, le cas échéant, ces observation retirées de l'analyse. Une autre raison peut être une inadéquation "intrinsèque" des données à la modélisation binomiale classique. Dans ce cas des solutions peuvent être apportées *via* une modélisation par un modèle "quasi-binomial".

Nous présenterons d'abord les outils de diagnostic des points isolés inadéquats avec les données et ensuite nous présenterons la façon de prendre en compte, de manière systématique, l'inadéquation au modèle binomial classique. Pour ce faire, nous aurons besoin de quelques notions que nous introduisons dans la suite.

La matrice d'incidence X est une matrice de dimension $N \times K$, dont l'élément (i, j) est la valeur de la j-ième variable explicative de la i-ième observation.

La matrice des poids W est une matrice diagonale de dimension, $N \times N$, fonction des paramètres β, telle que son élément (i, i), noté w_i soit $w_i = 1/Var(Y_i)$.

La matrice H, appelée *la hat matrice*, est définie, à partir des matrices X et W, comme :

$$\boldsymbol{H = W^{\frac{1}{2}} X (X' W X)^{-1} X' W^{\frac{1}{2}}} \tag{4.12}$$

Nous avons que $HY = \hat{Y}$ (d'où le nom de *hat matrice*) et les valeurs de H interviennent dans le calcul de l'importance d'une observation dans l'estimation. En particulier, les

éléments de la diagonale de la matrice H, h_i reflètent l'influence potentielle de l'observation Y_i sur l'estimation des paramètres du modèle.

Le résidu est la partie de l'observation que le modèle n'explique pas et il est calculé comme une fonction de l'écart entre la valeur observée de la variable réponse et sa valeur estimée.

On défini les *résidus bruts* comme la différence entre la valeur observée et la valeur estimée de la variable réponse :

$$r_i = y_i - n_i\hat{p}_i \tag{4.13}$$

Les résidus bruts sont difficilement comparables car le nombre de tirages, n_i, différent pour les différentes observations, influence la variance. Les résidus peuvent être rendus comparables en divisant les résidus bruts par l'écart type de la variable réponse. Ces résidus sont appelés les *résidus de Pearson* :

$$r_{S_i} = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}} \tag{4.14}$$

La somme de leurs carrés donne la statistique X^2 de Pearson :

$$X^2 = \sum_{i=1}^{N} r_{S_i}^2 \tag{4.15}$$

Néanmoins, les résidus de Pearson ne sont pas normalisés. Pour obtenir les résidus normalisés, les résidus sont divisés par leur écart-type, et on peut montrer que $\sigma(y_i - \hat{y}_i) = \sqrt{n_i\hat{p}_i(1-\hat{p}_i)(1-h_i)}$, où h_i est l'i-ième élément de la diagonale de la matrice H. En normalisant les résidus bruts nous obtenons *les résidus de Pearson standardisés* :

$$r_{P_i} = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)(1-h_i)}} \tag{4.16}$$

Lorsque le modèle est bien ajusté, les résidus de Pearson standardisés suivent une loi normale $N(0,1)$ (Pregibon (1981)). Dans ce cas, les observations sont situées approximativement dans l'intervalle $(-1.96, 1.96)$, l'intervalle de confiance à 95%. Cette propriété peut être utilisée dans le diagnostic de la structure des données.

Avant d'introduire une mesure d'adéquation du modèle aux données, rappelons que *le modèle saturé* est le modèle qui correspond "parfaitement" aux données, à savoir le modèle qui contient autant de paramètres que d'observations. Ici, le modèle saturé est défini par :

$$\hat{p}_i = \frac{y_i}{n_i} \tag{4.17}$$

pour $1 \leq i \leq N$.

Le modèle saturé sert ici pour définir *la déviance*, qui mesure l'écart ou la "déviation" du modèle par rapport aux données (Collett (1991)). La déviance est définie comme la statistique du rapport de vraisemblance entre le modèle proposé et le modèle saturé. Si on note L_S la valeur de maximum de vraisemblance pour le modèle saturé et L_M la valeur de maximum de vraisemblance pour le modèle ajusté ($L_M = L(\hat{\beta})$) alors on définit la déviance comme :

$$Dev = -2\log(\frac{L_M}{L_S}) \tag{4.18}$$

Elle peut être réécrite sous la forme (Collett (1991)) :

$$Dev = 2\sum_{i=1}^{N} \left(y_i \log(\frac{y_i}{\hat{y}_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \hat{y}_i}\right) \tag{4.19}$$

qui fait apparaître la comparaison des valeurs de réponses observées avec les valeurs prédites.

Une fois le modèle défini, certaines observations ont des propriétés particulières. On peut les classer en trois catégories (non exclusives entre elles) :

1. les points leviers
2. les points aberrants (ou points extérieurs)
3. les points influents

Les points leviers sont des points ayant potentiellement une grande influence sur l'estimation du fait des valeurs de leurs variables explicatives (sans considérer la valeur de la variable réponse). La recherche des points leviers est associée à la notion de fonction de *l'effet de levier* ou *leverage*, en anglais. Dans le cadre du modèle linéaire, le *leverage* est une fonction des variables explicatives qui mesure le potentiel de chaque donnée d'affecter l'estimation du modèle (figure 4.1). Dans ce cadre, le leverage d'un point x_i est défini comme la distance entre x_i et le centroïde $\overline{x} = \frac{1}{n}\sum_{i=1}^{N} x_i$ et dépend donc uniquement du plan d'expérience.

Dans le cas du modèle linéaire généralisé (et de la régression binomiale en particulier), la définition du leverage devient plus compliquée pour les raisons de différences de variance entre les différentes observations. En conséquence, l'interprétation du leverage devient moins évidente. Ici, nous pouvons nous contenter de dire que le leverage représente la distance entre un point et le reste des points en tenant compte du poids de

chaque point. En pratique, le leverage d'une observation i est défini comme h_i, le i-ième élément de la diagonale de la matrice H (équation 4.12). On considère que l'observation i est élevée en leverage si $h_i > \frac{2K}{N}$ (McCullagh et Nelder (1989b)).

Alors que la notion des points leviers porte sur la valeur des variables explicatives, *les points aberrants* sont particuliers de part leur réponse, qui n'est pas en adéquation avec la valeur prédite par le modèle. Cette inadéquation avec le modèle est reflétée par les résidus et les points sont considérés comme aberrants par rapport au modèle choisi lorsque leurs résidus sont élevés (Collett (1991)).

La présence des points leviers et des points aberrants n'est pas gênante tant qu'ils n'influencent pas l'estimation du modèle de manière importante. Une observation est *influente* si son omission change d'une manière importante le modèle estimé (Collett (1991)). Il s'agit donc de mesurer la différence entre $\hat{\beta}$ et $\hat{\beta}_{(i)}$, où $\hat{\beta}_{(i)}$ désigne le vecteur des paramètres estimés sans la i-ième observation (McCullagh et Nelder (1989a)). La distance mesurant cette différence est appelée *la distance de Cook* et est définie ar (Collett (1991)) :

$$D_i = \frac{1}{K}(\hat{\beta} - \hat{\beta}_{(i)})^T X^T W X (\hat{\beta} - \hat{\beta}_{(i)}) \tag{4.20}$$

Elle peut être approximée par :

$$D_i \approx \frac{h_i r_{P_i}^2}{K(1-h_i)} \tag{4.21}$$

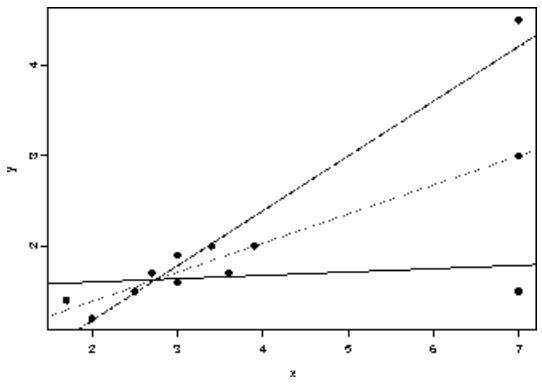


FIG. 4.1 – Les points leviers, $x = 7$, dans le cas d'une régression linéaire : le changement de réponse d'un point éloigné du cluster des points peut avoir une influence beaucoup plus importante sur l'estimation que les changements de réponse des points se trouvant dans le cluster des données.

Cette dernière expression permet d'interpréter l'influence d'un point en termes du leverage et de son résidu. La valeur de la fonction $\frac{h_i}{1-h_i}$ est inférieure à 1 pour les valeurs de leverage comprises entre 0 et 0.5. Elle est peu croissante pour les valeurs de h_i entre 0.8 et 1. Pour les valeurs de h_i se rapprochant de 1 elle tend vers $+\infty$. Par conséquent, plus le leverage sera élevé, et notamment entre 0.8 et 1, plus le point sera potentiellement influent. La distance de Cook de l'observation i est également directement proportionnelle avec le carré de son résidu de Pearson standardisé. Ainsi, si le résidu est petit mais le leverage elevé, le point peut être influent. Inversement, si le résidu est elevé mais le leverage entre 0 et 0.5 son influence sera diminuée. L'expression de la distance est pondérée par $\frac{1}{K}$ afin d'obtenir l'influence moyenne sur tous les paramètres. D'un point de vue pratique, la règle usuelle est de considérer comme influents les points ayant une distance de Cook supérieure à 1 (McCullagh et Nelder (1989a)).

Lorsque nous effectuons une analyse des résidus on s'intéresse surtout à isoler les observations influentes.

Surdispersion

Il se peut que l'inadéquation entre le modèle et les données ne soit pas dûe à un certain nombre d'observations isolées mais caractérise plutôt l'ensemble d'observations. Lorsque nous considérons que le modèle logistique est tout de même adapté pour modéliser les données en question, et lorsque leur variabilité est plus grande que celle prédite par le modèle nous avons à faire à *la surdispersion.*

La surdispersion dans les données peut provenir de sources diverses. Elle reflète généralement le manque d'indépendance ou manque d'homogénéité dans les données. Plusieurs modèles de survenue de la surdispersion peuvent être adoptés.

- *Les clusters* McCullagh et Nelder (1989a) proposent une explication possible de la surdispersion qui serait dûe à l'absence de la prise en compte des variables explicatives supplémentaires nécessaires pour expliquer les données. Dans le cas des variables explicatives nominales chacune des catégories contiendrait, en fait, des clusters correspondant aux sous-catégories. La variance serait plus élevée au sein de chacune des catégories car les sous-catégories n'ont pas été modélisées.
- *Corrélation entre les réponses binaires* Dans le cas d'une corrélation entre les observations la variance observée ne correspondrait pas à la variance sous le modèle binomial. Une corrélation positive entraîne une plus grande variabilité que celle prédite par le modèle binomial (Collett (1991)).

Lorsque nous pouvons attribuer la survenue de la surdispersion à une des sources évoquées, elle peut être modélisée en considérant que le paramètre de la loi binomiale corespondant à la probabilité de réponse de l'observation i est une variable aléatoire π_i d'espérance $E(\pi_i) = p_i$ et de variance :

$$Var(\pi_i) = \phi p_i(1 - p_i) \tag{4.22}$$

où $\phi > 0$ est un paramètre inconnu, que l'on appellera *le paramètre de dispersion*. Un tel modèle est appelé *modèle quasi-binomial*. On peut montrer (Collett (1991)) que l'espérance et la variance de la variable de réponse Y_i s'écrivent alors :

$$E(y_i) = n_i p_i \tag{4.23}$$

$$Var(y_i) = n_i p_i(1 - p_i)[1 + (n_i - 1)\phi] \tag{4.24}$$

Le paramètre de dispersion ϕ peut être estimé par (McCullagh et Nelder (1989a)) :

$$\hat{\phi} = \frac{1}{N - K} \sum_{i=1}^{N} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1 - \hat{p}_i)} \tag{4.25}$$

Sa valeur peut également s'exprimer en fonction de la statistique X^2 de Pearson (à l'aide de l'équation 4.15) :

$$\hat{\phi} = \frac{X^2}{N - K} \tag{4.26}$$

4.1.1.4 Tester les hypothèses sous le modèle de régression logistique

Avant d'introduire les tests effectués dans le cadre de la régression logistique nous faisons un rappel des notions principales employées dans le cadre des tests statistiques.
Définition d'un test Sur la base d'un échantillon, on observe N réalisations d'une variable X qui suit une loi $P_\theta, \theta \in \Theta$, et pour $\Theta_0, \Theta_1 \subset \Theta$ on veut déterminer parmi deux hypothèses $\mathcal{H}_0 : \theta \in \Theta_0$ (hypothèse nulle) et $\mathcal{H}_1 : \theta \in \Theta_1$ (hypothèse alternative) celle qui est vraie. On va alors tester $\mathcal{H}_0$ contre $\mathcal{H}_1$ et décider de rejeter ou non $\mathcal{H}_0$, le rejet de $\mathcal{H}_0$ impliquant l'acceptation de $\mathcal{H}_1$. Le test est fondé sur une fonction des valeurs de la variable X observées dans l'échantillon, appelée *la statistique de test*.

Les tests statistiques présentés ci-après s'appuient sur la méthodologie des tests des modèles emboîtés, à savoir dans le cas $\Theta_0 \subset \Theta_1$.
Degré de signification d'un test Lorsque nous voulons tester l'hypothèse $\mathcal{H}_0$ contre

$p-valeur$	$\mathcal{H}_0$
$P > 0.1$	$\mathcal{H}_0$ pas rejetée
$0.05 < P \leq 0.1$	légère preuve contre $\mathcal{H}_0$
$0.01 < P \leq 0.05$	une évidence modérée contre $\mathcal{H}_0$
$0.001 < P \leq 0.01$	forte preuve contre $\mathcal{H}_0$
$P \leq 0.001$	$\mathcal{H}_0$ rejetée

TAB. 4.2 – Les différentes p-valeurs et leur implication sur l'acceptation ou rejet de l'hypothèse $\mathcal{H}_0$. Les valeurs indiquées dans ce tableau correspondent à des probabilités de 10%, 5%, 1%, 0.1% de rejeter $\mathcal{H}_0$ à tort.

l'hypothèse $\mathcal{H}_1$ au regard des données il n'est pas possible de démontrer que l'hypothèse $\mathcal{H}_1$ est vraie et que $\mathcal{H}_0$ est fausse. La méthode de test consiste à supposer que l'hypothèse $\mathcal{H}_0$ est vraie, et à s'intéresser à la distribution de la statistique de test pertinente pour les données analysées, sous l'hypothèse $\mathcal{H}_0$.

Du point de vue pratique, les résultats des tests sont exprimés en *p-valeur* qui est la probabilité d'observer une valeur de la statistique plus exceptionnelle que la valeur observée sur l'echantillon de test, sous $\mathcal{H}_0$. Plus la *p-valeur* est petite, plus la preuve contre l'hypothèse $\mathcal{H}_0$ est forte et le test significatif (l'interprétation habituelle des différentes *p-valeurs* est indiquée dans le tableau 4.2).

Dans le cadre de ce travail nous aurons besoin des tests statistiques pour décider de deux questions :

1. Le modèle que l'on ajuste est-il en adéquation avec les données ?
2. Existe-t-il une différence entre les paramètres estimés ? Dans le cas des variables explicatives nominales, cette question se traduit en question d'effet de groupe, à savoir si les différents groupes associés aux différentes modalités des variables explicatives sont caractérisés par des lois différentes (équation 4.9).

Les tests statistiques permettant de décider de ces questions dans le cadre des modèles linéaires généralisés sont présentés ci-après.

Test d'ajustement du modèle

Le test d'adéquation est effectué afin de décider si le modèle ajusté est conforme aux données. Il revient à comparer le modèle ajusté au modèle saturé (Maccario (1998)), qui comporte un paramètre pour chaque observation (donc N paramètres). Le nombre de

degrés de liberté pour le test est égal à la différence de nombre de degrés de liberté (ddl) chacun des modèles, c'est à dire $N-K-1$, où N est le nombre d'observations et K le nombre de variables explicatives. Le résultat classique sur les propriétés asymptotiques de test de maximum de vraisemblance pour les modèles emboîtés (McCullagh et Nelder (1989c)) indique que 2 fois la différence de leurs log-vraisemblances suit asymptotiquement la loi du χ^2 à $N-K-1$ degrés de liberté. Du fait que $Dev = -2\log(\frac{L_M}{L_S})$, nous avons :

$$Dev \sim \chi^2_{N-K-1} \tag{4.27}$$

Par conséquent, pour tester l'adéquation, on utilise la p-valeur calculée sur la déviance observée Dev_{obs} :

$$p = P(\chi^2_{(N-K-1)ddl} < Dev_{obs}) \tag{4.28}$$

Pour un χ^2 à $N-K-1$ degrés de liberté nous avons :

$$E(Dev) = N-K-1 \tag{4.29}$$

Cette dernière expression donne la règle *ad hoc* pour évaluer l'adéquation du modèle

$$Dev \approx N-K-1 \tag{4.30}$$

Test de l'effet de groupe

Dans le cadre des modèles linéaires généralisés il est possible de tester différentes caractéristiques des paramètres estimés. Ainsi nous pouvons tester si deux paramètres estimés sont égaux entre eux ($\mathcal{H}_0 : \beta_k = \beta_l$), si un des paramètres est nul ($\mathcal{H}_0 : \beta_k = 0$) ou si tous les paramètres sont égaux ($\mathcal{H}_0 : \beta_1 = \beta_2 = \ldots = \beta_K = 0$). Dans ce dernier cas :

$$\mathcal{H}_0 \quad : \quad \beta_1 = \beta_2 = \ldots = \beta_K = 0 \tag{4.31}$$

$$\mathcal{H}_1 \quad : \quad \text{il existe au moins un } \beta_i \text{ non nul} \tag{4.32}$$

Dans le cas des variables explicatives nominales, on appelle ce test le *test de l'effet de groupe*. Il permet de conclure s'il existe au moins une modalité de la variable explicative pour laquelle le paramètre p_i associé est différent du paramètre correspondant à la modalité de référence.

Ces tests sont effectués à l'aide de tests des modèles emboîtés (McCullagh et Nelder (1989c)). En particulier, afin d'effectuer le test de l'effet de groupe, nous pouvons voir

le modèle M_0 à paramètres égaux, $\beta_1 = \beta_2 = \ldots = \beta_K = 0$ comme le cas spécial du modèle général M à K paramètres $\beta_1, \ldots, \beta_K$. Dans ce cas, comme les paramètres β sont estimés par maximum de vraisemblance, nous avons que la statistique de rapport de vraisemblance suit un χ^2 à $ddl_0 - ddl = K$ degrés de liberté, où ddl_0 et ddl sont le nombre de degrés de liberté des modèles M_0 et M, respectivement, et $L(M_0)$ et $L_(M)$ sont respectivement les valeurs maximales des vraisemblances des modèles M_0 et M. Nous avons donc que :

$$-2 \log \frac{L(M_0)}{L_(M)} \sim \chi^2_K \tag{4.33}$$

Notons que, dans le cas du test de l'effet de groupe, nous avons que $ddl_0 - ddl = K$.

Lorsque nous cherchons plus particulierement à examiner la différence entre deux paramètres spécifiques le test des modèles emboîtés peut à nouveau être utilisé. Sans perdre en généralité, nous pouvons prendre l'exemple d'une variable explicative nominale à 3 modalités et nous pouvons nous intéresser à savoir si β_1 est égal à β_2 mais différent de β_3. Nous définirons des nouvelles hypothèses :

$$\mathcal{H}'_0 : \beta_1 = \beta_2, \beta_3 \text{ quelconque} \tag{4.34}$$

$$\mathcal{H}'_1 : \beta_1 \neq \beta_2, \beta_3 \text{ quelconque} \tag{4.35}$$

Le modèle M sera défini comme précédemment et un modèle M'_0 correspondant à l'hypothèse $\mathcal{H}'_0$. Là encore, la statistique de test de rapport de maximum de vraisemblance de ces deux modèles suit asymptotiquement un χ^2 à $ddl - ddl'_0 = 1$ degrés de liberté, où ddl et ddl'_0 sont le nombre de degrés de liberté des modèles M et M'_0, respectivement. :

$$-2 \log \frac{L(M'_0)}{L(M)} \sim \chi^2_1 \tag{4.36}$$

4.1.2 Application

4.1.2.1 Modélisation du G+C% génomique par le modèle de régression logistique

Nous utilisons la régression logistique pour modéliser le G+C% génomique d'un organisme, et pour mettre en évidence, sous ce modèle, la différence présumée entre le G+C% des ARNnc et le G+C% du reste de génome.

Les études évoquées précédemment (Klein *et al.* (2002), Schattner (2002)) considéraient le G+C% dans deux groupes : les ARNnc et le reste du génome. Néanmoins, une pression évolutive forte agit sur les séquences codantes. De ce fait, nous pouvons supposer que leur composition a été influencée de façon particulière et que celle-ci diffère de la composition du reste du génome. Le cadre GLM permettant une modélisation aisée dans plus de deux groupes, nous considérons un troisième groupe correspondant aux séquences codantes, disponibles à partir des annotations des génomes. On prendra soin lors de l'application de vérifier que l'introduction du troisième groupe est pertinente.

Nous définissons donc trois groupes génomiques : l'ensemble des séquences d'ARNnc, l'ensemble des séquences codantes et l'ensemble des séquences du reste du génome (cet ensemble sera défini plus tard de façon précise). On notera ces groupes G_0, G_1 et G_2, respectivement.

Les problèmes modélisés par une régression logistique se présentent sous la forme d'une variable de réponse qui peut avoir deux valeurs : *succès* et *echec*. Pour modéliser le G+C% dans une séquence génomique, nous considérerons que, si un des nucléotides G ou C est rencontré, il s'agit d'un succès et si c'est un des nucléotides A ou T, qu'il s'agit d'un echec :

$$y = \begin{cases} 1 & \text{si G,C} \\ 0 & \text{si A,T} \end{cases}$$

Dans la modélisation que nous proposons, les séquences génomiques représenteront les individus statistiques, au nombre de N. La variable réponse Y_i prend la valeur du nombre de G+C dans la i-ième séquence et nous supposons que la variable Y_i suit une loi de binomiale de paramètre p_i, c'est-à-dire, que la probabilité d'avoir un G ou un C pour le nucléotide i est de p_i. Nous voulons expliquer le $(G+C)\%$ par l'appartenance à un groupe en introduisant une variable explicative nominale X correspondant à un des trois groupes définis.

Transformé ainsi, nous pouvons reformuler le problème comme étant l'analyse de la relation entre la probabilité du succès dans une séquence et les variables explicatives d'appartenance à chacun des trois groupes. Notons que, en modélisant ce problème par une régression logistique nous faisons l'hypothèse d'indépendance des bases dans les séquences ainsi que de l'égalité de la probabilité de G+C dans toute séquence au sein du même groupe. Ce sont des hypothèses fortes, mais la prise en compte de la surdispersion permet de les relâcher.

Dans le codage disjonctif complet, pour représenter une variable nominale à 3 modalité nous avons besoin de 2 variables explicatives x_1, x_2 qui prennent veleurs $(x_1, x_2) = (0, 0)$ si le nucléotide appartient à G_0, valeur $(x_1, x_2) = (1, 0)$ si il est dans le groupe G_1 et $(x_1, x_2) = (0, 1)$ si il est dans le groupe G_2.

Le modèle s'écrit alors comme :

$$logit(\hat{p}_i) = \hat{\mu} + \hat{\beta}_1 x_1^i + \hat{\beta}_2 x_2^i \tag{4.37}$$

Une fois les paramètres de la régression logistique estimés, nous avons, d'après l'équation 4.5 et si l'on utilise le codage disjonctif complet :

$$\begin{aligned} G_0 : \hat{p}_0 &= \frac{\exp(\hat{\mu})}{1 + \exp(\hat{\mu})} \\ G_1 : \hat{p}_1 &= \frac{\exp(\hat{\mu} + \hat{\beta}_1)}{1 + \exp(\hat{\mu} + \hat{\beta}_1)} \\ G_2 : \hat{p}_2 &= \frac{\exp(\hat{\mu} + \hat{\beta}_2)}{1 + exp(\hat{\mu} + \hat{\beta}_2)} \end{aligned} \tag{4.38}$$

Le modèle posé ainsi, nous l'avons d'abord appliqué sur le génome de *Staphylococcus aureus N315*. C'est un génome A+T riche (G+C% génomique moyen 33%) déjà étudié dans le laboratoire (voir Boisset *et al.* (2007), par exemple). De ce fait, des indications existent suggérant que la discrimination entre les ARNnc et le reste du génome sur la base de G+C% est possible dans ce génome. Tout en produisant un résultat sur le génome de *S.aureus*, cette étude permet de mettre au point la méthode afin de l'appliquer ensuite au génome de notre organisme d'intérêt, *Ralstonia solanacearum*.

Nous présentons dans la suite les résultats de ces deux analyses. Pour les effectuer, nous avons utilisé la fonction `glm()` et différentes fonctionnalités associées de R (R Development Core Team (2008)) pour l'ajustement et l'analyse du modèle logistique.

4.1.2.2 Application au génome de *Staphylococcus aureus*

Nous pouvons, à présent, estimer les paramètres du modèle sur les données du génome de *Staphylococcus aureus N315*. Comme expliqué précédemment, les données sont partagées en trois groupes : ARN, COD et AUTRE. L'hypothèse que nous cherchons à confirmer ou infirmer est que la composition en $G+C$ est différente dans chacun de ces groupes. Cette composition sera reflétée par les paramètres qui seront estimés dans le modèle et un test sur l'effet de groupe permettra de trancher si la différence en $G+C$ observée dans les groupes est significative.

Les données Les données présentes dans le groupe ARN proviennent d'une part de NCBI (version de Février 2008) pour les ARNr, ARNt et tmRNA (79 ARNnc) et d'autre part de Rfam (version 8.1), pour les autres ARNnc annotés (32 ARNnc). Notons que cet ensemble ne contient pas, en raison d'absence de Rfam, l'ARN III, un ARNnc de *S.aures* (Boisset *et al.* (2007)).

Les séquences codantes sont extraites des fichiers d'annotation du NCBI et l'ensemble AUTRE est constitué comme le complément de l'union des ensembles ARN et COD, dans leurs coordonnées absolues, par rapport au brin direct du génome.

Le résumé des caractéristiques de chacun de ces ensembles est donné dans le tableau 4.3. Le graphique 4.2 représente les distributions estimées de G+C% dans chacun des 3 groupes. Dans les groupes COD et AUTRE les distributions du G+C% se répartissent symétriquement par rapport aux G+C% moyens. Dans le groupe ARN, le G+C% présente une bimodalité que nous tenterons d'expliquer dans la suite. Notons encore que dans le groupe AUTRE certaines séquences présentent un G+C% proche ou égal à 0. Les séquences à G+C% nul sont au nombre de 93 et elle sont toutes de longueur inférieure à 20nt. Elles correspondent probablement aux régions intergéniques séparant les gènes organisés en opérons (les gènes co-transcrits, espacés par les séquences intergéniques courtes).

Classe	Nbr.	Nbr nt.	Long. moy.	sd	G+C% moy.	sd	G+C% moy. pondéré (*)
ARN	110	33022	300.2	651.97	0.509	0.097	0.494
COD	2588	2339644	904.3	762.66	0.333	0.033	0.336
AUTRE	2403	444688	185.05	205.50	0.265	0.081	0.272
Total	5101	2817354					

TAB. 4.3 – *Staphylococcus aureus* : résumé des caractéristiques des groupes ARN, COD et AUTRE. (*) : la moyenne pondérée est calculée avec la formule $\frac{\sum_{i=1}^{N_G} y_i}{\sum_{i=1}^{N_G} n_i}$, avec N_G le nombre d'observations dans le groupe G.

Estimation

Le modèle de régression logistique M est ajusté sur les données décrites donnant les paramètres estimés :

$$\hat{\mu} = -0.026, \hat{\beta}_1 = -0.651, \hat{\beta}_2 = -0.948 \tag{4.39}$$

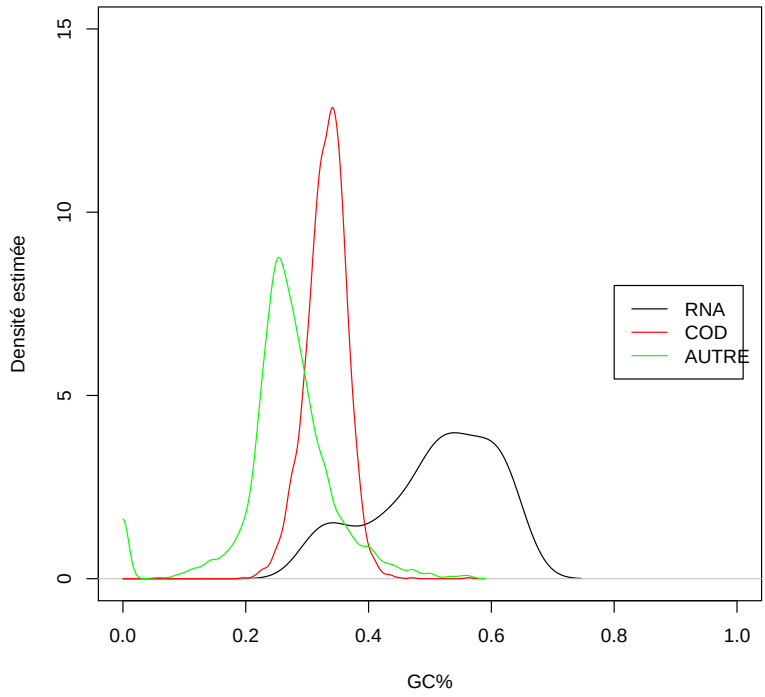


FIG. 4.2 – *Staphylococcus aureus* : distribution estimée de G+C% dans les groupes ARN, COD et AUTRE. Dans les groupes COD et AUTRE le G+C% se repartie symétriquement par rapport à leurs G+C% moyens. Dans le groupe ARN, le G+C% présente une bimodalité qui sera expliquée dans la suite.

donnant des valeurs de $G+C\%$ prédites pour chaque groupe :

$$\hat{p}_{ARN} = 0.493, \hat{p}_{COD} = 0.336, \hat{p}_{AUTRE} = 0.274 \tag{4.40}$$

La déviance du modèle est égale à 13671 pour 5098 degrés de liberté et conformément à ceci ($Dev > ddl$) le test d'ajustement rejette l'hypothèse $\mathcal{H}_0$ avec $p - valeur < 10^{-8}$, signifiant que le modèle n'est pas adéquat. Nous avons donc du procéder à la recherche des éventuels points atypiques.

Analyse des résidus

Nous analysons ici l'écart entre les valeurs observées et les valeurs prédites des points isolés, ainsi que leur influence sur l'estimation du modèle, dans le cas des données de *S. aureus*.

La figure 4.3 représente les distances de Cook des observations en fonction du numéro de l'observation. Toutes les distances sont inférieures à 1 et par conséquent aucun point n'est pas influent dans le modèle.

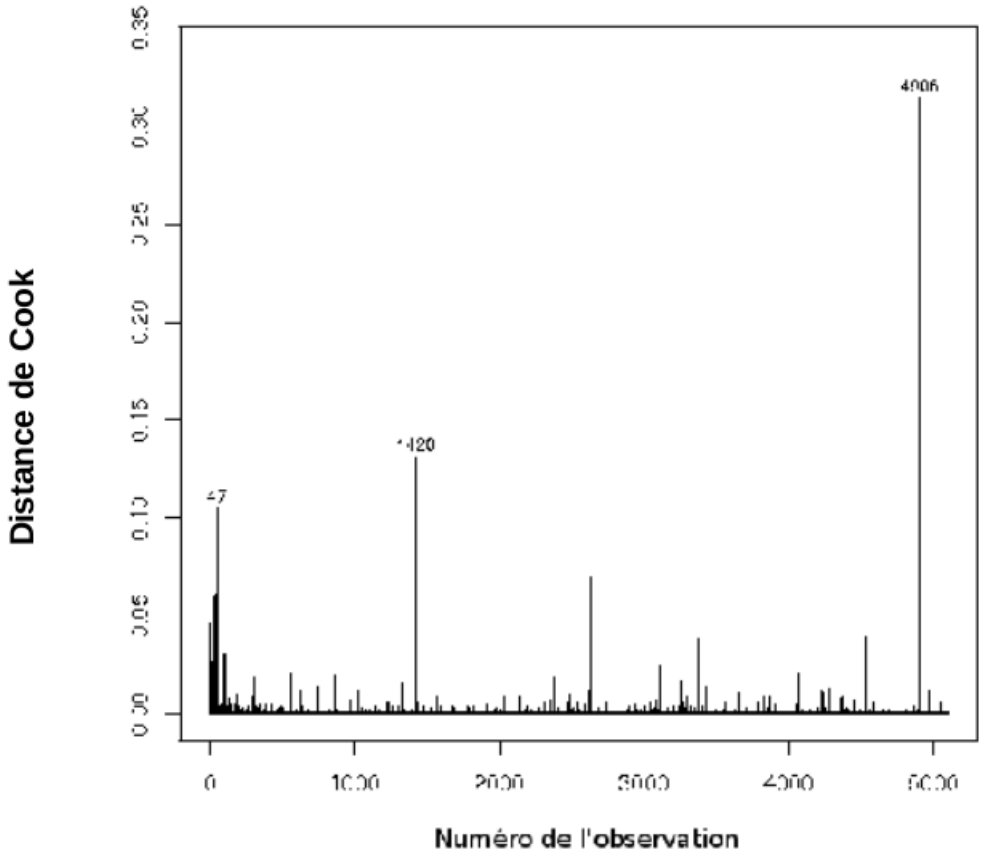


FIG. 4.3 – Les distances de Cook de l'ensemble des observations, pour le modèle de régression décrit dans l'équation 4.39.

Néanmoins, certains points se détachent du reste des données par leur influence. Ce sont les points 47, 1420 et 4906 , figure 4.3 et tableau 4.4). L'observation 47 correspond à la séquence d'un ARN riboswitch (appartenant à la famille des T-box). Les observations 1420 et 4906 appartiennent aux groupes COD et AUTRE, respectivement, et de façon intéressante contiennent toutes les deux des répétitions.

Afin de déterminer quelle est la part des résidus et quelle est la part du leverage dans les distances de Cook observées nous traçons deux graphiques : un représentant la distance de Cook en fonction des résidus de Pearson standardisés et l'autre en fonction

Num. obs.	G+C	Longueur	Pos. début	Pos. fin	Brin	Groupe	Description
47	0.35	301	1696617	1696918	-	ARN	T-box
1420	0.36	20138	1437931	1458069	-	COD	GeneID :1124105, Rep 24
4906	0.36	3167	2559319	2562486	+	AUTRE	Rep 43

TAB. 4.4 – L'analyse des points influents de la figure 4.3. Rep 24 et Rep 43 correspondents à des répétitions présentes dans la CRISPRdb (Grissa *et al.* (2007)), la base des données des CRISPR.

du leverage (figure 4.4).

Pour les points les plus influents (les points 1, 2, 3, 4 et 5 de la figure 4.4) les résidus de Pearson standardisés ont une valeur élevée tandis que leur leverage est bas. Ces graphiques suggèrent que dans le modèle ajusté, les points influents dépendent essentiellement des résidus de Pearson standardisés. En conséquence, dans la suite nous raisonnerons en termes des résidus de Pearson standardisés et de la distance de Cook, sans tenir compte du leverage.

Jusque là nous avons raisonné en terme des points les plus influents globalement. Cependant, nous pouvons également nous intéresser aux points influents dans chacun des groupes. Leur identification peut aider à mieux connaître la composition de chacun des trois groupes.

Figure 4.5 présente la distance de Cook en fonction des résidus de Pearson standardisés. Même si elle est globalement peu élevée, la distance de Cook varie selon les groupes. Sauf quelques points influents déja identifiés (tableau 4.4), les valeurs de la distance de Cook dans les groupes COD et AUTRE varient peu (valeur au-dessous de 0.02). En revanche, une fraction du groupe ARN présente une distance de Cook plus élevée (encadré, groupe ARN, figure 4.5). Ces points s'écartent du modèle en raison de leurs résidus plus élevés et pourraient être à la cause de la formation de bimodalité dans la distribution de G+C% dans le groupe ARN (figure 4.2). Ils ont donc été isolés (tableau 4.5). Parmi les 18 points isolés, se détachant du reste des observations du groupe ARN par leur composition en G+C% plus basse, 17 d'entre eux, que ça soit les riboswitch ou les séquences "leader", appartiennent aux familles d'ARNnc agissant en tant que régulateur en *cis* se trouvant dans les région 5' des gènes adjacents (nous les appelerons *les 5' cis-régulateurs*). Parmi ces points se trouve notamment la totalité des membres de la famille T-box de l'ensemble ARN.

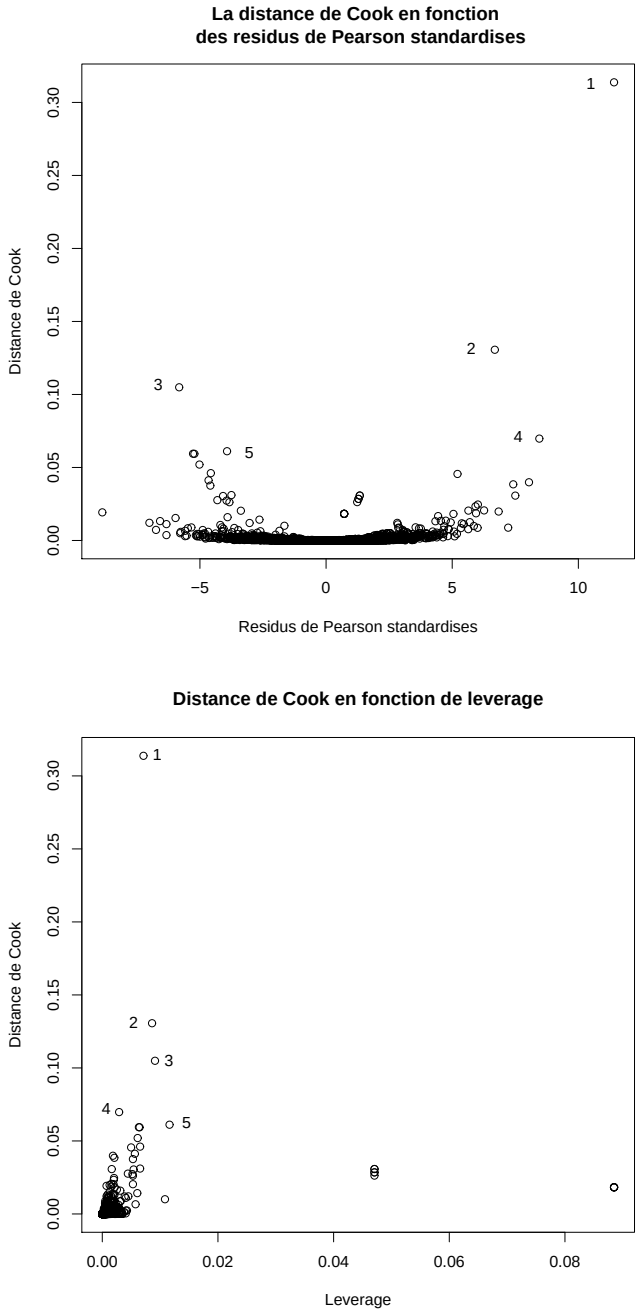


FIG. 4.4 – La distance de Cook en fonction des résidus de Pearson standardisés et du leverage. Les numéros des points ont été introduits pour faire la correspondance entre les deux graphiques et ne correspondent pas aux numéros d'observations.

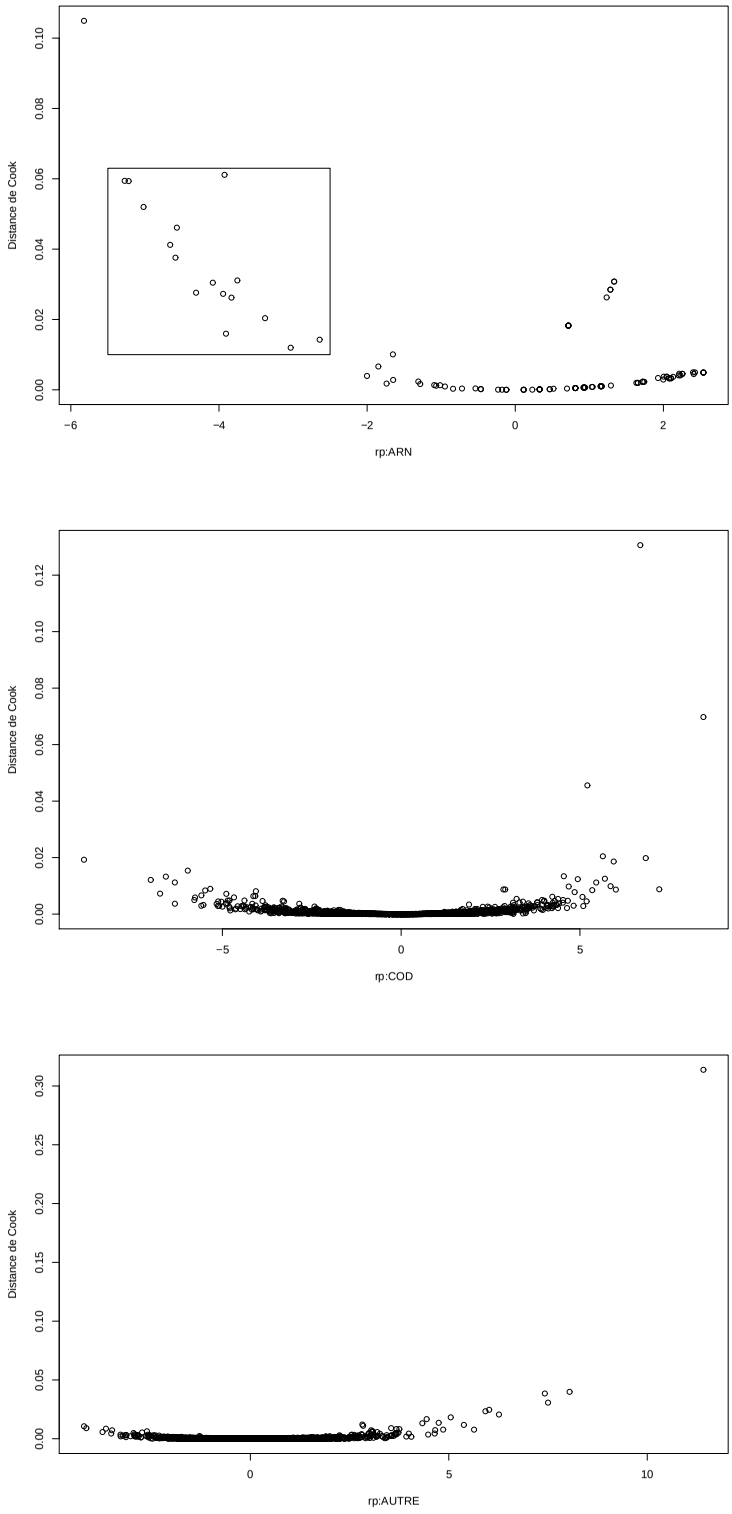


FIG. 4.5 – Les graphiques de la distance de Cook en fonction des résidus de Pearson standardisés, par groupe.

Num. obs.	G+C	Longueur	Pos. début	Pos. fin	Brin	Famille
47	0.32	301	1696617	1696918	-	T-box riboswitch
50	0.31	209	1716121	1716330	-	T-box riboswitch
35	0.31	213	1114012	1114225	+	T-box riboswitch
51	0.31	202	1773882	1774084	-	T-box riboswitch
40	0.32	186	1372336	1372522	+	T-box riboswitch
49	0.31	175	1713742	1713917	-	Lysine riboswitch
1	0.33	216	12487	12703	+	T-box riboswitch
27	0.31	146	577527	577673	+	L10 leader
45	0.33	197	1649276	1649455	-	T-box
37	0.34	172	1171095	1171267	+	T-box riboswitch
42	0.39	384	1483784	1484168	-	RNase P
6	0.29	103	430797	430900	+	Purine riboswitch
41	0.34	175	1399486	1399661	+	Lysine riboswitch
99	0.36	216	2212090	2212306	-	GlmS
39	0.36	175	1372133	1372308	+	T-box riboswitch
31	0.35	128	995898	996026	+	yybP-ykoY
5	0.39	200	407703	407903	-	T-box riboswitch
44	0.38	97	1576866	1576963	-	Glycine riboswitch

TAB. 4.5 – L'analyse des points du groupe ARN ayant la valeur de résidu de Pearson standardisée inférieure à -2. Les points sont ordonnés par ordre croissant de valeur des résidus de Pearson standardisés.

Ces derniers résultats nous ont suggéré que l'ensemble des élements 5' cis-régulateurs chez *S. aureus* pourrait avoir une composition particulière. Et en effet, la totalité des éléments 5' cis-régulateurs, au nombre de 27, présente une composition plus basse en G+C% que le reste du groupe ARN, et celle-ci semble être rapprochée de la composition de l'ensemble COD (figure 4.6, et tableau 4.6).

La présence des séquences 5' cis-régulatrices dans l'ensemble ARN explique la bimodalité dans la distribution de G+C% dans ce groupe (figure 4.2). En effet, ces points retirés de l'ensemble ARN, la bimodalité disparaît (figure 4.7).

En dehors du nuage des points du groupe ARN présentant la distance de Cook plus élevée, nous constatons (figure 4.5) que les résidus de Pearson sont contenus dans

Sous-groupe ARN	Moyenne G+C%	Min.	Max	sd
5' cis-régulateurs	0.382	0.309	0.479	0.053
Reste	0.550	0.298	0.645	0.067

TAB. 4.6 – *S. aureus* : le résumé de la composition en G+C% dans les sous-groupes des éléments 5' cis-régulateurs et le reste du groupe ARN. Le groupe des 5' cis-régulateurs contient 27 observations et le reste du groupeARN en contient 83.

l'intervalle de confiance $(-1.96, 1.96)$ uniquement pour le groupe ARN, pour les deux autres groupes cet intervalle étant plus étendu allant d'environ -5 à 5. Ceci suggère que, dans les groupes COD et AUTRES, la variabilité des données est sous-estimée et explique la non-adéquation du modèle binomial. Afin de prendre en compte la variabilité non-binomiale des données nous devons procédér à la modélisation de la surdispersion.

Modélisation de la surdispersion La valeur estimée du coefficient de dispersion $\hat{\phi}$ est 2.654619. La prise en compte de la surdispersion dans le modèle ne change pas des valeurs prédites des coefficients de régression et des G+C% prédits. Ils restent donc

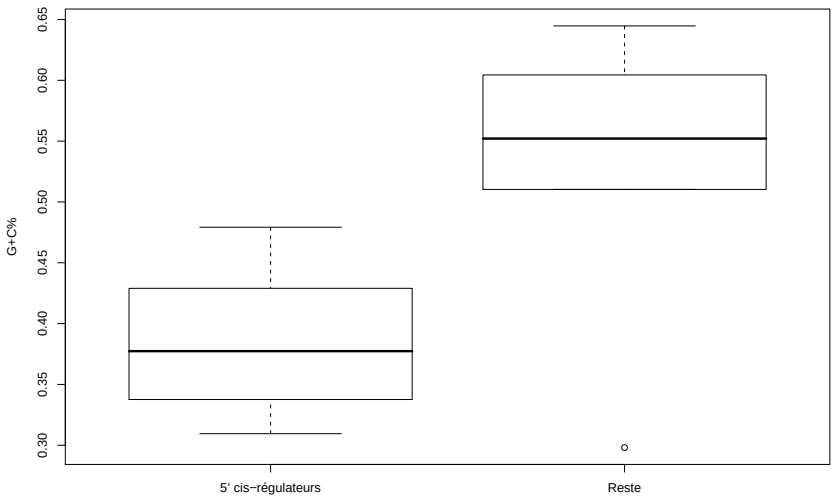


FIG. 4.6 – *S. aureus* : les boxplot des distribution de G+C% au sein du groupe ARN, d'une part dans le sous-groupe des 5' cis-régulateurs et d'autre part dans le reste du groupe. Les 5' cis-régulateurs sont au nombre de 27 et les autres éléments du groupe ARN sont au nombre de 83.

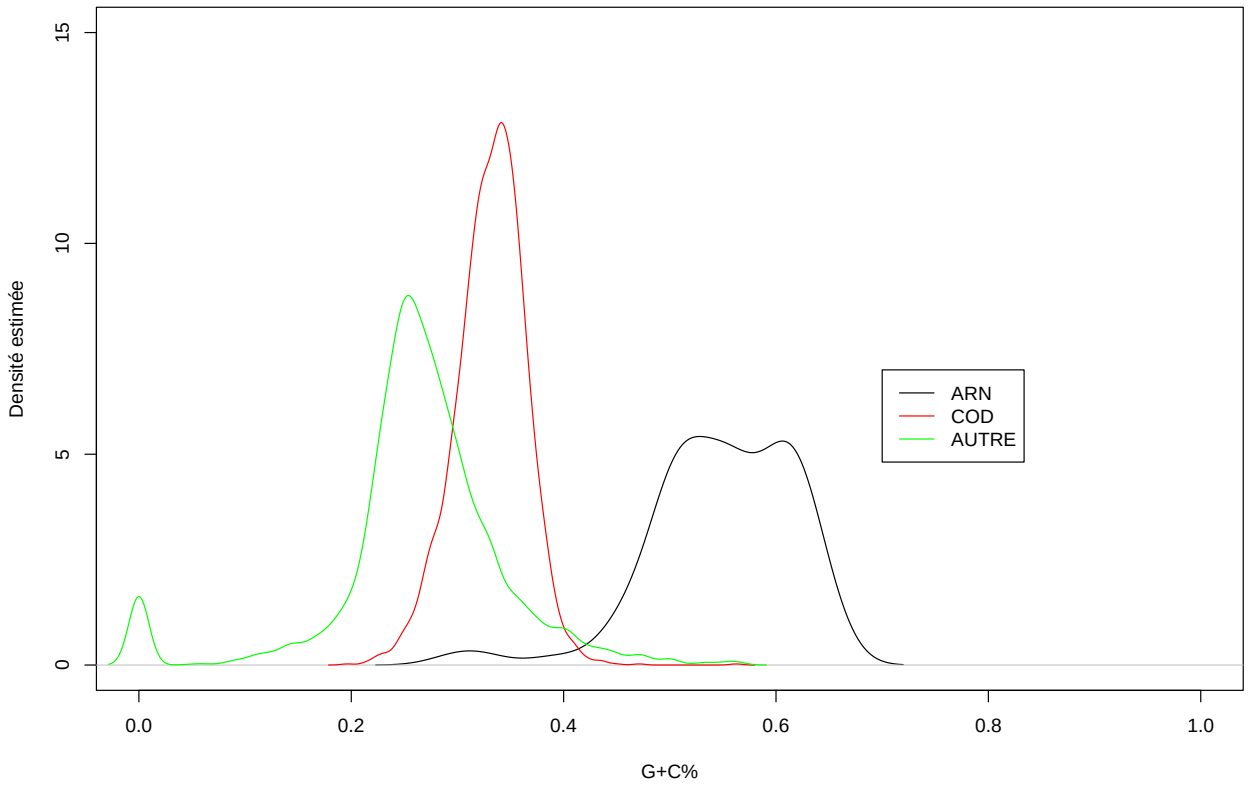


FIG. 4.7 – *Staphylococcus aureus* : distribution estimée de G+C% dans les groupes ARN, COD et AUTRE.

identiques aux valeurs prédites dans l'équation 4.39.

Le modèle s'ajuste bien aux données par construction de la prise en compte de la surdispersion. La variance prédite par le modèle change et ceci est reflété dans le test de l'effet de groupe qui est très significatif (p-$valeur < 10^{-8}$). Ceci permet de conclure qu'au moins un des groupes a un G+C% significativement différent des deux autres.

Nous nous intéressons plus particulièrement au groupe ARN. Pour cette raison nous devons nous assurer que son G+C% est différent de chacun des deux autres, ce qui est fait à l'aide du test d'égalité de deux paramètres (cf. équation 4.36). Le test est très significatif (p-$valeur < 10^{-8}$), ce qui permet de conclure que le G+C% du groupe ARN est significativement différent des deux autres, sous le modèle logistique.

4.1.2.3 Application au génome de *Ralstonia solanacearum*

Nous appliquons à présent la régression logistique sur les données du chromosome de *Ralstonia solanacearum*. Seule séquence du chromosome a été utilisée, étant donnée que la majorité des ARNnc connus de *R. solanacearum* se trouvent sur ce réplicon (voir tableau 1.2). Le plasmide ayant la composion similaire à celles du chromosome, les résultats obtenus pourront être extrapolés sur ce deuxième replicon. Les ensembles ARN, COD et AUTRE ont été générés de la même façon que les ensembles de séquences correspondants de *S. aureus* (décrits dans la section précédente). Le résumé des caractéristiques de chacun des groupes est donné dans le tableau 4.7. Les G+C% moyens par séquence (moyenne pondérée) sont différents entre les différents groupes, mais cette différence est moins importante que dans le cas de *S. aureus*. De plus, elle est pratiquement estompée dans le cas de la moyenne globale en G+C%.

Classe	Nbr.	Nbr nt.	Long. moy.	sd	G+C% moy.	sd	G+C% moy. ponédérée (*)
ARN	69	18482	267.85	618.39	0.605	0.044	0.559
COD	3440	3257032	946.81	739.42	0.672	0.045	0.677
AUTRE	3062	445399	145.46	201.56	0.609	0.102	0.619
Total	6571	3720913					

TAB. 4.7 – *Ralstonia solanacearum* : résumé des caractéristiques des groupes ARN, COD et AUTRE. (*) : la moyenne pondérée est calculée avec la formule $\frac{\sum_{i=1}^{N_G} y_i}{\sum_{i=1}^{N_G} n_i}$, avec N_G le nombre d'observations dans le groupe G.

La figure 4.8 représente les distributions estimées de G+C% dans chacun des 3 groupes. Les distributions dans les groupes ARN et COD semblent régulières et séparées entre elles. En revanche, la distribution dans l'ensemble AUTRE est étendue et recouvre une partie de la distribution des ARN, suggérant qu'une discrimination des séquences d'ARN des autres éléments du génome sur la seule base de G+C% n'est pas envisageable. Cette conclusion sera confirmée dans la suite par les tests effectués dans le cadre du modèle logistique.

Modèle estimé

Le modèle de régression logistique a été estimé sur les données de *R.solanacearum* donnant les valeurs des coefficients de régression :

$$\hat{\mu} = 0.231, \hat{\beta}_1 = 0.509, \hat{\beta}_2 = 0.249 \tag{4.41}$$

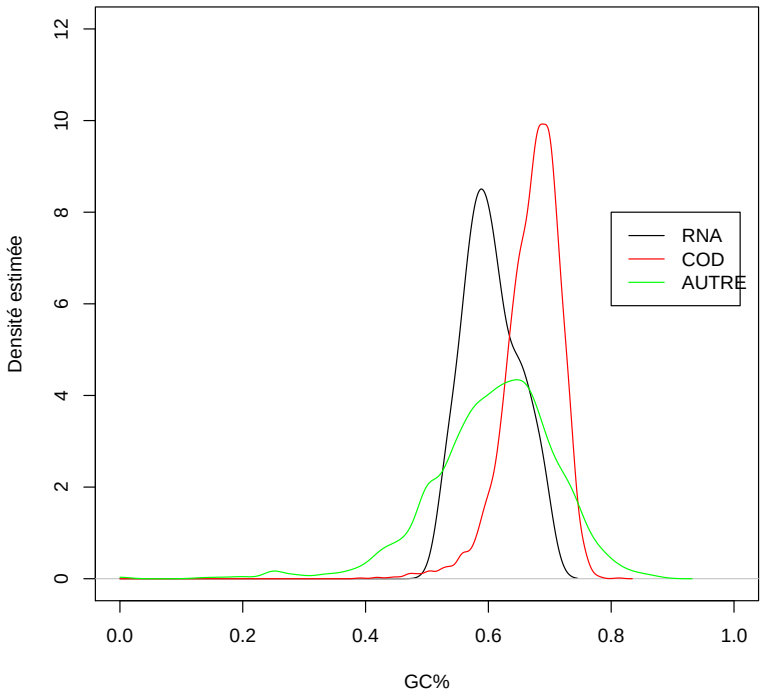


FIG. 4.8 – *Ralstonia solanacearum* : distribution estimée de G+C% dans les groupes ARN, COD et AUTRE

avec les valeurs de $G+C\%$ par groupe prédites :

$$\hat{p}_{ARN} = 0.557, \hat{p}_{COD} = 0.677, \hat{p}_{AUTRE} = 0.618 \tag{4.42}$$

A l'instar des données de *S.aureus*, les données de *R. solanacearum* sont surdispersées avec un facteur de dispersion estimé $\hat{\phi} = 5.04$.

Analyse des résidus et de la surdispersion

Six points influents ont été identifiés, correspondant tous aux ARNr (figure non présentée). Ces observations présentent une composition en G+C plus basse que le reste du groupe, leur moyenne en G+C% étant autour de 0.53. Leur séquences sont longues de plusieurs milliers de nucléotides et influent grandement l'estimation des paramètres.

Afin d'ajuster le modèle à la majorité des ARNnc présents dans le groupe ARN, nous avons retiré les points correspondant aux ARNr de l'étude. L'ensemble de données ainsi modifié présente toujours la surdispersion dans le modèle ajusté et ce modèle prédit les valeurs de G+C% par groupe suivants :

$$\hat{p}_{ARN} = 0.612, \hat{p}_{COD} = 0.677, \hat{p}_{AUTRE} = 0.618 \tag{4.43}$$

La comparaison de ces valeurs estimées avec les valeurs des paramètres estimés sur les données de départ (équation 4.42) montrent que la composition des séquences des ARNr qui ont été retirés influencent fortement l'estimation du G+C% dans le groupe ARN. Etant donné les valeurs rapprochées de $\hat{p}_{ARN} = 0.612$ et $\hat{p}_{AUTRE} = 0.618$ nous pouvons difficilement envisager une modélisation du G+C% de *R. solanacearum* par les trois groupes proposés. Ceci est confirmé par les résultats de test d'égalité de deux paramètres (équation 4.36) qui n'est pas significatif, lorsque nous tenons compte de la surdispersion.

Ces résultats montrent que, dans le cas de *R. solanacearum*, le G+C% seul ne peut pas être utilisé comme critère pour discriminer les trois groupes proposés.

Notons encore que les ARNnc correspondant aux 5' cis-régulateurs présentent un G+C% plus élevé que le reste du groupe ARN (figure 4.9 et tableau 4.8) et leur composition ne semble pas être proche de la composition de l'ensemble COD, comme dans le cas de *S. aureus*.

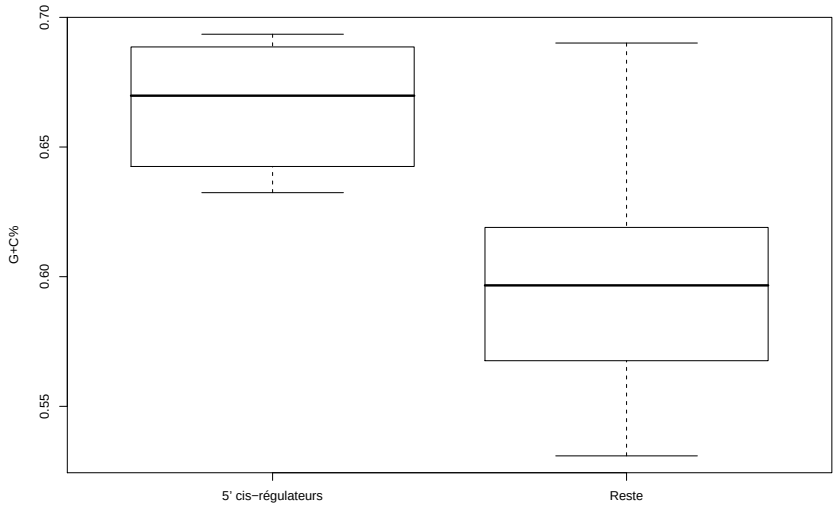


FIG. 4.9 – *R. solanacearum* : les boxplot des distribution de G+C% au sein du groupe ARN, d'une part dans le sous-groupe des 5' cis-régulateurs et d'autre part dans le reste du groupe. Les 5' cis-régulateurs sont au nombre de 6 et les autres éléments du groupe ARN sont au nombre de 63.

Sous-groupe ARN	Moyenne G+C%	Min.	Max	sd
5' cis-régulateurs	0.666	0.632	0.693	0.027
Reste	0.599	0.531	0.690	0.041

TAB. 4.8 – *R. solanacearum* : le résumé de la composition en G+C% dans les sous-groupes des éléments 5' cis-régulateurs et le reste du groupe ARN.

4.1.3 Discussion

Lorsque nous nous intéressons à la composition des génomes il est nécessaire de disposer des outils statistiques permettant de tirer les conclusions sur l'égalité ou les différences en composition des différents groupes génomiques. Même si cette tâche semble fondamentale dans le contexte d'étude de la composition des génomes, une démarche canonique n'existe pas et elle fait particulièrement défaut dans la plupart des travaux consacrés à l'étude de la composition des ARNnc (chapitre 3).

Nous venons de présenter et de proposer l'utilisation du cadre théorique des modèles linéaires généralisés pour étudier la composition en G+C dans différents groupes génomiques, et en particulier le groupe ARN. Ce cadre générique offre, d'une part, la possibilité de tester de façon aisée différentes hypothèses d'égalité entre les paramètres et d'autre part des outils exploratoires permettant d'isoler des éléments atypiques et potentiellement intéressants.

Staphylococcus aureus Les résultats de l'utilisation du modèle linéaire généralisé pour modéliser la composition en G+C suggèrent que ce paramètre peut être utilisé pour discriminer les ARNnc dans le génome de *S.aureus*, les tests l'affirmant étant très significatifs.

Ce résultat a été exploité dans un travail sur l'utilisation de biais de composition et les modèles de Markov cachés dans la recherche des ARNnc chez *S. aureus*.

Ce travail soulève la question intrigante du rapport entre la composition en G+C et les séquences répétées, dans le génome de *S. aureus*. En effet, des séquences s'écartant de la distribution prédite par le modèle et qui ne sont pas des ARNnc contiennent des séquences répétées. Toutes ces séquences présentent un G+C% plus élevé que le pourcentage prédit par le modèle.

Notons encore que les résultats que nous avons présentés se sont basés sur les données

ne contenant pas l'ARN III, ARN de *S. aureus*, 515nt de long, ayant une composition basse en G+C. Cette observation change le modèle ajusté et se présente comme un point influent, mais les conclusions sur les effets de groupe restent inchangés même en présence de cette observation (résultats non-présentés).

Ralstonia solanacearum L'application des GLM sur les données de *R. solanacearum* n'a pas permis de conclure qu'une discrimination est possible entre les ARN et le reste du génome sur la base de la composition. La modélisation de la composition en G+C ne semble pas être suffisante. Par conséquent, nous étudierons dans la suite une modélisation plus complexe, tout d'abord en modélisant la composition lettre par lettre et ensuite en introduisant une dépendance des lettres successives.

Néanmoins, nous avons pu observer que, chez *R. solanacearum*, le groupe ARN présente une composition en moyenne plus basse que les deux autres groupes et que les ARN ribosomiques présentent un G+C% particulièrement bas, plus bas que le reste des séquences du groupe ARN. Il est intéressant de noter que cette observation concernant le génome G+C% riche de *R. solanacearum* est à l'oppositions de la tendance des ARNnc dans les organismes A+T riches hyperthermophiles, où les ARNnc se démarquent par leur G+C% plus élevé.

Composition des éléments 5' cis-régulateurs Dans les deux génomes analysés la composition des éléments 5' cis-régulateurs, comprenant les riboswitch, présente une composition différente des autres ARNnc et se rapporoche de la composition des régions codantes. Malgré leur structure secondaire riche les approchant des ARNnc, la composition de ces éléments dans les génomes analysés semble avoir évolué selon le schéma des séquences codantes en amont desquelles elles se trouvent. Une étude plus vaste dans l'ensemble des éléments 5' cis-régulateurs annotés dans les génomes pourrait permettre de conclure s'il s'agit d'une caractéristique universelle de ces éléments.

Surdispersion des données Dans le cas de la modélisation de G+C% dans le génome de *S. aureus* tout comme celle de *R. solanacearum*, à l'aide de la régression logistique, nous avons pu identifier une surdispersion des données par rapport au modèle. Il est possible que la source de cette surdispersion est liée à la corrélation entre les réponses binaires (voir les sources possibles de la surdispersion dans la section 4.1.1.3).

Dans le cas d'une corrélation positive entre les observations nous aurons que la survenue d'un nucléotide X entraîne la survenue plus probable d'un nucléotide Y et elle entraînerait une plus grande variabilité des réponses binaires que celle prédite par le modèle. En effet, on peut imaginer des différens liens existants entre les nucléotides d'une séquence faisant que leur survenue n'est pas indépendante. C'est pourquoi, dans la suite, nous tenterons d'améliorer leur modélisation en tenant compte de la succession des nucléotides.

4.2 Enrichissement du modèle : de G+C% au modèle Markov

Jusqu'ici, nous avons utilisé le G+C% pour discriminer les différents groupes génomiques. L'utilisation de ce critère implique plusieurs hypothèses sur la composition des séquences, que l'on se propose ici de mettre à l'épreuve des données. D'abord, l'utilisation du G+C% fait l'abstraction des éventuelles différences entre la composition en G et en C et entre la composition de A et de T. Or lorsqu'une telle différence existe, il peut être intéressant de l'exploiter pour plus finement caractériser les séquences étudiées. Dans la suite, nous vérifierons cette interchangeabilité du G et C (A et T).

Ensuite, résumer la composition de la séquence au seul G+C% laisse supposer que les nucléotides successifs sont indépendants. On vérifiera si cette hypothèse est vérifiée sur les données en comparant ce modèle à nucléotides indépendants à un modèle markovien permettant de modèliser la succession des nucléotides le long de la séquence.

Notons que le modèle de régression logistique n'a pas pu être utilisé pour modéliser ce problème car, dans le cas de la prise en compte de chacun des nucléotides ou de la dépendance entre les nucléotides, la séquence ne peut plus être modèlisée sous la forme d'une réponse binaire. A la différence des tests effectués dans le cadre de la régression logistique, les tests que nous emploierons ici supposent une distribution homogène au sein de chacun des groupes. Ils ne tiennent pas compte non plus du nombre et de la longueur des séquences au sein de chacun des groupes.

Dans la suite nous présentons ces différentes analyses sur le chromosome de *R. solanacearum*.

4.2.1 Evaluation de la pertinence de l'usage du G+C%

4.2.1.1 Utilisation du G+C% ou de la composition en nucléotides

Nous proposons ici de vérifier si l'utilisation de la distribution de chaque nucléotide est plus pertinente que l'utilisation du G+C%.

On note $\{P_i^{Gr}, i \in \{A,C,G,T\}, Gr \in \{ARN, CODANT, AUTRE\}$ la probabilité d'observer le nucléotide i dans le groupe Gr. Savoir si une perte de l'information a lieu en considérant le G+C% plutôt que la distribution de chaque nucléotide revient à tester les hypothèses suivantes :

$$H0 \quad : \quad P_A = P_T \text{ et } P_C = P_G \tag{4.44}$$

$$\text{contre} \tag{4.45}$$

$$H1 \quad : \quad P_A \neq P_T \text{ ou } P_C \neq P_G \tag{4.46}$$

Remarquons que cela permet aussi de tester l'adéquation de la composition de la séquence à la deuxième loi de Chargaff (Chargaff (1951); Rudner *et al.* (1968); Forsdyke et Mortimer (2000)).

Pour faire ce test, on utilisera à nouveau la théorie des tests des modèles emboités (voir section 4.1.1.4). On note M_0 le modèle associé à l'hypothèse H_0, et M_1 le modèle associé à l'hypothèse alternative H_1 : M_0 est bien un sous-modèle de M_1. La statistique de test sera celle du log-rapport de vraisemblance :

$$T = -2(\log L_0 - \log L_1)$$

où L_0 est la vraisemblance maximale sous l'hypothèse nulle H_0, et L_1 est la vraisemblance maximale sous l'hypothèse alternative. Le nombre de degrés de liberté du modèle M_1 vaut 3 (car la somme des P_i est nulle), et le nombre de degrés de liberté du modèle M_0 vaut 1. Sous H_0, T suit asymptotiquement une loi de χ^2 dont le nombre de degrés de liberté vaut 3-1=2.

Les calculs permettant d'établir les maxima de vraisemblance de chacun de ces modèles sont présentés dans l'annexe B. Soulignons le facteur 1/2 dans les estimateurs du maximum de vraisemblance des paramètres du modèle M_0 obtenus : $\hat{P}_A = \hat{P}_T = \frac{N_A+N_T}{2N}$ et $\hat{P}_C = \hat{P}_G = \frac{N_C+N_G}{2N}$

		$\widehat{P_A}$	$\widehat{P_C}$	$\widehat{P_G}$	$\widehat{P_T}$	p-value
CODANT	M_1	.161	.337	.340	.162	
	M_0	.162	.338	.338	.162	$< 10^{-10}$
AUTRE	M_1	.189	.306	.308	.197	
	M_0	.193	.307	.307	.193	$< 10^{-10}$
ARN	M_1	.214	.284	.275	.227	
	M_0	.220	.280	.280	.220	.005

TAB. 4.9 – Comparaison des taux de GC aux distributions de chacun des nucléotides dans chacun des groupes chez *R. solanacearum* : test de $H_0 : P_A = P_T$ et $P_C = P_G$

Les vraisemblances maximales obtenues sont :

$$\log L_0 = (N_C + N_G) \log \left(\frac{N_C + N_G}{2N} \right) + (N_A + N_T) \log \left(\frac{N_A + N_T}{2N} \right) \quad (4.47)$$

$$\log L_1 = N_A \log \left(\frac{N_A}{N} \right) + N_C \log \left(\frac{N_C}{N} \right) + N_G \log \left(\frac{N_G}{N} \right) + N_T \log \left(\frac{N_T}{N} \right) \quad (4.48)$$

Du fait du nombre important de nucléotides dans chacun des groupes, l'approximation asymptotique de la loi de T par une distribution χ^2 peut être effectuée.

Le test a été effectué sur les données de chaque groupe de séquences. Les résultats sont présentés dans la table 4.9. H_0 est rejeté pour les groupes "ARN", "CODANT" et "AUTRE" (p-value respectives de 5.10^{-3}, et inférieure à 10^{-10}).

4.2.1.2 Discrimination des groupes à l'aide de la distribution en nucléotides

Nous voulons, à présent, vérifier si les distributions en nucléotides obtenues dans chacun des groupes, diffèrent significativement. Cela revient à tester les hypothèses suivantes :

$$H0 : P_U^{ARN} = P_U^{CODANT} = P_U^{AUTRE} \text{ pour tous les nucléotides } U \in \{A, C, G, T\} \quad (4.49)$$

$$H1 : \text{Il existe au moins un groupe où } P_U \text{ diffèrent pour au moins un } U \quad (4.50)$$

Là encore, les deux modèles associés à chacune de ces deux hypothèses sont emboités, et nous utiliserons un test de rapport de vraisemblance.

Afin de s'intéresser plus particulièrement aux groupes ARN, nous avons également testé si la distribution dans ce groupe diffère significativement de la distribution dans

Questions posées	Statistique	d.d.l.	p-value
Au moins un des groupes diffèrent	7960.7	6	$< 10^{-10}$
ARN diffèrent de CODANT	1125.9	3	$< 10^{-10}$
ARN diffèrent de AUTRE	228.2	3	$< 10^{-10}$

TAB. 4.10 – Comparaison des distributions de composition en nucléotides dans les différents groupes chez *R. solanacearum*

chacun des deux autres groupes. Tester si les distributions des nucléotides différent dans les groupes ARN et CODANT revient à formuler les hypothèses suivantes :

$$H0' \quad : \quad P_U^{ARN} = P_U^{CODANT}, \text{ et } P_U^{AUTRE} \text{ quelconque, pour } U \in \{A, C, G, T\} \quad (4.51)$$

$$H1' \quad : \quad P_U^{ARN} \neq P_U^{CODANT} et, P_U^{AUTRE} \text{ pour au moins un } U \in \{A, C, G, T\} \quad (4.52)$$

On adaptera facilement ces hypothèses pour le cas de la comparaison du groupe ARN au groupe AUTRE. Là encore, les modèles correspondant aux hypothèses $H0'$ et $H1'$ sont emboités.

La table 4.10 présente les résultats pour chacun de ces tests. On constate que les différences de composition en nucléotides des groupes ARN, CODANT et AUTRES sont très significatives.

4.2.2 Prise en compte de la dépendance entre nucléotides successifs

Les résultats présentés jusqu'ici supposaient l'indépendance des nucléotides successifs dans la séquence. Dans cette section, on propose de vérifier si cette hypothèse est correcte sur la séquence de *R. solanacearum* ou s'il est préférable de considérer que la présence d'un nucléotide donné dépend des nucléotides le précédant dans la séquence. Pour cela, on considèrera comme alternative à l'indépendance, une dépendance markovienne des nucléotides, c'est à dire que le nucléotide observé à la position i de la séquence dépend directement de celui observé à la position $i - 1$.

Ensuite, la discrimination des différents groupes sera étudiée à l'aide de ce modèle markovien.

4.2.2.1 Présentation de la modélisation markovienne de la séquence

Tout d'abord, nous introduirons les notations, définitions et propriétés qui nous serons utiles pour traiter le problème de la modélisation markovienne des séquences. Des détails complémentaires peuvent être trouvés dans Durbin *et al.* (1998a) ou Robin *et al.* (2003), par exemple.

La séquence est notée $X_1, ..., X_n$ où $X_i = A, C, G$ ou T, est le nucléotide observé à la position i. La modélisation markovienne suppose que X_i ne dépend directement que du nucléotide le précédant, X_{i-1}, c'est-à-dire que :

$$P(X_i|X_1, ..., X_{i-1}) = P(X_i|X_{i-1}) \text{ (propriété de Markov)} \tag{4.53}$$

Le modèle sera donc spécifié par la donnée des probabilités $P(v|u), u \in \{A, C, G, T\}, v \in \{A, C, G, T\}$; ces probabilités sont désignées par le terme de probabilités de transition. Afin de spécifier complètement le modèle, la distribution de la première lettre de la séquence ("qui n'a pas de passé") doit également être indiquée ; cette distribution sera désignée par le terme de *distribution initiale*. Ici, nous avons défini la dépendence markovienne d'ordre 1. Lorsque X_i dépend de k nucléotides le précédant, la propriété de Markov s'écrit de façon analogue et la chaîne de Markov correspondante est dite alors d'ordre k.

Sous certaines conditions, généralement vérifiées lors de la modélisation des séquences d'ADN on peut calculer les probabilités d'observer une lettre donnée, indépendamment de son contexte ; cette probabilité est dénommée *distribution stationnaire* et sera notée $\mu(u), u \in \{A, C, G, T\}$ (pour les conditions de l'existence, voir par exemple Robin *et al.* (2003)). Notons que $\mu(.)$ ne constitue pas un paramètre supplémentaire du modèle, car c'est une fonction déterministe des probabilités de transition. On dira que la séquence est à l'état stationnaire, si X_1 suit la distribution stationnaire.

Sous ce modèle de Markov, la vraisemblance d'une séquence donnée se calcule à l'aide de propriétés élémentaires de probabilités.

$P(X_1, ..., X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)...P(X_n|X_1, ..., X_{n-1})$

Or, $P(X_i|X_1, ..., X_{i-1}) = P(X_i|X_{i-1})$ donc,

$P(X_1, ..., X_n) = P(X_1)P(X_2|X_1)P(X_3|X_2)...P(X_n|X_{n-1})$

Le membre de droite de cette égalité ne fait intervenir que les probabilités de transition et la loi de X_1. Si on suppose que la séquence est à l'état stationnaire, la vrai-

semblance de cette séquence est donnée par :

$$P(X_1, ..., X_n) = \mu(X_1) \times \prod_{u \in \{A,C,G,T\}, v \in \{A,C,G,T\}} P(v|u)^{N_{uv}} \tag{4.54}$$

où N_{uv} désigne le nombre d'occurence du dinucléotide "uv" dans la séquence.

Afin de compléter cette introduction sur la modèlisation markovienne d'une séquence, on précise que l'estimateur du maximum de vraisemblance de $P(v|u)$ est donné par :

$$\widehat{P(v|u)} = \frac{N_{uv}}{N_u} \tag{4.55}$$

4.2.2.2 Utilisation d'un modèle d'indépendance ou d'une dépendance markovienne des nucléotides

On se propose ici d'étudier si les données de la séquence de *R. solanacearum* sont compatibles avec l'hypothèse d'indépendance des nucléotides successifs, dans chacun des trois groupes génomiques définis. Pour cela, nous comparerons les ajustements à un modèle où les nucléotides sont indépendants (modèle noté M_0), à un modèle markovien de succession des nucléotides (modèle noté M_1).

On utilisera, ici encore, la théorie des tests des modèles emboîtés, étant donné que le modèle M_0 est un cas particulier du modèle M_1. Les hypothèses testées sont définies comme suit :

$$H0 \quad : \quad P(v|A) = P(v|C) = P(v|G) = P(v|T) \text{ pour } v \in (A, C, G, T) \tag{4.56}$$

$$H1 \quad : \quad \text{Il existe au moins un couple } (u, u') \text{ tel que } P(v|u) \neq P(v|u') \tag{4.57}$$

La statistique de test utilisée vaudra, comme précédemment, deux fois le log-rapport de vraisemblance de ces deux modèles, le nombre de degrés de liberté du test étant la différence entre les nombres de degrés de liberté de chacun de ces modèles. Pour le modèle M_0, le nombre de degrés de liberté vaut 3, et pour M_1, il vaut $4 \times (4-1) = 12$. Sous H_0, la statistique de test suit donc asymptotiquement une loi de χ^2 à 9 degré de liberté.

Les tests ont été effectués groupe par groupe, et les résultats obtenus sont présentés dans la table 4.11.

On constate que l'hypothèse d'indépendance des nucléotides est largement rejetée pour les groupes CODANT et AUTRE, et est rejetée avec une p-value de $4.^{-5}$ pour le groupe ARN.

Questions posées	Statistique	d.d.l.	p-value
CODANT	154000.1	9	$< 10^{-10}$
AUTRE	15793.3	9	$< 10^{-10}$
ARN	36.0	9	$4.^{-5}$

TAB. 4.11 – Test de l'hypothèse d'indépendance des nucléotides successifs dans la séquence de *R. solanacearum*

4.2.2.3 Discrimination des groupes à l'aide de la distribution markovienne des nucléotides

La section précédente a montré que les données de la séquence de *R. solanacearum* sont plus compatibles avec un modèle prenant en compte une dépendance entre nucléotides successifs qu'un modèle supposant les différents nucléotides indépendant. A présent, nous examinons si les différents groupes ARN, CODANT, AUTRES peuvent être discriminés sur la base d'une modélisation markovienne des séquences.

La démarche utilisée ici sera tout à fait similaire à celle employée et présentée dans la section 4.2.1.2. Le lecteur est invité à s'y référer pour avoir plus de détails sur la démarche suivie ici.

Les hypothèses testés ici sont :

$$H0 \quad : \quad P^{ARN} = P^{CODANT} = P^{AUTRE} \tag{4.58}$$

$$H1 \quad : \quad \text{Il existe au moins un groupe où } P^G \text{ différent} \tag{4.59}$$

où P^G désigne l'ensemble des paramètres du modèle de markov du groupe G.

Afin de s'intéresser plus particulièrement au groupe ARN, nous avons également testé si la distribution dans ce groupe diffère significativement de la distribution dans chacun des deux autres groupes. Tester si la distribution des nucléotides diffère dans les groupes ARN et CODANT revient à formuler les hypothèses suivantes :

$$H0' \quad : \quad P^{ARN} = P^{CODANT}, \text{ et } P^{AUTRE} \text{ quelconque} \tag{4.60}$$

$$H1' \quad : \quad P^{ARN} \neq P^{CODANT}, \text{ et } P^{AUTRE} \text{ quelconque} \tag{4.61}$$

On adaptera facilement ces hypothèses pour le cas de la comparaison du groupe ARN au groupe AUTRES.

La table 4.12 présente les matrices d'émission estimées par maximum de vraisemblance (équation 4.55) dans chacun des trois groupes. La table 4.13 présente les résultats

	ARN					CODANT					AUTRES			
	A	C	G	T		A	C	G	T		A	C	G	T
A	0.204	0.294	0.281	0.221	A	0.161	0.304	0.305	0.230	A	0.260	0.256	0.248	0.235
C	0.211	0.294	0.276	0.218	C	0.191	0.255	0.411	0.143	C	0.201	0.281	0.365	0.154
G	0.214	0.266	0.290	0.230	G	0.172	0.421	0.259	0.148	G	0.191	0.364	0.282	0.163
T	0.229	0.286	0.248	0.237	T	0.068	0.365	0.401	0.165	T	0.106	0.299	0.330	0.266

TAB. 4.12 – Matrices d'émission estimée dans chacun des trois groupes : ARN, CODANT, et AUTRES.

Questions posées	Statistique	d.d.l.	p-value
Au moins un des groupes diffèrent	17473.3	24	$< 10^{-10}$
ARN diffèrent de CODANT	1114.5	12	$< 10^{-10}$
ARN diffèrent de AUTRE	2557.7	12	$< 10^{-10}$

TAB. 4.13 – Comparaison des distributions de composition en nucléotides dans les différents groupes chez *R. solanacearum* en considérant l'existance d'une dépendance markovienne entre les nucléotides successifs.

pour chacun de ces tests. On constate que les différences de composition en nucléotides des groupes ARN, CODANT et AUTRES sont très significatives lorsque la dépendance markovienne des nucléotides le long de la séquence est prise en compte.

4.2.3 Discussion

L'analyse des données de *R. solanacearum* montre que l'utilisation du GC% pour décrire les séquences des trois groupes génomiques n'est pas équivalente à l'utilisation de chaque nucléotide séparément, qui n'est, à son tour, pas équivalente à l'utilisation des di-nucléotides (ou une dépendance d'ordre 1) pour décrire cette composition.

Sous l'hypothèse d'une distribution homogène dans les groupes, nous montrons que l'utilisation de la composition en nucléotides et en di-nucléotides permet de discriminer les trois groupes génomiques. Or, nous avons vu (section 4.1.3) que l'hypothèse d'homogénéité au sein des groupes n'est pas vérifiée dans le génome de *R. solanacearum*. Néanmoins, la technique des GLM n'étant pas adaptée pour l'analyse de la composition en nucléotides ou en di-nucléotides, nous n'avons pas effectué une analyse prenant en compte la non-homogénéité dans les groupes. De ce fait, les conclusions de la présente analyse restent partielles.

Dans la section 5.3, nous les mettons à l'épreuve des données génomiques de *R. solanacearum* en utilisant un modèle d'indépendance des nucléotides et un modèle de dépendance markovienne d'ordre 1, *via* les modèles de Markov cachés, après avoir introduit les notions de base sur ces modèles.

Chapitre 5

Segmentation du génome sur la base de la composition en nuclétides

Les résultats présentés dans le chapitre précédent (notamment, dans la section 4.1.2.3) suggèrent que, lorsque la variabilité dn composition au sein des groupes est prise en compte, la discrimination entre les ARNnc et le reste du génome, sur la seule base de G+C%, paraît impossible dans le génome de *R. solanacearum*. Nous avons donc proposé une modélisation plus riche, en nucléotides ou en di-nucléotides.

Les résultats exposés dans la partie consacrée à la modélisation markovienne montrent qu'une telle modélisation diffère de la modélisation du G+C%.

Nous avons montré également que, sous l'hypothèse d'homogénéité au sein du groupe, une discrimination entre les groupes génomiques est possible. Même si les groupes ne sont pas homogènes, les *p-valeurs* des tests obtenues, confortent la possibilité de discrimination entre les groupes.

Dans le chapitre précédent nous avons donc établi l'existence d'un biais de composition au sein des ARNnc de *R. solanacearum*, du moins lorsque la nous supposons que la composition au sein des groupes est homogène. Dans le présent chapitre nous proposons de l'utiliser pour la détection des ARNnc. Dans ce contexte, le problème de la détection des ARNnc s'apparente au problème de segmentation du génome sur la base de l'existence d'un biais de composition. Pour cela, nous utiliserons la modélisation par chaînes de Markov cachées.

Dans la suite, nous présentons d'abord la notion de chaînes de Markov cachées ainsi que la modélisation des séquences génomiques par ces modèles. Une fois ces notions

familières, nous introduisons plus formellement la définition des chaînes de Markov cachées. Nous présentons les algorithmes de reconstruction des chemins cachés, en portat une attention particulière à l'algorithme de Viterbi. Une étude par simulations analysera sa pertinence dans le cadre de la segmentation utilisant le biais de composition dans les séquences nucléiques. Enfin, nous présenterons les résultats de segmentation pour le génome de *R. solanacearum*.

5.1 Modèles de Markov cachés

Un modèle de Markov caché modèlise deux processus emboîtés : un processus observable et un processus caché. Le processus caché est modélisé par une chaîne de Markov homogène d'ordre 1 et ses réalisations sont appelées les *états cachés*.

Pour expliquer l'idée derrière les modèles de Markov cachés, Rabiner (1989) propose l'exemple d'un système de K urnes $\{S_1, S_2, \ldots, S_K\}$ contenant chacune un grand nombre de balles de M différentes couleurs $\{X_1, \ldots, X_M\}$. On connaît le nombre de balles de chaque couleur dans chacune des urnes, $b_s(X)$ étant la proportion des balles de couleur X dans l'urne s. Quelqu'un choisit une urne initiale au hasard et tire une balle dans cette urne. Nous ne pouvons pas voir quelle urne a été choisie. La couleur de la balle est ensuite enregistrée. L'urne suivante est choisie selon un processus aléatoire et ce choix reste caché. Toute la procédure est ensuite répétée n fois, générant une séquence de n balles observées.

Lorsque le choix des urnes est dicté par un modèle de Markov homogène, le processus décrit est l'exemple d'une chaîne de Markov cachée. Les urnes représentent les états cachés et la suite des couleurs des balles tirées représente le processus observé. Cet exemple illustre deux aspects du processus stochastique des chaînes de Markov cachées : les états cachés et les observations (respectivement les urnes et les couleurs dans notre exemple).

De façon générale, un modèle de Markov caché est caractérisé par (Rabiner (1989)) :

1. L'ensemble d'états cachés du modèle. Dans les applications des modèles de Markov cachés, les états cachés ont une signification liée aux objets du problème étudié (dans l'exemple précédent les états cachés correspondent aux urnes).

2. L'ensemble observable dit aussi *alphabet* (les couleurs, dans l'exemple précédent).

3. L'ensemble de probabilités de passer d'un état caché à l'autre.Le modèle suppose une transition markovienne entre les états, c'est-à-dire que le choix du futur état ne dépend directement que de l'état présent.
4. Les probabilités des observations dans les états cachés. Dans les exemples, ces probabilités correspondent aux fréquences des balles de chacune des couleurs dans chacune des urnes.
5. La distribution initiale sur les états. En effet, dans l'exemple des urnes, la première urne est choisie selon un processus aléatoire (indépendant du passé qui n'existe pas encore) et dans un modèle formel, ce processus doit être défini.

Trois questions fondamentales doivent être résolues pour l'utilisation pratique des modèles de Markov cachés (Rabiner (1989)) :

1. Problème 1 : Etant donné une séquence d'observations O et un modèle de Markov caché λ, comment calculer la probabilité de la séquence observée sous le modèle λ, $\mathbb{P}(O|\lambda)$.
2. Problème 2 : Etant donné une séquence d'observations O et un modèle de Markov caché λ, comment choisir la séquence d'états cachés optimale pour un critère donné qui "explique" le mieux les données. Ce problème est connu sous le nom de *reconstruction du chemin caché* ou *segmentation*.
3. Problème 3 : Pour une séquence d'observations O et un modèle λ comment trouver les paramètres de λ qui maximisent $\mathbb{P}(O|\lambda)$.

Les solutions classiques à ces questions seront décrites au fur et à mesure dans le texte.

Notons également que les modèles de Markov cachés sont souvent désignés par le nom HMM de l'anglais Hidden Markov Model. Nous utiliserons cette appellation abrégée dans la suite du texte.

Représentations graphiques des modèles de Markov cachés Les HMM sont communément représentés par les graphes de dépendances entre les états cachés. Nous introduisons ici cette représentation, qui sera utilisée par la suite.

Graphe des états cachés

Le modèle des états cachés est un modèle de Markov homogène d'ordre 1. Ce modèle peut être présenté sous la forme d'un graphe dont les sommets correspondent aux états cachés et les arêtes correspondent aux transitions entre les différents états.

La figure 5.1 présente le graphe des états cachés d'un modèle HMM à 3 états. Les probabilités de transition d'un état à un autre sont données sur les arcs du graphe.

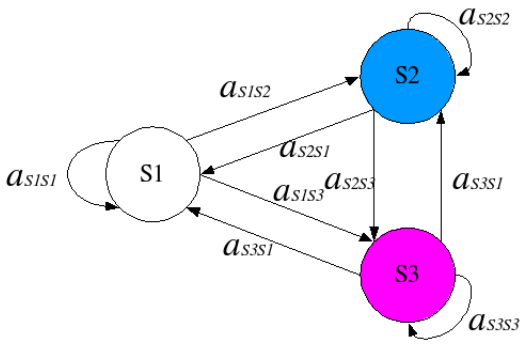


FIG. 5.1 – Graphe des états d'un modèle de Markov à 3 états. Un arc allant du sommet S_i au sommet S_j représente la possibilité d'une transition de l'état S_i à l'état S_j et les étiquettes $a_{S_iS_j}$ représentent les probabilités de transition correspondantes.

Graphe de la séquence d'états cachés

Le graphe des états cachés (fig. 5.1) n'explicite pas l'aspect "temporel" de la séquence cachée. Cet aspect des HMM est représenté par le graphe de la séquence des états cachés (fig. 5.2), visualisant les transitions possibles entre les états cachés le long de la séquence observée.

Notons que le graphe des états cachés ou de la séquence des états cachés, définit ce que l'on appelle la topologie du modèle. En effet, ces graphes n'explicitent pas les probabilités d'émission associées aux différents états cachés, mais la nature des transitions entre les différents états. Ainsi, ils ne suffisent pas à spécifier entièrement un modèle HMM.

Séquence générée Un exemple de la séquence d'observations générée par un HMM décrit dans la figure 5.1, dont les probabilités de transition et d'émission sont spécifiées dans le graphique 5.3 (respectivement en A) et B)) sur un alphabet $X = \{a, c, g, t\}$ est représenté en (C) de la figure 5.3). Lorsqu'on parle des plages d'un état caché d'une chaîne de Markov cachée on se réfère aux blocs successifs correspondant chacun à un état différent.

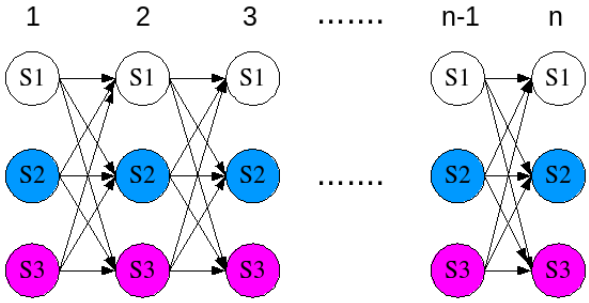


FIG. 5.2 – Le graphe de la séquence des états cachés d'un modèle de Markov caché à 3 états. A chaque position de 1 à N, la séquence observée peut se trouver dans l'un des trois états cachés. Les transitions possibles entre les états de la position i à la position j sont représentées par les flèches.

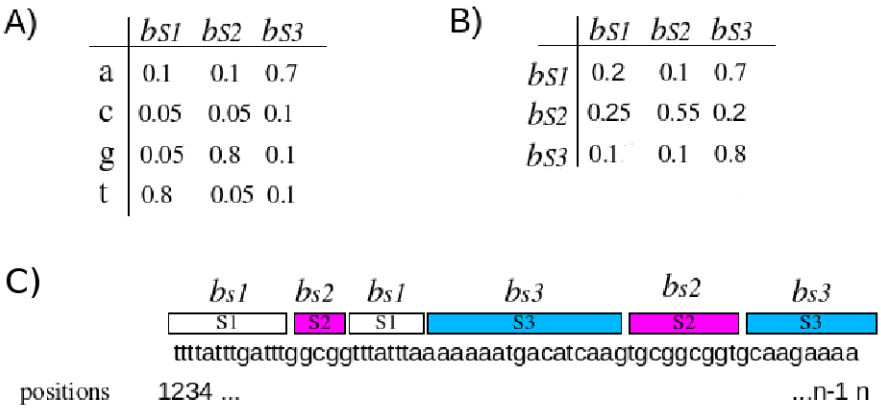


FIG. 5.3 – Exemple de séquence générée par un HMM. A) Probabilités d'émission B) Probabilités de transition C) Suite des états cachés (non observés) et séquence générée.

5.1.1 Modélisation des séquences génomiques par les modèles de Markov cachés

Les modèles de Markov cachés et les modèles dérivés sont largement employés pour modéliser les séquences génomiques. Parmi leurs applications les plus connues figurent la détection de gènes codant pour des protéines (Burge et Karlin (1997a)), la prédiction de l'appartenance à une famille protéique (Bateman *et al.* (1999)), la détection de promoteurs (Ohler *et al.* (2001)), la prédiction de la structure secondaire des protéines (Krogh *et al.* (2001)), la détection d'îlots CpG (Durbin *et al.* (1998b)) ou encore la recherche des membres des familles ARNnc (Durbin *et al.* (1998b)). Dans le domaine de la prédiction des ARNnc, Klein *et al.* (2002) ont utilisé avec succès les HMM pour la détection des ARNnc chez les archées A+T-riches hyperthermophiles et Tjaden (2007) propose l'utilisation d'un HMM comme élément d'un modèle d'integration de données hétérogènes (voir chapitre 3).

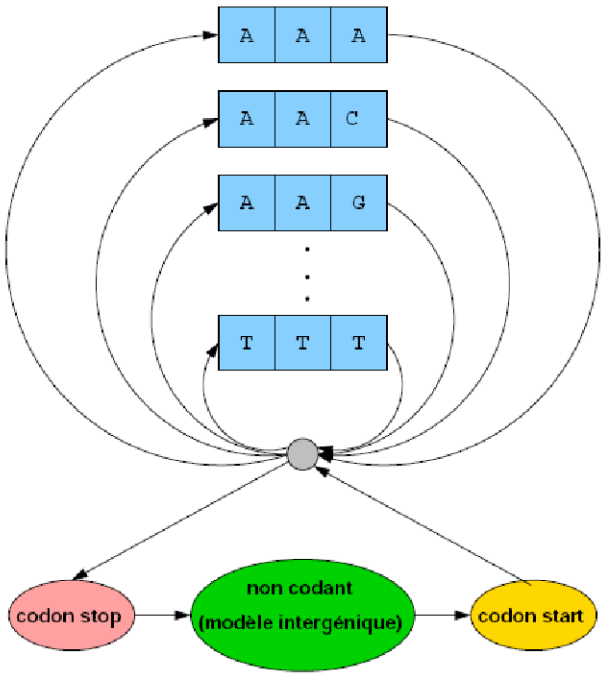


FIG. 5.4 – Un détecteur de gènes HMM "simple" conçu pour le génome d'*E.coli* (Krogh *et al.* (1994), image reprise de http ://www2.lifl.fr/ touzet/DESSBI/Cours/genes.pdf). Les probabilités de transition ne sont pas indiquées. Chaque codon se voit attribuer une probabilité d'émission estimée.

La figure 5.4 donne le graphe des états cachés du modèle utilisé par Krogh *et al.* (1994) pour la détection de gènes dans le génome d'*E. coli*. Plusieurs aspects de l'utilisation des HMM dans la modélisation des séquences génomiques y sont illustrés. D'abord, chaque état n'est pas relié à tous les autres, ce qui signifie qu'il y a des transitions interdites. Par exemple, il n'est pas possible d'avoir le premier nucléotide du codon start après le dernier nucléotide du codon stop ; ces deux codons sont supposés être séparés par une région intergénique. Ensuite, certaines transitions sont "obligatoires" (de probabilité 1) : par exemple, pour passer dans l'état "intergénique" on est obligé de passer par l'état correspondant au codon stop.

5.1.2 Définition et notations

Dans cette section, la présentation s'appuit sur la thèse de P. Nicolas (Nicolas (2003)). Nous présentons ici la définition formelle d'un HMM dans le cas d'un nombre fini d'états cachés.

5.1.2.1 Définition du processus caché

Comme dit précédemment un HMM modélise deux processus emboîtés le processus caché et le processus observé.

Soit $\mathcal{S}$ un ensemble fini, $|\mathcal{S}| = K$, appelé *l'ensemble des états cachés*. Le premier processus, dit le processus caché $S_1, \ldots, S_N$, est un modèle de Markov homogène d'ordre 1, dont les réalisations sont des membres de l'ensemble $\mathcal{S}$. Nous noterons $\{a(u,v)\}_{u,v\in S}$ ses probabilités de transition et $\{a_0(u)\}_{u\in S}$ sa distribution initiale. Nous avons donc, pour tout $u, v \in S$:

$$\begin{cases} a(u,v) &= \mathbb{P}(S_{n+1} = v | S_n = u) \quad \text{(propriété d'indépendance conditionnelle des états cachés)} \\ a_0(u) &= \mathbb{P}(S_1 = u) \qquad\qquad\qquad\qquad\qquad\qquad \text{(distribution initiale)} \end{cases}$$

Rappelons que nous avons, pour tout $u \in S$, $\sum_{v\in S} a(u,v) = 1$ et $\sum_{u\in S} a_0(u) = 1$.

Ainsi, la probabilité d'apparition d'une suite (s_1^N) particulière est :

$$\begin{aligned} \mathbb{P}(S1 = s_1, \ldots, S_N = s_N) &= \mathbb{P}(S_N = s_N | S_{N-1} = s_{N-1}) \cdot \mathbb{P}(S_{N-1} = s_{N-1} | S_{N-2} = s_{N-2}) \\ &\quad \cdots \mathbb{P}(S_2 = s_2 | S_1 = s_1) \cdot \mathbb{P}(S1 = s_1) \\ &= a_0(s1) \cdot \prod_{i=1,\ldots,N-1} a(s_i, s_{i+1}) \end{aligned}$$

Longueurs des plages d'états Nous pouvons à présent calculer la probabilité d'apparition, dans le processus caché, d'un même état s, l fois de suite.

$$\mathbb{P}(S_{i+1} = s, \ldots, S_{i+l} = s, S_{i+l+1} \neq s | S_i = s) = a(s,s)^l \cdot (1 - a(s,s)) \tag{5.1}$$

La probabilité d'observer une plage d'état s de longueur l suit donc une loi géométrique de paramètre $1 - a(s,s)$.

Si on note L_s la variable aléatoire associée à la longueur de la plage d'état s, son espérance vaut

$$E(L) = \frac{1}{1 - a(s,s)} \tag{5.2}$$

Cette propriété permet d'interpréter les probabilités de transition en terme de longueurs des plages générées.

Remarque Dans une séquence issue d'un HMM , la longueur des plages de chacun des états cachés du modèle suit une loi géométrique. Cette propriété inhérente aux HMM a présenté un obstacle à l'application de HMM dans la modélisation des génomes, étant donné que leurs longueurs ne suivent pas a loi géométrique.

Plusieurs modèles alternatifs ont été introduits afin de contourner cet écueil et ils ont été intégrés à la plupart des détecteurs de gènes du type HMM tel que Glimmer ou Genscan par exemple (Burge et Karlin, 1997b; Salzberg *et al.*, 1998; Delcher *et al.*, 1999)).

5.1.2.2 Définition du processus observé

Le deuxième processus, le processus observé, correspond à la séquence observée. Soit $\mathcal{X}$ (avec $|\mathcal{X}| = M$) *l'alphabet* du HMM. Le processus observé est défini comme $X_1, \ldots, X_n$, où X_i est à valeurs dans l'ensemble $\mathcal{X}$. Chaque observation, dite aussi *émission*, de ce processus est associée à un état caché et dépend de cet état caché, qui est qualifié d'*état sous-jacent*.

Les lois d'émissions conditionnellement à l'état caché sous-jacent peuvent être fixées librement en fonction des problèmes et des données traitées. Il est également possible d'introduire une dépendance directe entre les observations, c'est-à-dire de supposer que la loi de chaque observation ne dépend pas seulement de l'état caché, mais également des autres observations précédentes : il est courant, par exemple, de supposer également un

modèle markovien de succession des observations, conditionnellement aux états cachés. Ici, nous présenterons les cas de l'indépendance des observations conditionnellement à l'état caché, et de dépendance markovienne d'ordre 1 entre les observations conditionnellement à l'état caché. Nous utiliserons ces deux modèles par la suite.

Cas d'indépendance conditionnellement à l'état caché La distribution de la séquence $X_1, \ldots, X_n$ observée est décrite conditionnellement au processus caché à l'aide des lois d'émission $\{b_s\}_{s\in\mathcal{S}}$:

$$b_s(x) = \mathbb{P}(X_i = x | S_i = s) \tag{5.3}$$

où $x \in \mathcal{X}$ et $\sum_{x\in\mathcal{X}} b_s(x) = 1$.

On se référera à ce modèle comme au modèle $M1M0$, $M1$ faisant référence au modèle de Markov d'ordre 1 correspondant au processus caché et $M0$ au modèle d'indépendance des observations conditionnellement à l'état caché.

Cas de dépendance markovienne Une généralisation du modèle du processus observé décrit ci-dessus est de considérer que les observations exhibent une dépendance markovienne conditionnellement au processus caché. Nous donnons ici la définition dans le cas où la dépendance markovienne est du même ordre pour tous les états cachés.

Si la dépendance markovienne est d'ordre r alors les lois d'émissions conditionnellement au processus caché, $\{b_s\}_{s\in\mathcal{S}}$ s'écrivent comme :

$$\begin{cases} b_s(w,x) &= \mathbb{P}(X_i = x | S_i = s, X_{i-1}\ldots X_{i-r} = w), i > r & \text{(propriété de Markov conditionnelle} \\ b_{0s}(w,x) &= \mathbb{P}(X_i = x | S_i = s, X_1 \ldots X_{i-1} = w), 1 < i \leq r & \text{(distribution initiale)} \\ b_0(x) &= \mathbb{P}(X_1 = x) & \text{(distribution initiale)} \end{cases}$$

Dans le cas de dépendance markovienne d'ordre r on parlera de modèle $M1Mr$.

La probabilité d'observer $(x_1^N) = (x_1, \ldots, x_N)$ conditionnellement au processus caché $(s_1^N) = (s_1, \ldots, s_N)$ s'écrit comme :

$$\begin{aligned}\mathbb{P}(X_1^N = x_1^N | S_1^N = s_1^N) &= \mathbb{P}(X_1 = x_1) \prod_{i=2}^{r} \mathbb{P}(X_i = x_i | S_i = s_i, X_2^{i-1} = x_2^{i-1}) \\ &\quad \cdot \prod_{i=r+1}^{N} \mathbb{P}(X_i = x_i | S_i = s_i, X_{i-r}^{i-1} = x_{i-r}^{i-1}) \\ &= b_0(x_1) \prod_{i=2}^{r} b_{0s_i}(x_1 ... x_{i-1}, x_i) \prod_{i=r+1}^{N} b_{s_i}(x_i, x_{i-r}^{i-1})\end{aligned}$$

Dans la suite, on notera une chaîne de Markov cachée $(S_1^N = s_1^N, X_1^N = x_1^N)$ de longueur N et $\lambda(\mathcal{S}, \mathcal{X}, \theta)$ le modèle correspondant, où $\mathcal{S}$ est l'ensemble des états cachés, $\mathcal{X}$ est l'alphabet de la séquence observée et, $\theta = (a, b)$ est l'ensemble des paramètres.

5.1.2.3 Vraisemblance d'une séquence sous le modèle de Markov caché

La vraisemblance des paramètres θ sur le processus $(S_1^N = s_1^N, X_1^N = x_1^N)$ pour un modèle λ est

$$\begin{aligned}\mathbb{P}_\theta(S_1^N = s_1^N, X_1^N = x_1^N) &= \mathbb{P}_\theta(S_1^N = s_1^N) \mathbb{P}_\theta(X_1^N = x_1^N | S_1^N = s_1^N) \\ = \left(a_0(s1) b_0(x_1) \prod_{i=2}^{r} b_{0s_i}(x_1 ... x_{i-1}, x_i) \right) &\cdot \left(\prod_{i=1}^{N-1} a(s_i, s_{i+1}) \cdot b_{s_i}(x_{i+1-r}^{i}, x_{i+1}) \right)\end{aligned}$$

Une estimation de l'ensemble des paramètres θ peut être obtenue en maximisant cette vraisemblance.

5.1.3 Estimation des paramètres

L'estimation des paramètres d'un HMM à partir d'une chaîne observée correspond au troisième problème posé par Rabiner (section 5.1, (Rabiner, 1989)). Nous présentons ici la méthodologie exposée dans Durbin *et al.* (1998a).

On distingue deux cas selon que les états cachés sont connus ou non. En effet, le problème n'est pas le même si l'on dispose uniquement de la séquence observée, et que l'on cherche à estimer l'ensemble des paramètres du HMM qui a généré cette séquence, ou bien si l'on dispose de la séquence observée et également de la séquence des états cachés lui correspondant. Ce deuxième cas de figure se produit si l'on dispose par exemple d'un ensemble d'apprentissage permettant de caractériser chacun des états cachés et leur transitions.

Le nombre de paramètres à estimer augmente avec la complexification du modèle. Pour chaque état caché on doit estimer les probabilités d'émission correspondantes, à savoir $|\mathcal{X}|^r(|\mathcal{X}|-1)$ paramètres. Dans le cas général, où toutes les transitions sont autorisées, nous devons estimer $|\mathcal{S}|(|\mathcal{S}|-1)$ paramètres de transition. Ainsi, le nombre total de paramètres à estimer en général est $|\mathcal{S}|(|\mathcal{S}|-1)+|\mathcal{S}||\mathcal{X}|^{r-1}(|\mathcal{X}|-1)$, sans compter les paramètres initiaux.

Dans un souci de clarté de l'exposé, les résultats présentés dans la suite le seront dans le cadre d'un modèle $M1M1$. La généralisation au cas $M1Mr$ ou $M1M0$ se fera facilement.

5.1.3.1 Estimation lorsque la séquence d'états cachés est connue

C'est le cas le plus simple et les résultats obtenus sont les plus intuitifs. En effet, si les états cachés sont connus, on imagine facilement pouvoir caractériser les lois d'émissions pour chaque état caché, ainsi que les probabilités de transition entre états cachés en nous reposant sur la propriété markovienne de cette séquence.

En effet, l'estimateur de maximum de vraisemblance des paramètres d'une chaîne de Markov S repose sur les comptages des occurences du mot s_1s_2 (mot composé des états cachés) pour deux états cachés s_1 et s_2 ; ce comptage sera noté $N^s_{s_1,s_2}$. L'estimateur de maximum de vraisemblance des probabilités de transition est alors donné, pour $(s1,s2)\in\mathcal{S}^2$, par :

$$\widehat{a_{s1s2}} = \frac{N^s_{s_1,s_2}}{\sum_{s_2\in\mathcal{S}} N^s_{s_1,s_2}}$$

De même, l'estimateur de maximum de vraisemblance des probabilités d'émission repose sur la théorie des chaînes de Markov. On obtient, pour les observations $(u,v)\in\mathcal{X}^2$ et les états cachés $s\in\mathcal{S}$:

$$\widehat{b_s(u,v)} = \frac{N_s(u,v)}{\sum_{v\in\mathcal{X}} N_s(u,v)}$$

où $N_s(u,v)$ est le nombre d'occurences du mot uv dans les séquences observées dans l'état caché s.

5.1.3.2 Estimation lorsque la séquence d'états cachés n'est pas connue

La maximisation de la vraisemblance lorsque les états cachés ne sont pas connus est plus complexe, et il n'y a pas de forme analytique simple des estimateurs dans ce

cas. La maximisation se fait au travers d'une démarche itérative, faisant intervenir la reconstruction des états cachés pour un jeu de paramètres donné, et l'estimation des paramètres pour un ensemble d'états cachés donné.

L'algorithme le plus utilisé pour estimer les paramètres d'un HMM dans ce cas, est l'algorithme "Expectation-Maximisation" (EM, Dempster *et al.* (1977)). Cet algorithme se nomme également Baum-Welch dans le cadre des HMM. Il s'agit d'un algorithme itératif, qui converge vers un maximum local de la vraisemblance. On se donne donc une valeur initiale pour les paramètres λ^0 et chaque nouvelle itération m conduit à produire un nouveau jeu de paramètres λ^m tel que $\mathbb{P}_{\lambda^m}(X_1^n = x_1^n) > \mathbb{P}_{\lambda^{m-1}}(X_1^n = x_1^n)$.

Une itération est constituée de deux étapes. La première étape, dite étape "E" consiste en un calcul d'espérance conditionnelle : $Q(\lambda|\lambda^{m-1}) = E_{\lambda^{m-1}}(\log \mathbb{P}_\lambda(x_1^n, s_1^n)|x_1^n)$. La deuxième étape, dite étape "M", consiste à trouver la nouvelle valeur des paramètres λ^m qui maximise cette quantité.

Le détail des calculs dépasse le cadre de cette thèse et peuvent être trouvés, par exemple, dans les ouvrages suivants : Nicolas (2003); Durbin *et al.* (1998b).

5.1.4 Segmentation des séquences : la reconstruction du chemin caché

Les méthodes de reconstruction du chemin caché répondent au deuxième problème énoncé par Rabiner (1989), décrit dans la section 5.1, à savoir : étant donné une séquence d'observations $O = o_1 o_2 \dots o_N$ et un modèle de Markov caché λ, comment choisir la séquence des états cachés correspondante, optimale pour un certain critère, ou qui "explique" le mieux les données ?

Dans une partie des problèmes modélisés par les HMM, l'utilisation du modèle se fait justement à travers la recherche du chemin caché "optimal" (par exemple, les problèmes de reconnaissance des formes ou de parole, voir Rabiner (1989)).

Pour cette raison, le problème de reconstruction du chemin caché apparaît comme fondamental dans l'utilisation pratique des HMM.

Les algorithmes de segmentation s'emploient à trouver le chemin caché le plus probable, c'est-à-dire, maximisant la probabilité jointe du processus observé et du processus caché (Durbin *et al.* (1998b)) :

$$s_1^N = \arg\max \mathbb{P}(s_1^N, x_1^n) \tag{5.4}$$

On se place ici dans le cadre, généralement utilisé, où les paramètres du modèle sont supposés connus.

La difficulté provient d'un nombre élevé de chemins possibles : dans le cas général de l'autorisation de toutes les transitions, il existe $|\mathcal{S}|^N$ chemins différents pour un modèle à $|\mathcal{S}|$ états cachés et une séquence de taille N. Leur énumération exhaustive n'est donc généralement pas calculable en temps réel.

Les algorithmes de reconstruction du chemin caché contournent cet écueil en tirant profit de la propriété de l'indépendance conditionnelle des états cachés (décrit dans 5.1.2.1). Deux algorithmes, l'algorithme de Viterbi et l'algorithme dit Forward-Backward, sont les plus utilisés pour cette tâche (Durbin *et al.* (1998b)). Le premier adopte le point de vue de la maximisation globale de la suite d'états cachés reconstruite en cherchant à maximiser la probabilité de la séquence entière, tandis que le deuxième suit un raisonnement plus local en maximisant la probabilité à chaque position sans garantir l'optimalité globale du chemin reconstruit.

Nous présenterons d'abord l'algorithme de Viterbi et l'algorithme Forward-Backward. Ensuite, nous présenterons une étude des popriétés de l'algorithme de Viterbi et les conditions de son utilisation, qui a été menée au cours de la thèse. Pour simplifier les notations et la lecture du document, on se place dans le cadre d'un modèle $M1M0$.

5.1.4.1 Algorithme de Viterbi

L'algorithme de Viterbi est un algorithme de recherche du plus court chemin dans un graphe sans cycle. Il reconstruit le chemin caché d'une séquence en recherchant le chemin caché le plus probable :

$$s_1^N = \arg\max_{s_1^N} \mathbb{P}(s_1^N, x_1^N) \tag{5.5}$$

En exploitant les propriétés de l'indépendance conditionnelle des états cachés d'une part, et des observations d'autre part (sections 5.1.2.1 et 5.1.2.2), la relation suivante peut être établie entre la probabilité jointe du chemin caché et de la séquence observée

à la position i et cette même probabilité à la position $i-1$:

$$\begin{aligned}
\mathbb{P}(s_1^i, x_1^i) &= \mathbb{P}(x_i|s_1^i, x_1^{i-1})\mathbb{P}(s_1^i, x_1^{i-1}) && (5.6)\\
&= \mathbb{P}(x_i|s_1^i, x_1^{i-1})\mathbb{P}(s_i|s_1^{i-1}, x_1^{i-1})\mathbb{P}(s_1^{i-1}, x_1^{i-1}) && (5.7)\\
&= \mathbb{P}(x_i|s_i, x_{i-r}^{i-1})\mathbb{P}(s_i|s_{i-1})\mathbb{P}(s_1^{i-1}, x_1^{i-1})\\
&= b_{s_i}(x_{i-r}^{i-1}, x_i)a(s_{i-1}, s_i)\mathbb{P}(s_1^{i-1}, x_1^{i-1})
\end{aligned}$$

Si on définit $P_i(s)$ comme

$$P_i(s) = \max_{s_1^{i-1} \in \mathcal{S}} \mathbb{P}(S_i = s, s_1^{i-1}, x_1^i) \tag{5.8}$$

alors, à l'aide de l'équation 5.6, nous pouvons déduire, entre $P_i(s)$ et $P_{i-1}(q)$,$q, s \in \mathcal{S}$, la relation de récurence suivante :

$$P_i(s) = \max_{s_1^{i-1} \in \mathcal{S}} (b_s(x_{i-r}^{i-1}, x_i)a(q, s)P_{i-1}(q)) \tag{5.9}$$

Afin de reconstruire le chemin caché, l'algorithme de Viterbi procède en deux étapes : la récurrence en avant et l'étape en arrière (schématisé sur la fig. 5.5).

Pour l'étape en avant, la valeur initiale P_1, définie comme $P_1 s = a_0(s)b_s(x_1)$ pour tout $s \in \mathcal{S}$ est calculée. Ensuite, les positions de 1 à N sont parcourues en calculant pour chaque position i et pour chaque état caché s la valeur de $P_i(s)$ à l'aide des P_{i-1} déjà calculés. Un pointeur est gardé de l'état s à la position i à l'état q à la position $i-1$, où q maximise $P_i(s)$ de la récurrence 5.9.

Pour l'étape en arrière, dite aussi de traceback, à partir de l'état $\hat{s}$ à la position N, maximisant $P_N(s)$, le chemin jusqu'à la position 2 est reparcouru en arrière où $\hat{s}_{i-1}$ prend la valeur de l'état pointée par $\hat{s}_i$, calculée dans l'étape en avant.

Complexité A chaque position i parcourue lors de l'étape en avant les valeurs de $P_i(s, q)$ sont calculés pour $s, q \in \mathcal{S}$. Dans l'étape en arrière la séquence est parcourue une fois. Par conséquent la complexité est de l'ordre $N \cdot |\mathcal{S}|^2$ pour une séquence de taille N.

5.1.4.2 Algorithme Forward-Backward

Une autre approche de la reconstruction du chemin caché d'une chaîne de Markov cachée est l'algorithme *Forward-Backward*. A la différence de l'algorithme de Viterbi, qui calcule le chemin le plus probable globalement, l'algorithme Forward-Backward

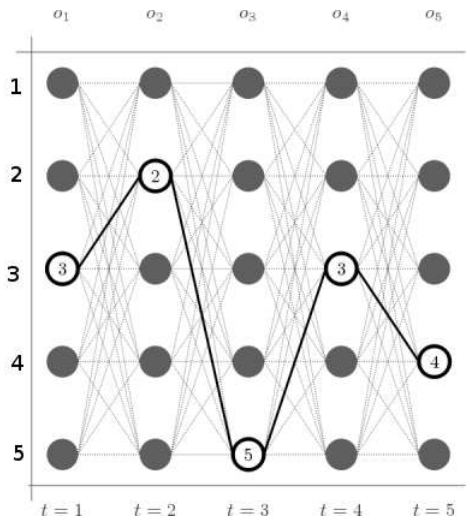


FIG. 5.5 – Le principe de l'algorithme de Viterbi pour une chaîne de Markov cachée à 5 états (représentés par les cercles). Les positions sont notées par t et les observations par o_t. Les $P_t(s)$ sont calculés à chaque position t et sont représentées par les arêtes en gris clair (ce procédé corespond à la réccurence en avant). Le chemin reconstruit (correspondant à l'étape en arrière) est $\hat{s}_1^5$ est $3-2-5-3-4$ (les arêtes entourés en noir). Repris de http ://upload.wikimedia.org/wikipedia/commons/7/71/Hmm-Viterbi-algorithm-med.png

attribue la valeur la plus probable de l'état caché à chaque position. Autrement dit, l'algorithme Forward-Backward reconstruit le chemin caché $\hat{S}_1^n = \hat{S}_1 \dots \hat{S}_n$ tel que pour tout $1 \leq i \leq n$:

$$\hat{S}_i = \arg\max_{k \in \mathcal{S}} \mathbb{P}(S_i = k | x_1^n) \tag{5.10}$$

où x est la séquence observée.

Le calcul de ces probabilités repose sur deux équations de récurrence, et l'algorithme est constitué de deux étapes, classiquement désignés sous le nom d'étape "Forward" et d'étape "Backward". Le lecteur intéressé aux détails de ces algorithmes est invité à consulter un des nombreux ouvrages à ce sujet (Durbin *et al.* (1998a), Robin *et al.* (2003), Nicolas *et al.* (2004)).

Notons qu'un procédé similaire permet de répondre à la première question énoncée par Rabiner, à savoir calculer la probabilité d'une séquence observée sous un modèle donné.

5.2 Etude d'une propriété de la reconstruction par algorithme de Viterbi

5.2.1 Introduction

L'algorithme de Viterbi est efficace et simple d'utilisation. Pour cette raison, il est communément utilisé en bioinformatique pour les tâches telles que la détection des gènes, la prédiction des promoteurs etc. (voir section 5.1.1). Toutefois, cet algorithme présente l'inconvénient de ne pas apporter d'information sur la "fiabilité" de la reconstruction. En effet, il garantit que le chemin le plus probable globalement est trouvé mais il ne permet pas de savoir à quel point le chemin prédit est plus probable que les autres chemins possibles. Ainsi, une mesure de pertinence du chemin reconstruit n'est pas donnée et les réponses aux questions ci-dessous ne sont pas apportées :

1. A quel point le chemin caché reconstruit à partir d'une chaîne de Markov cachée représente-t-il le "vrai" chemin, c'est à dire le chemin qui a été emprunté par le modèle pour générer la séquence observée ?
2. Le chemin caché reconstruit représente-t-il un chemin caché qui aurait pu être produit par le HMM utilisé ?

Dans un cadre théorique, les propriétés de la reconstruction des états cachés par algorithme de Viterbi ont été étudiées par Merhav et Ephraim (1991); Kogan (1988) et Caliebe et Roesler (2002). Un des résultats est que le chemin reconstruit jusqu'à une position t ne dépend pas de l'observation à la position n, lorsque n tend vers l'infini (avec $t \ll n$). Néanmoins, ces résultats théoriques restent encore insuffisants pour permettre des conclusions sur la pertinence de la prédiction du chemin caché dans le cadre d'une application en bioinformatique.

Lors de l'utilisation des HMM pour modéliser les ARNnc de *R. solanacearum*, nous avons utilisé l'algorithme de Viterbi pour reconstruire le chemin caché. Nous avons alors pu observer une incohérence apparente entre le fait que la distribution des longueurs des plages reconstruites d'un état d'une chaîne de Markov cachée suit une loi géométrique dépendant uniquement du paramètre de transition associé à cet état caché (voir section 5.1.2.1) et nos résultat qui suggéraient que cette distribution était dépendante des autres paramètres du modèle [1]. N'ayant pas trouvé de référence bibliographique

[1] Je tiens à remercier Sophie Schbath d'avoir soulevé ce problème lors du premier comité de thèse.

décrivant ce comportement de l'algorithme de Viterbi nous avons effectué une étude par simulation, afin d'évaluer la sensibilité de l'algorithme aux paramètres d'émission conditionnellement aux états cachés.

Si la segmentation par l'algorithme de Viterbi n'est pas "pertinente" dans le cas des séquences simulées, nous pouvons nous attendre à ce qu'elle le soit encore moins dans le cas des séquences génomiques réelles. C'est pourquoi, dans l'étude réalisée, nous avons utilisé les chaînes de Markov cachées simulées. De plus, l'uilisation des séquences simulées permet de contrôler librement les différents paramètres du modèle générant les séquences et de connaître la suite des états cachés (le "vrai" chemin). Ce dernier point permet de confronter la reconstruction au "vrai" chemin.

Le protocole utilisé et les résultats de cette étude sont présentés dans la suite du texte, dans le cas de l'algorithme de Viterbi et comme point de comparaison dans le cas de l'algorithme Forward-Backward.

5.2.2 Protocole de l'étude

Nous nous plaçons dans le cadre du modèle $M1M0$ à 2 états cachés s_1 et s_2, et nous étudions le comportment de l'algorithme de Viterbi en terme de la distribution des longueurs des plages reconstruites (DLPR dans la suite du texte), dans le cas des séquences simulées, générées par un HMM.

Nous avons simulé les séquences de longueur 100.000 à 900.000 (par pas de 200.000). Nous voulons évaluer la sensibilité de l'algorithme de Viterbi aux paramètres d'émission conditionnellement aux états cachés (voir section précédente). C'est pourquoi les différents HMM utilisés pour générer ces séquences diffèrent entre eux par les probabilités d'émission conditionnellement aux états cachés. Nous analyserons la reconstruction de l'état s_2.

Les probabilités de transition sont les mêmes pour tous les HMM utilisés, et ont été choisies comme des valeurs typiques que nous avons utilisées dans la modélisation des séquences génomiques dans le cadre de la détection des ARNnc. Les séquences générées sont caractérisées par des états "longs" et des transitions rares entre les états. Les valeurs des probabilités de transitions sont données dans le tableau 5.1. Dans le cas des transitions plus fréquentes, dans la reconstruction par algorithme de Viterbi, un des états prennait le dessus et le résultat de la reconstruction consistait alors en une seule

	s_1	s_2
s_1	0.985	0.015
s_2	0.99	0.01

TAB. 5.1 – Les probabilités de transition entre les états cachés (s_1 et s_2), utilisés pour simuler les séquences.

plage couvrant l'ensemble de la séquence.

La DLPR de l'état s_2 est géométrique de paramètre $1 - a_{s_1,s_1} = 0.01$ (équation 5.1) et la longueur moyenne des plages est située autour de 100nt (d'après l'équation 5.2).

Quant aux probabilités d'émission conditionnellement aux deux états cachés des modèles utilisés, nous avons utilisé la combinaison de quatre lois différentes données dans le tableau 5.2. La combinaison de ces lois d'émission permet de comparer la reconstruction dans le cas où les deux lois d'émission sont assez proches (par exemple b_1 et b_2) ou assez éloignées (par exemple b_1 et b_4).

	A	C	G	T
b_1	0.1	0.4	0.4	0.1
b_2	0.2	0.3	0.3	0.2
b_3	0.3	0.2	0.2	0.3
b_4	0.4	0.1	0.1	0.4

TAB. 5.2 – Les différentes lois d'émission conditionnelles aux états cachés utilisées pour générer les séquences.

Notons que d'autres valeurs, ou un plus grand nombre de valeurs balayant l'espace de lois d'émission, auraient pu être choisies. Cependant, les expérimentations à l'aide de quatre lois d'émission choisies ayant suffi à déceler la tendance globale dans les résultats, une étude plus exhaustive n'a pas été jugée nécessaire dans un premier temps.

Les chaînes de Markov cachées ont été simulées à l'aide du logiciel SHOW (Nicolas *et al.* (2004)).

5.2.3 Résultats

Nous avons d'abord évalué la sensibilité de la reconstruction par les algorithmes de Viterbi et Forward-Backward à la longueur de la séquence générée, afin d'écarter une

éventuelle influençe de ce paramètre sur la reconstruction. Ensuite nous avons évalué leur sensibilité aux différentes lois d'émission associées aux deux états cachés.

Nous présentons ces résultats dans la suite.

Influence de la longueur de la séquence Le graphique 5.6 présente les DLPR obtenues par reconstruction des séquences simulées. Seulement les résultats obtenus pour les lois d'émission $b_1 = (0.1, 0.4, 0.4, 0.1)$ et $b_2 = (0.4, 0.1, 0.1, 0.4)$ sont présentés, étant donné que des résultats similaires ont été obtenus pour les autres combinaisons des lois d'émission. Le tableau 5.3 présente complémentairement les longueurs moyennes des plages reconstruites.

Quatre constats peuvent être faits à partir de ces résultats. Premièrement, la longueur moyenne des plages reconstruites diffère systématiquement de la longueur attendue sous le modèle ; la longueur attendue se situe aux environs de 100nt, alors que les longueurs moyennes des plages reconstruites se situent aux environs de 125nt pour l'algorithme de Viterbi, et de 115nt pour l'algorithme "Forward-Backward". Deuxièmement, les DLPR ne sont que très peu sensibles à la longueur de la séquence. Ceci est en accord avec le résultat théorique obtenu par Caliebe et Roesler (2002) dans le cas de Viterbi (voir section 5.2.1). Troisièmement, les deux dernières remarques sont vraies pour les plages reconstruites en utilisant l'algorithme de Viterbi tout autant qu'en utilisant l'algorithme "Forward-Backward" et quatrièmement, le biais est moins important pour l'algorithme "Forward-Backward".

	longueurs des séquences simulées				
	100 000	300 000	500 000	700 000	900 000
Simulés	100.8	99.8	99.7	100.1	100.3
Viterbi	125.6	124.8	123.9	124.9	125.3
Forward-Backward	116.2	115.8	113.9	114.5	115.5

TAB. 5.3 – Longueurs moyennes des plages simulés et reconstuites de l'état s_2, obtenus avec les lois d'émission $b_1 = (0.1, 0.4, 0.4, 0.1)$ et $b_2 = (0.4, 0.1, 0.1, 0.4)$.

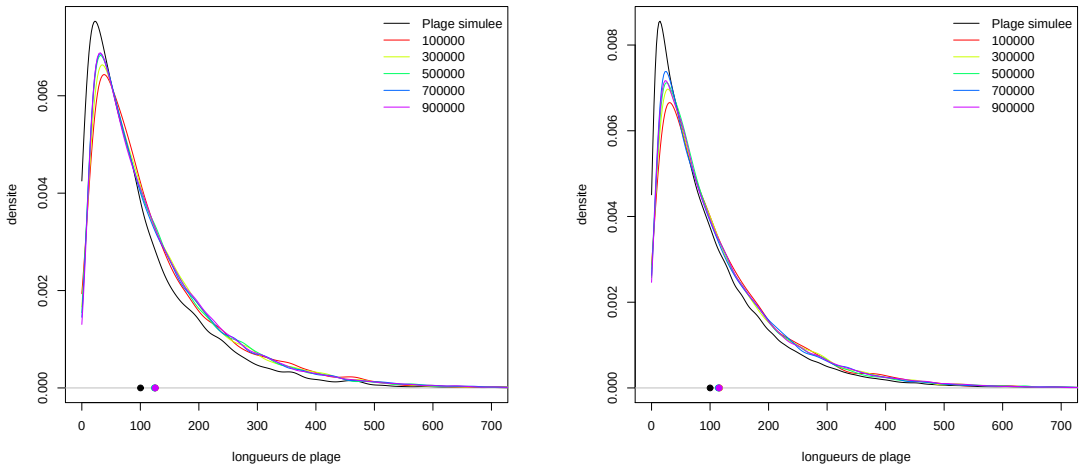
Influençe des lois d'émission Les résultats du paragraphe précédent indiquent que la reconstruction du chemin caché ne dépend pas de la longueur de la séquence

générée. Par conséquent, ici nous fixerons la taille de la séquence à 900 000, afin d'exclure la possibilité d'un comportement "non-asymptotique" dû à la longueur trop courte des séquences simulées et nous procédons par évaluation de l'influençe des valeurs des probabilités d'émission sur la reconstruction des états cachés par les algorithmes de Viterbi et "Forward-Backward".

Le graphique 5.7 présente les DLPR par l'algorithme de Viterbi et l'algorithme "Forward-Backward" pour des lois d'émission conditionnellement à l'état s_1 valant b_1 et les lois d'émission de l'état s_2 valant tour à tour b_2, b_3 et b_4. Le tableau 5.4 complète ces résultats en donnant les valeurs des longueurs de plages moyennes obtenues pour l'ensemble des simulations.

La figure 5.7a suggère que la DLPR dépend fortement des lois d'émission, pour l'algorithme de Viterbi. Ceci est également le cas, dans une moindre mesure, pour l'algorithme "Forward-Backward", ce qui est visible sur la figure 5.7b.

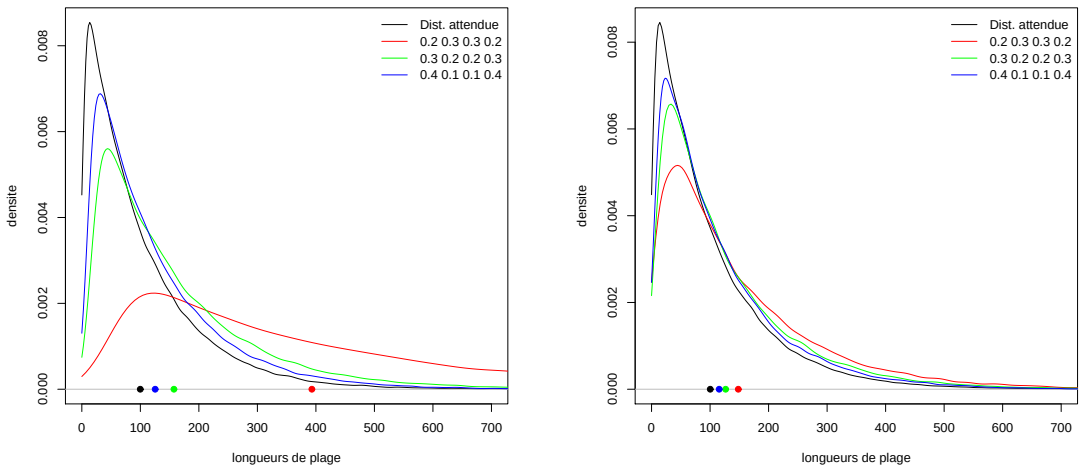
Dans les graphiques et dans le tableau des longueurs moyennes, une tendance peut être observée : plus les deux lois d'émission conditionnellement aux états cachés sont



(a) (b)

FIG. 5.6 – Les DLPR de l'état s_2, obtenues avec les lois d'émission s_1 : $b_1 = (0.1, 0.4, 0.4, 0.1)$ et s_2 : $b_2 = (0.4, 0.1, 0.1, 0.4)$. Les différentes couleurs correspondent au différentes longueurs des séquences simulés. Les points au-dessus de l'axe des abscisses représentent les longueurs moyennes. a) Reconstruction par l'algorithme de Viterbi b) Reconstruction par l'algorithme "Forward-Backward".

différentes entre elles, plus la reconstruction est proche de la "vraie" distribution. Pour les lois d'émission les plus proches, à savoir b_2 et b_3, ou b_3 et b_4, la longueur moyenne des plages reconstruites est 3.8 (pour b_2 et b_3) et 4.5 fois plus élevée (pour b_3 et b_4) que la valeur attendue sous le modèle. La même tendance peut être constatée pour toutes les combinaisons des lois d'émission (tableau 5.4) pour la reconstruction par Viterbi ainsi que pour la reconstruction par "Forward-Backward", dans une moindre mesure. Notons néanmoins que b_2 et b_3 sont aussi proches l'une de l'autre que b_3 et b_4 le sont, mais les DLPR et les valeurs moyennes associées ne sont pas égales (450.2 pour le premier couple et 357.4 pour le deuxième couple). Ceci laisse présager que la "distance" entre les lois d'émission ne suffit pas à elle seule pour expliquer la différence entre les DLPR et la vraie distribution.



(a) (b)

FIG. 5.7 – Distributions des longueurs des plages simulées et reconstuites de l'état s_2, obtenues avec les lois d'émission de l'état s_1 : $b_1 = (0.1, 0.4, 0.4, 0.1)$, les lois d'émission de l'état s_2 valant respectivement $b2 = (0.2, 0.3, 0.3, 0.1)$, $b3 = (0.3, 0.2, 0.2, 0.1)$ et $b4 = (0.4, 0.1, 0.1, 0.4)$. Les différentes couleurs correspondent aux différentes lois de l'état s_2. Les points au-dessus de l'axe des abscisses représentent les longueurs moyennes. a) Reconstruction par l'algorithme de Viterbi b) Reconstruction par l'algorithme "Forward-Backward".

Des observations précédentes suggèrent notamment que plus les lois d'émission conditionnellement aux deux états cachés sont éloignées plus la DLPR s'approche de la

s_1 \ s_2	Viterbi				"Forward-Backward"			
	b_1	b_2	b_3	b_4	b_1	b_2	b_3	b_4
b_1	–	393.1	157.4	125.3	–	148.2	126.5	115.5
b_2	359.2	–	450.2	156.2	145.9	–	149.2	125.2
b_3	156.6	455.5	–	357.4	126.3	150.6	–	146.9
b_4	125.2	158.7	381.1	–	115.1	125.7	147.2	–

TAB. 5.4 – Longueurs moyennes des longueurs des plages simulées et reconstuites de l'état s_2.

distribution géometrique des longueurs des plages de la vraie distribution. La section suivante explore cette relation à l'aide d'un grand nombre de HMM aux lois d'émission différentes et d'une distance entre ces lois.

5.2.3.1 L'influençe de la distance entre les lois d'émission sur la reconstruction du chemin caché par algorithme de Viterbi

Dans cette section nous examinons le lien entre la différence entre les lois d'émission conditionnellement aux deux états du HMM et l'écart entre les DLPR à la vraie distribution associée. Pour ce faire nous devons définir une distance entre les lois d'émission ainsi qu'une distance entre les distributions des longueurs de plages comparées.

Nous définissons d'abord les distances considérées pour ensuite présenter l'approche utilisée afin d'examiner leur lien.

5.2.3.2 Distance en variation totale pour évaluer la différence entre les lois d'émission

La distance en variation totale est une distance entre deux lois de probabilité et dans le cadre des lois d'émission du modèle M1M0, aux états s_1 et s_2. Elle est définie par :

$$d_{VT} = \frac{1}{2} \sum_{x \in \mathcal{A}} |s_1(x) - s_2(x)| \tag{5.11}$$

Par exemple, les distances en variation totale entre les lois d'émission utilisées dans les simulations précédentes varient de 0.2 à 0.6 ; elles sont données dans le tableau 5.5.

	b_1	b_2	b_3	b_4
b_1	–			
b_2	0.2	–		
b_3	0.4	0.2	–	
b_4	0.6	0.4	0.2	–

TAB. 5.5 – Distances en variation totale entre les différentes lois d'émission des HMM utilisés dans la section 5.2.2.

5.2.3.3 Ecart entre la vraie distribution et la distribution des plages reconstruites

Pour mesurer cet écart nous avons eu recours à une mesure définie à l'aide d'un estimateur de paramètre du "temps passé" dans l'état considéré. En supposant que la séquence reconstruite est une chaîne de Markov cachée et que la DLPR est alors géometrique, nous estimons d'abord le paramètre de cette loi qui donne la valeur du paramètre de transition correspondant à l'état considéré (équation 5.1). Ensuite, les distributions des plages, vraies et reconstruites, cette dernière supposée géometrique, étant entièrement définies par le paramètre de la loi géometrique associée, la différence entre ces paramètres peut être prise comme une mesure *ad hoc* de l'écart entre ces distributions.

Nous introduisons, à présent, l'estimateur du paramètre de transition, à partir de la DLPR.

La longueur des plages de l'état S générées par un HMM suit une distribution géométrique de paramètre $1-a_{S,S}$ (équation 5.1). Nous proposons de construire un estimateur de la probabilité de rester dans l'état S, noté $\widehat{a_{S,S}}$, à partir de cette observation.

Soit $X_1, \ldots, X_n$ une suite de variables aléatoires suivant une loi géométrique de paramètre p. Nous avons, alors

$$\mathbb{P}(X_i = x_i) = p \cdot (1-p)^{x_i - 1}, x \in \mathbb{N} \tag{5.12}$$

La vraisemblance d'observer $X_1 = x_1, \ldots, X_n = x_n$, sous le modèle s'écrit alors comme une fonction du paramètre p :

$$\begin{aligned} L(p) &= (1-p)^{x_1-1} \cdot p \ldots (1-p)^{x_n-1} \cdot p \\ &= (1-p)^{\sum_{i=1}^{n} x_i - n} \cdot p^n \end{aligned} \tag{5.13}$$

Chercher l'estimateur de maximum de vraisemblance revient à chercher l'estimateur de log-vraisemblance et nous pouvons passer à l'echelle logarithmique :

$$\log L(p) = n \log p + (\sum_{i=1}^{n} x_i - n) \cdot \log(1-p) \tag{5.14}$$

Le maximum de la fonction de vraisemblance étant au zéro de la dérivée, nous avons que pour $p = \hat{p}$:

$$(\log L(p))' = n \cdot \frac{1}{p} - (\sum_{i=1}^{n} x_i - n) \cdot \frac{1}{1-p} = 0 \tag{5.15}$$

et donc :

$$\hat{p} = \frac{n}{\sum_{i=1}^{n} x_i} \tag{5.16}$$

Nous obtenons donc que l'estimateur de vraisemblance du paramètre de la loi géométrique est donné par $\frac{1}{\overline{X}_{S,S}}$, où $\overline{X}_{S,S}$ est la moyenne des longueurs de toutes les plages de l'état S. L'estimateur que nous retiendrons pour le paramètre de transition est alors (d'après l'équation 5.1) :

$$\widehat{a}_{S,S} = 1 - \frac{1}{\overline{X}_{S,S}} \tag{5.17}$$

Cet estimateur sera utilisé pour mesurer l'écart entre la vraie distribution des plages, dont le paramètre est connu, et la DLPR.

Dans la suite on se propose d'étudier cet écart en fonction de la distance en variation totale entre les lois d'émission conditionnellement aux états cachés.

5.2.3.4 L'écart entre les paramètres de transition en fonction de la distance entre les lois d'émission dans la reconstruction par algorithme de Viterbi

Nous voulons maintenant étudier la différence entre le paramètre de transition a_{s_2,s_2} et le paramètre estimé par l'équation 5.17 à partir de la DLPR, en fonction des lois d'émission conditionnellement aux états cachés. Pour ce faire, nous proposons l'expérience suivante :

1. Nous simulons les séquences à l'aide des HMM, modèles M1M0 à 2 états. Les paramètres des lois d'émission sont tirés aléatoirement dans une loi uniforme et renormalisés pour que leur somme soit égale à 1. On utilise les mêmes paramètres de transition que ceux utilisés précédemment : $a_{s_1s_1} = 0.985, a_{s_1s_2} = 0.015, a_{s_2s_2} =$

$0.99, a_{s_2s_1} = 0.01$. Les séquences de trois longueurs différentes seront simulées : $n = 1000, 10000$ et 100000.

2. Pour chacune des séquences nous calculons la distance en variation totale entre deux lois d'émission (définie comme 5.11).
3. Pour chacune des séquences nous calculons la différence entre la valeur du paramètre de transition $a_{s_2s_2}$ et $\hat{a}_{s_2s_2}$, la valeur du paramètre estimée à partir du DLPR par algorithme de Vietrbi. (équation 5.17).
4. Nous traçons le graphique de différence des paramètres en fonction de la distance entre les lois d'émission pour toutes les séquences simulées.

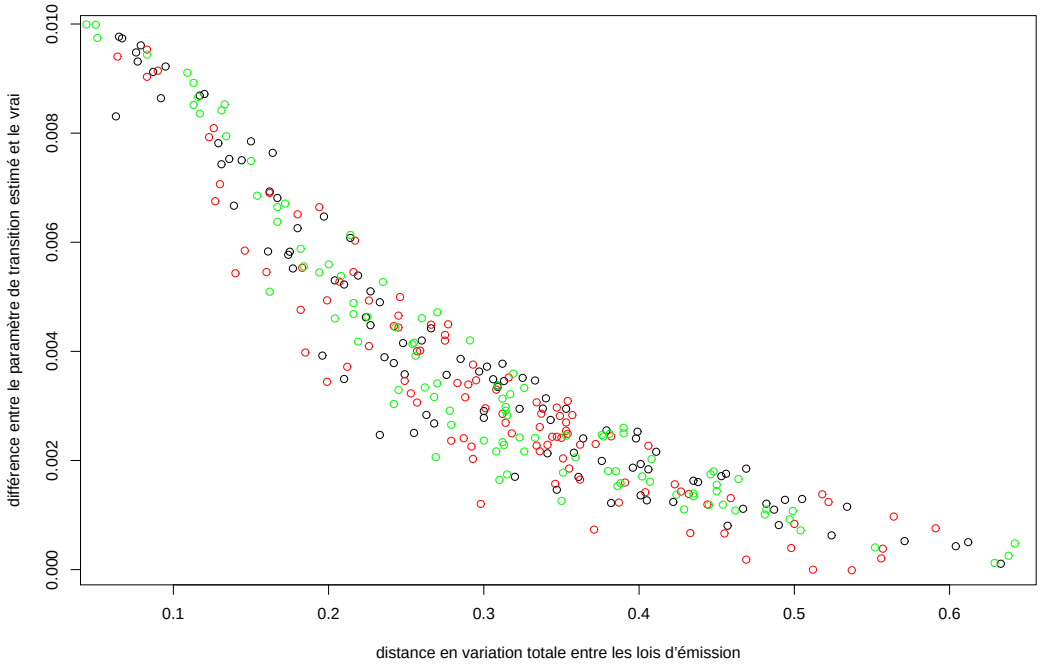


FIG. 5.8 – Adéquation entre le paramètre de transition $a_{s_2s_2}$ et celui estimé après reconstruction par l'algorithme de Viterbi $\hat{a}_{s_2s_2}$, en fonction de la distance entre les lois d'émissions. Les lois d'émission sont tirées dans une distribution uniforme renormalisée. Les points noirs, rouges et verts représentent les valeurs pour les séquences simulées de longueur 1000, 10.000 et 100.000, respectivement.

Les résultats sont présentés sur la figure 5.8. Ce graphique montre que plus la distance entre les lois d'émission conditionnellement aux états cachés est grande, plus

l'écart entre les paramètres $a_{s_2s_2}$ et $\hat{a}_{s_2s_2}$ augmente, confirmant ainsi la tendance observée dans 5.2.3. Cette tendance semble être particulièrement accentuée entre les valeurs 0 et environ 0.3 en distance en variation totale, indiquant que la DLPR est particulièrement éloignée de la vraie distribution lorsque la distance entre les lois d'émission se situe dans l'intervalle $[0, 0.3]$. La différence entre le paramètre de transition estimé et le vrai paramètre est, selon ces résultats, toujours positive ou nulle.

5.2.4 Discussion

Cette étude de simulations montre que, dans le cadre de l'utilisation d'un modèle HMM M1M0 à deux états, la DLPR dépend de la différence des lois d'émission de chacun des états, à la différence des plages d'une chaîne de Markov caché. La chaîne reconstruite peut alors ne pas correspondre au modèle qui a été utilisé pour la reconstruction. Nos résultats suggèrent également que la reconstruction ne dépend pas de la longueur de la chaîne simulée, conformément aux résultats de Caliebe et Roesler (2002).

Cet écart a été observé dans le cas de l'utilisation de l'algorithme de Viterbi mais également dans le cas de l'utilisation de l'algorithme "Forward-Backward", dans une moindre mesure cependant. En particulier, nous avons observé que plus les lois d'émission étaient proches plus l'écart entre la vrai valeur et la valeur estimée de paramètre de transition est faible. Cette tendance a été confirmée par l'étude par simulation qui a exploré de façon plus systématique l'influençe des différentes combinaison de deux lois d'émission. Cette deuxième étude n'a pas été réalisée dans le cas de reconstruction par algorithme Forward-Backward, mais elle serait nécessaire afin de pouvoir comparer rigoureusement la reconstruction par ces deux algorithmes en terme des DLPR.

Pour mesurer l'écart entre les valeurs des paramètres de transition, vrai et estimé, nous avons défini une distance *ad hoc*. L'utilisation d'une distance entre les distributions empiriques devrait être utilisée dans une éventuelle étude plus approfondie.

Ici, nous avons exploré l'influençe de la différence entre les paramètres d'émission sur la reconstruction du chemin caché, mais nous avons vu que seul, il ne suffit pas à expliquer l'écart entre DLPR et vrai distribution (section 5.2.3). Le deuxième paramètre de transition, pourrait également avoir une influençe, et une autre étude de simulation devrait être effectuée afin de répondre à cette question.

Une étude portant sur des modèles d'ordre différent ou à nombre supérieur d'états

n'a pas été menée. Néanmoins, on peut raisonnablement s'attendre à ce que la différence des lois d'émission ait un impact sur les lois de transition empiriques du chemin reconstruit.

Cette étude par simulation soulève également la question de la possibilité de déterminer, de façon théorique, la DLPR par algorithme de Viterbi. Cette question nous semble d'autant plus pertinente que les DLPR empiriques que nous avons pu observer (figure 5.7) des distributions "régulières" qui se ressemblent entre elles. De plus (résultats non montrés), d'après les résultats de simulations, les DLPR des chaînes différentes générées à l'aide du même HMM présentent la même distribution, indiquant qu'il pourrait s'agir d'une loi dépendante des paramètres du HMM.

Dans le cadre de l'utilisation des HMM pour la prédiction des ARNnc chez *R. solanacearum* ces résultats suggèrent qu'une utilisation de l'algorithme de Viterbi n'est pas appropriée, étant donnée la faible différence entre les lois d'émission estimées pour l'état ARN et le reste du génome (tableau 4.12). L'algorithme "Forward-Backward" de reconstruction est mieux adapté à cette fin, car la distribution des longueurs des plages est moins dépendante de la différence des lois d'émission et la séquence reconstruite est plus proche d'une séquence issue du modèle utilisé pour la segmentation.

Plus généralement, il nous semble que l'utilisation de l'algorithme de Viterbi exige une vigilence quant aux valeurs des paramètres du modèle utilisé pour la reconstruction. Cet inconvénient de l'algorithme de Viterbi bien connu de la communauté statistique semble être moins connu dans la communauté bio-informatique.

5.3 Application des HMM pour la recherche des ARNnc dans le génome de *Ralstonia solanacearum*

Nous présentons ici les résultats de l'utilisation des HMM, et plus particulierement l'algorithme Forward-Backward, pour la recherche des ARNnc dans le chromosome de *Ralstonia solanacearum*, en se basant sur l'hypothèse de l'existence d'une différence de composition entre les ARNnc et le reste du génome de *R. solanacearum*. Cette hypothèse est appuyée par l'étude présentée dans la section 4.2. Cette étude suggère également que l'utilisation du G+C% n'est pas suffisante pour discriminer l'ensemble des ARNnc du groupe des séquences intergéniques (groupe AUTRE) et par conséquent, nous utilisons

une modélisation d'abord en nucléotides et puis en di-nucléotides.

Tous les calculs d'estimations et de segmentations ont été faits à l'aide du logiciel SHOW (Structured HOmogeneities Watcher, Nicolas *et al.* (2004)).

5.3.1 Modèle utilisé

5.3.1.1 Topologie

Conformément aux trois ensemble considérés AUTRE, CODANT et ARN utilisés pour mettre en évidence l'étude de la composition des ARN, nous proposons une modélisation par trois états cachés correspondant à ces groupes.

Dans le cadre de cette étude nous recherchons les ARNnc se trouvant dans les régions intergéniques. Afin de diminuer le nombre de paramètres à estimer nous pouvons contraindre le modèle à ce que les transitions entre les états COD et ARN soient impossibles : entre une région prédite comme codante et une région prédite comme ARN il doit toujours se trouver une région AUTRE les séparant. En réalité, ceci n'est pas toujours vérifié, les riboswitch, se trouvant dans les régions 5'UTR des gènes, faisant l'exception. Si toutefois nous nous focalisons sur les ARNnc transcrits, se trouvant au milieu des IGR, cette hypothèse est valable, d'autant plus que les riboswitch de *R. solanacearum* ne semblent pas avoir une composition se démarquant de la composition du reste du génome (voir section 4.1.2.3).

Le graphe des états cachés (section 5.1) du modèle que nous proposons est donné sur la figure 5.9.

Notons que la topologie du modèle est simplificatrice quand à la modélisation des séquences codantes : elle est loin de prendre en compte toutes les subtilités de composition de ces séquences (biais du troisième codon, séquences START et STOP etc.).

5.3.1.2 Les paramètres

Les lois d'émission dans chacun des états cachés ont été estimés par maximum de vraisemblance à l'aide des comptages des nucléotides (tab. 5.6) et des di-nucléotides (tab. 5.7, 5.8 et 5.9).

Les paramètres de transition non nuls ont été estimés par l'algorithme EM pour le modèle M1M0 (section 5.1.3) et sont présentés dans le tableau 5.10. Les mêmes paramètres de transition ont été repris pour le modèle M1M1, afin de rendre les résultats

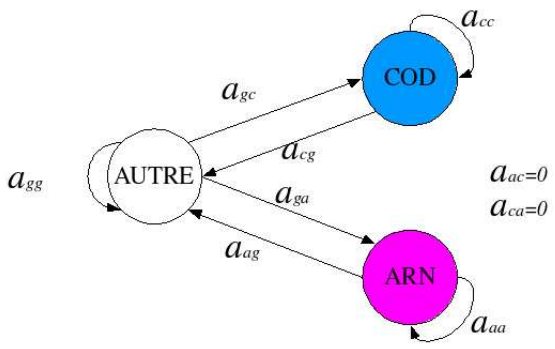


FIG. 5.9 – Graphe des états cachés utilisé pour la segmentation du génome de *R. solanacearum*. Les états cachés correspondent aux séquences codantes (COD), les ARNnc (ARN) et le reste du génome (AUTRE). Les probabilités de transition entre états COD et ARN sont nulles.

état	A	C	G	T
COD	0.1605285	0.3365921	0.3403506	0.1625287
AUTRE	0.1891398	0.3060180	0.3080573	0.1967848
ARN	0.2135506	0.2841981	0.2750858	0.2271655

TAB. 5.6 – La matrice d'émission estimée, du modèle M1M0 pour chacun des états.

de segmentation à l'aide des deux modèles , M1M0 et M1M1, comparables. Notons que la longueur moyenne des plages ARN attendue selon ce modèle se situe autour de 100nt (d'après les calculs présentés dans la section 5.5).

5.3.2 Résultats

Le tableau 5.11 récapitule les résultats de segmentation utilisant les modèles M1M0 et M1M1. 553 et 729 régions correspondent respectivement à l'état ARN. Dans un grand nombre de cas, les segments identifiés par la segmentation M1M0 "englobent" plusieurs segments plus courts identifiés par M1M1, ce qui explique le nombre plus élevé des segments M1M1 et la longueur moyenne plus élevée des segments M1M0. Cependant, ceci n'est pas toujours le cas, et certaines des prédictions d'un des deux

COD	A	C	G	T
A	0.16068760	0.3038993	0.3049934	0.2304197
C	0.19103037	0.2548348	0.4108662	0.1432687
G	0.17196330	0.4206495	0.2589375	0.1484497
T	0.06836089	0.3650046	0.4013393	0.1652952

TAB. 5.7 – La matrice d'émission estimée, du modèle M1M1 pour l'état COD.

AUTRE	A	C	G	T
A	0.2604111	0.2564265	0.2479955	0.2351669
C	0.2006508	0.2810123	0.3648156	0.1535213
G	0.1910854	0.3640116	0.2822662	0.1626368
T	0.1058371	0.2987533	0.3298148	0.2655948

TAB. 5.8 – La matrice d'émission estimée, du modèle M1M1 pour l'état AUTRE.

modèles ne se supersposent pas avec les prédictions du deuxième modèle. Notons que la longueur moyenne des plages reconstruites est considérablement plus élevée que la longueur attendue, se situant autour de 100nt (section 5.3.1.2). Ainsi, les plages reconstruites sont, en moyenne, plus longues que les ARNnc connus dans le chromosome de *R. solanacearum* (voir tableau 4.7), et la majorité des segments ainsi identifiés ont une longueur supérieure à 200nt. Nous constatons cependant que, les segments issus du modèle M1M1 sont plus proches en longueur des ARNnc connus que les segments prédits par le modèle M1M0.

De façon générale, la couverture du génome est plus faible dans le cas M1M1 mais en même temps plus spécifique que celle dans le cas M1M0, étant donné que plus d'ARNnc connus ont été retrouvés par le modèle M1M1 (56 contre 52).

Tous les segments reconstruits appartenant à l'état ARN ne se trouvent pas dans les régions intergéniques. Effectivement, une partie importante des segments est localisée dans les régions codantes qui couvrent 90% du génome (5.12). Les segments reconstruits couvrent 8% (M1M0) et 4% (M1M1) des régions codantes.

Dans le cas des deux modèles, la majorité des ARNnc retrouvés représentent les ARNnc "structuraux", les ARNr et les ARNt. Les ARNr, au nombre de six, ont été retrouvés par les deux modèles. Pour le modèle M1M0, 43 et pour le modèle M1M1 46, des 54 ARNt ont été retrouvés.

ARN	A	C	G	T
A	0.2038345	0.2941473	0.2805247	0.2214934
C	0.2113761	0.2943119	0.2759633	0.2183486
G	0.2139364	0.2660962	0.2901385	0.2298289
T	0.2292660	0.2855100	0.2478551	0.2373689

TAB. 5.9 – La matrice d'émission estimée, du modèle M1M1 pour l'état AR.

	COD	AUTRE	ARN
COD	0.999788	0.000212416	0
AUTRE	0.00093448	0.998454	0.000611197
ARN	0	0.00205836	0.997942

TAB. 5.10 – Les probabilités de transition entre différents états. Les paramètres non nuls ont été estimés par l'algorithme EM.

5.3.2.1 Evaluation de la spécificité et de la sensibilité

La spécificité et la sensibilité ont été évaluées dans le cas de la classification des segments et dans le cas de la classification des nucléotides. Dans le premier des cas, nous regardons combien de segments reconstruits correspondent aux ARNnc connus. Nous considérons qu'un segment est bien prédit lorsqu'il chevauche la séquence d'un ARNnc connu. Dans le deuxième cas, nous considérons la qualité de la prédiction en terme de nucléotides : un nucléotide à une position donnée est bien prédit si le nucléotide à cette position appartient à un ARNnc connu dans le génome de *R. solanacearum*. Sur la tableau 5.13 les résultats sont évalués en terme de spécificité et de sensibilité, définis comme :

$$\text{spécificité} : \frac{VP}{VP+FP}, \text{ sensibilité} : \frac{VP}{VP+FN}$$

où VP désigne *les vrais positifs* (le nombre de segments ou nucléotides bien prédits), FP désigne *les faux positifs* (les segments ou nucléotides prédits qui ne correspondent pas aux ARNnc connus) et FN désigne *les faux négatifs* (le nombre d' ARNnc ou nucléotides qui ne sont pas prédits).

Dans les deux cas, les prédictions M1M0 et M1M1 sont assez sensibles mais très peu spécifiques. Pour les deux modèles la sensibilité est meilleure dans en terme de nucléotides. Ceci est probablement dû au fait que les ARN ribosomiques, ayant une

Modèle	nombre de segments	nombre de segments > 200nt	longueur moyenne segments	longueur médiane segments	couverture génome	nb ARN retrouvés	nombre nt ARN / nombre nt prédit ARN
M1M0	553	413	787	456	11%	52	7.32
M1M1	729	337	351	178	6%	56	5.08

TAB. 5.11 – Le tableau récapituatif des résultats de segmentation du génome de *R. solanacearum* à l'aides des modèle M1M0 et M1M1.

Modèle	% de nt se trouvant dans les rég codantes	% de régions codantes couvertes	% des IGR couvert
M1M0	64%	8%	34%
M1M1	53%	4%	25%

TAB. 5.12 – Le tableau récapitulant la localsation géomiques des segments ARN.

composition particulière par rapport au reste du génome (voir section 4.1.3), ont tous été identifiés par les deux modèles HMM. Leur longueur compte davantage dans le calcul de la sensibilité en terme de nucléotide qu'en terme de segments. En revanche, la spécificité en terme de nucléotides est très faible, et ceci peut probablement s'expliquer par la présence des segments prédits longs qui ne correspondent pas aux ARNnc connus.

5.3.2.2 Analyse des ARNnc identifiés

Les ARNnc retrouvés par les deux modèles correspondent d'une part aux ARN ribosomiques, 6 en tout, qui sont retrouvés systématiquement, probablement du fait de leur composition particulière. La majorité des ARN de transfert est également retrouvée par les deux modèles (44 sur 54 pour le modèle M1M0 et 47 sur 54 pour le modèle M1M1). Les ARNnc qui n'ont été identifiés par aucun des deux modèles correspondent aux autres familles d'ARNnc et à cinq ARNt (tableau 5.15).

La plupart des prédictions à l'aide des deux modèles est commune (51 prédictions). Le modèle M1M1 prédit correctement plus d'ARNnc que le modèle M1M0, mais certains ARNnc sont identifiés par le modèle M1M0 seulement (tableau 5.14). Notons que l'ARNt de la proline est identifié une fois par le modèle M1M0 et une fois par le modèle M1M1.

	M1M0	M1M1
segments		
VP	52	56
FN	19	15
FP	501	673
Spécificité	0.09	0.08
Sensibilité	0.73	0.79
nucléotides		
VP	16230	16526
FN	2425	2042
FP	419321	239576
Spécificité	0.0001	0.06
Sensibilité	0.87	0.89

TAB. 5.13 – La spécificité et la sensibilité des sementations M1M0 et M1M1, évaluées à l'aide des vrais positifs (VP), faux positifs (FP) et faux négatifs (FN). L'évaluation a été faite en terme de segments retrouvés et en terme de nucléotides retrouvés.

ARN	modèle	pos. debut	pos. fin
Ser ARNt	M1M0	2344958	2345044
Pro ARNt	M1M0	1645665	1645738
Thr ARNt	M1M1	2306303	23063
Gly ARNt	M1M1	751244	751314
Pro ARNt	M1M1	1700427	1700500
Ala ARNt	M1M1	3508442	3508514
Val ARNt	M1M1	1833444	1833516

TAB. 5.14 – Les ARNnc qui ont été identfiés par un des deux modèles.

5.3.3 Discussion

Les segmentations du chromosome de *R. solanacearum* à l'aide des modèles M1M0 et M1M1, à 3 états cachés, ont généré respectivement553 (M1M0) et 729 (M1M1) segments.

La longueur moyenne de ces segments est plus élevée que la longueur attendue sous

ARN	pos. début	pos fin
TPP	132357	132458
yybP-ykoY	997648	997783
yybP-ykoY	2274739	2274896
SRP	1259539	1259640
Cobalamin rioswitch	2608997	2609253
Glycine ribositch	3545686	3545809
Glycine riboswitch	3545580	3545672
RNase P	525998	526276
Arg ARNt	1018652	1018723
Leu ARNt	1170093	1170180
Ser ARNt	1230833	1230923
Met ARNt	2399622	2399694
Leu ARNt	2634813	2634894

TAB. 5.15 – Les ARNnc de *R. solanacearum* qui n'ont pas été identifiés par aucun des deux modèles

le modèle M1M0 et M1M1. Malgré l'utilisation de l'algorithme "Forward-Backward" pour la reconstruction du chemin caché, il est probable que, au regard des conclusions de la section 5.5, ces résultats peuvent être attribués au faible écart entre les lois d'émission associées aux différents états. Le même problème est, du moins en partie, responsable de la faible spécificité des résultats obtenus. Pour ces raisons une piste possible d'amélioration de la spécificité de la prédiction à l'aide des HMM serait de déterminer les paramètres de transition en utilisant un autre procédé que l'algorithme EM, en recherchant, par exemple, empiriquement, avec un processus de balayage, les paramètres augmentant la spécificité sans diminuer la sensibilité.

Un nombre important des segments reconstruits se trouve dans les régions codantes. Ce résultats était attendu, en tout cas en terme de G+C% et donc possiblement en terme de nucléotides et di-nucléotides, car les distrubutions de G+C% des ARNnc et des séquences codantes ne sont pas entièrement séparées (voir la figure 4.8 de la section 4.1.2.3). Ce résultat est également conforme au fait que les régions codantes ne peuvent pas être entièrement décrites par leur composition en nucléotides mais que des modèles plus élaborés sont nécessaires pour leur modélisation. De ce fait, une deuxième façon

d'améliorer la spécificité des prédictions serait de modéliser les séquences codantes par trois états cachés correspondant aux trois positions du codon, comme ceci est fait habituellement. En revanche, le chevauchement d'une partie des segments prédits avec les régions codantes offre la possibilité d'éliminer ces prédictions "faux positifs"de l'ensemble des candidats ARNnc. Ainsi, dans le cas où nous voudrions analyser les résultats obtenus pour la recherche des candidats ARNnc plus de la moitié de ceux-ci serait rejetée en raison de leur appartenance aux régions codantes.

Quant à la comparaison des résultats de segmentation à l'aide des deux modèles, il aparaît que le modèle M1M1 est plus performant en sensibilité et en spécificité. Ceci est conforme aux observations affirmant qu'un modèle prenant en compte les di-nucléotides successifs serait mieux adapté pour caractériser les ARNnc, probablement en raison de l'importance des empilements de bases au sein de séquences d'ARNnc structurées (Clote *et al.* (2005)).

L'approche *ab initio* présentée permet d'identifier 52 (modèle M1M0) et 56 (modèle M1M1) des 69 ARNnc connus dans le chromosome de *R. solanacearum* et 58 sur 69 ARNnc, si l'on considère l'union des prédictions à l'aide des deux modèles. Toutefois, la quasi-totalité de ces ARNnc correctement prédits se limite aux ARN structuraux, les ARNt et les ARNr, dont la prédiction peut s'effectuer de façon efficace par d'autres outils adaptés à la recherche de ces familles spécifiques. Les ARNnc qui n'ont pas pu être identifiés correspondent, pour leur majorité, aux riboswitch. Quant à la détection des ARNnc proprement dits, l'approche HMM a permis d'en retrouver deux (l'ARN 6S et l'ARN *suhB*) tandis que les ARN SRP et RNase P n'ont pas été retrouvés. Ces résultats suggèrent tout d'abord que les ARN structuraux (c'est-à-dire les ARNt et les ARNr) possèdent une composition spécifique, différente du reste de génome. D'autre part, au regard de ces résultats, les riboswitch ne semblent pas être caractérisés par un biais en composition. Pour les 4 autres ARNc, deux ont été correctement prédits et deux ont été omis des prédictions. Nous sommes donc dans l'impossibilité de conclure si l'utilisation du biais de composition est pertinente pour la recherche des ARNnc dans *R. solanacearum*. Le renforcement des arguments pour penser qu'un biais existe pourrait venir des prédictions de nouveaux ARNnc dans ce génome.

Notons encore qu'une modélisation plus élaborée, en terme d'états cachés, pourrait donner lieu à des prédictions plus pertinentes et plus spécifiques. En effet, le génome de *R. solanacearum* étant caractérisé par la présence de 93 régions acquises probablement

par transfert horizontal et dont la composition en G+C% varie de 50% à 70% (les régions ACUR, Salanoubat *et al.* (2002)) une modélisation plus élaborée pourrait consister à modéliser ces régions par plusieurs états cachés ou à les masquer afin de les exclure de l'analyse.

Enfin, les résultats obtenus par segmentation sur la base de biais de composition n'ayant pas été concluants, ils n'ont pas été analysés de façon à déterminer les candidats ARNnc pour la validation biologique. Dans la suite dans le suite du travail de thèse nous avons utilisé une approche comparative afin de proposer des candidats ARNnc dans le génome de *R. solanacearum*.

Troisième partie

RNAsim : une approche comparative pour la détection des ARNnc

Introduction

L'analyse des résultats de l'approche par bias de composition n'a pas été menée au bout et nous avons envisagé une autre approche afin de réaliser notre objectif scientifique, l'identification des ARNnc candidats dans le génome de *R. solanacearum*. Au cours de la thèse, le contexte biologique a changé : deux nouvelles souches de génome de *R. solanacearum* ont été séquencées et nous avons eu à notre disposition leurs séquences. Cette nouvelle donne permettait l'exploitation de leur conservation, pour identitifier des nouveaux ARNnc. Par conséquent, le choix de méthode s'est naturellement porté sur une approche comparative de détection des ARNnc.

Dans cette partie, je présenterai d'abord un état de l'art sur les méthodes comparatives existantes. Ensuite, je présenterai un outil de détection des ARNnc par approche comparative, nommé RNAsim, auquel j'ai contribué durant ma thèse. Enfin, je présenterai une synthèse des résultats de l'utilsation de RNAsim pour la détection des ARNnc dans le génome de *R. solanacearum*.

Chapitre 6

Approche comparative pour la détection des ARNnc

6.1 Etat de l'art

Les approches comparatives pour la détection des ARNnc peuvent être réparties en trois catégories, selon l'objectif recherché et les méthodes mises en place pour l'accomplir (schématisé sur la figure 6.1).

Un premier objectif consiste à rechercher les ARNnc membres d'une famille déjà connue et caractérisée par sa structure secondaire. Les approches mettent généralement en oeuvre une modélisation par SCFG (Stochastic Context-Free-Grammars) de la structure secondaire consensus de la famille d'ARN recherchée (Durbin *et al.* (1998b)). Parmi les outils dévéloppés avec cet objectif nous pouvons citer RSEARCH (Klein et Eddy (2003)) ou FastR (Zhang *et al.* (2005)). Notons que la base de données de référence recensant les membres des différentes familles d'ARNnc connues, Rfam (Griffiths-Jones *et al.* (2003)), est basée sur INFERNAL ("INFERence of RNA ALignment"), le modèle sous-jacent à RSEARCH. La pluspart des ARNnc de *R. solanacearum* connus à ce jour ont donc été identifiés de cette façon là.

Ces méthodes sont utilisées pour l'annotation des familles d'ARNnc connues mais elles sont impuissantes quand il s'agit de recherche des ARNnc appartenant à des familles non connues.

Deux autres catégories d'approches n'exigent pas de connaissances *a priori* sur une famille d'ARNnc partuculière et sont, de ce fait, plus adaptées à notre question d'intérêt.

La deuxième catégorie s'appuie sur une conservation en séquence qui est ensuite examinée pour détecter une éventuelle conservation en structure. Dans les outils suivant cette démarche nous pouvons citer QRNA (Rivas et Eddy (2001)), RNAz et son ancienne version AlifoldZ, un outil qui a été originellemnt conçu pour la prédiction de la structure secondaire et ensuite étendu à la prédiction des ARNnc (Washietl *et al.* (2005)), ddbRNA (di Bernardo *et al.* (2003)) et MSARi (Coventry *et al.* (2004)).

QRNA exploite l'hypothèse que les alignements des séquences d'ARNnc présentent des mismatch mais conservent la structure secondaire (mutations compensatoires). QRNA a été utilisé avec succès pour la prédiction des ARNnc dans plusieurs génomes bactériens (Rivas *et al.* (2001), Silvaggi *et al.* (2006a), del Val *et al.* (2007)) et eucaryotes (McCutcheon et Eddy (2003)). RNAz propose une autre approche combinant la prise en compte de la stabilité thermodynamique estimée d'un repliement commun obtenu à partir de l'alignement multiple des séquences (décrit dans Washietl et Hofacker (2004)) et l'existence des mutations compensatoires au sein de ces repliements. Cet outil a été testé sur les ARNnc eucaryotes de *S. cerevisiae* et *C. elegans* et a été utilisé avec suc-

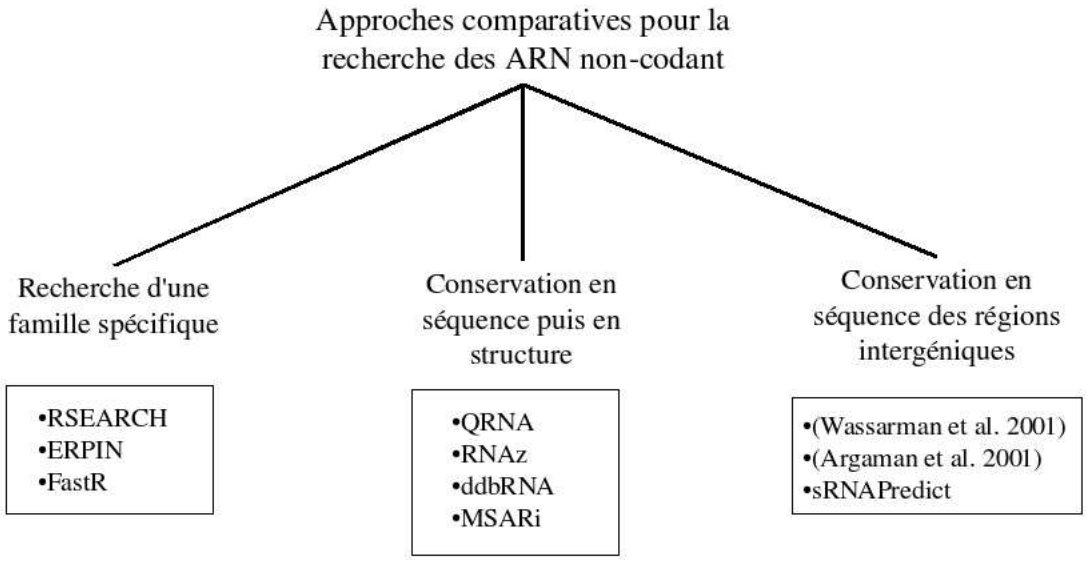


FIG. 6.1 – Les approches comparatives pour la recherche des ARNnc peuvent être réparties en trois groupes selon l'objectif recherché et les méthodes mises en place pour l'accomplir.

cès, en combinaison avec QRNA, pour la détection des ARNnc dans *S. meliloti* (del Val *et al.* (2007)). ddbRNA et MSARi, à l'aide d'une méthodologie différente, mettent en oeuvre l'idée présente dans RNAz.

Ces études, implémentant des modèles complexes, ont l'inconvénient d'un temps de calcul élevé mais présentent surtout un nombre important de paramètres à régler. Ces paramètres, lorsqu'ils sont utilisés avec leurs valeurs par défaut sont estimés sur des ensembles de test et peuvent être biaisés de ce fait. Un autre inconvénient apparaît lorsqu'on veut comparer plus de deux génomes. QRNA est conçu pour analyser les alignements deux-à-deux, tandis que RNAz, ddbRNA et MSARi nécessitent, en entrée, des alignements multiples. Cet écueil peut, dans le cas des trois derniers outils, être contourné à l'aide des outils d'alignements multiples de génomes entiers (par exemple MultiZ, Blanchette *et al.* (2004)).

La troisième catégorie d'approches comparatives pour la détection des ARNnc, s'appuie sur des stratégies moins complexes, d'un point de vue méthodologique. Il s'agit des travaux ayant permis de détecter des ARNnc sur la seule base de la conservation en séquence, en tenant compte de la localisation génomique des régions conservées et de leur co-localisation avec divers signaux de présence des ARNnc, tels que les terminateurs ou les promoteurs. Parmi ces travaux, nous pouvons citer deux travaux pionniers sur la recherche systématique des ARNnc, Wassarman *et al.* (2001) et Argaman *et al.* (2001)), qui ont permis la détection et la validation d'un grand nombre d'ARNnc chez *E. coli*. Dans ces travaux les analyses ont, pour la plupart, été faites manuellement. Le premier outil intégrant les résultats de conservation entre deux génomes et de prédiction des différents éléments indicateurs de la présence des ARNnc dans un processus automatisé est sRNAPredict (Livny *et al.* (2005)). Une deuxième version de cet outil, sRNAPredict2, est capable de gérer les prédiction redondantes, permettant ainsi son utilisation dans le cas de la comparaison de plus de deux génomes (Livny *et al.* (2006)). sRNAPredict a été utilisé avec succès dans plusieurs études portant sur la recherche systématique des ARNnc dans les génomes bactériens (Livny *et al.* (2005), Livny *et al.* (2006), Valverde *et al.* (2008)).

Les différentes études de recherche systématique des ARNnc dans les génomes bactériens seront présentées plus amplement dans la partie suivante du manuscript (section 10).

Je finirai cette présentation des méthodes comparatives existantes en décrivant plus

particulièrement l'idée et le fonctionnement de QRNA qui a été intégré dans l'outil de prédiction que j'ai utilisé au cours de ma thèse pour prédire des ARNnc candidats de *R. solanacearum*.

6.1.0.1 QRNA

L'idée formalisée dans le modèle sous-jacent à QRNA (Rivas *et al.* (2001)) est que le caractère des mismatch dans un alignement de séquences homologues sera différent selon qu'il s'agit de séquences codantes, de séquences d'ARNnc (ou autre élément structuré) ou de séquences appartenant au reste du génome (illustré sur la figure 6.2). Dans le cas des séquences codantes, le principe de QRNA est de considérer que les mutations devraient plus souvent correspondre à des mutations synonymes ou à celles engendrant des acides aminés proches. En revanche, dans le modèle correspondant aux séquences d'ARNnc, la présence des mutations compensatoires dans les éléments structurés est attendue. Le reste du génome est modélisé par un modèle de probabilités d'appariement indépendante de la position.

Pour chaque alignement soumis à QRNA, un score est calculé en fonction de chacun des modèles présentés ci-dessus. Le plus élevé permet d'associer un état à un alignement.

Un inconvénient de QRNA est le grand nombre de paramètres liés à chacun des modèles le composant. Ces paramètres ont été optimisés en partie en étudiant la spécificité et la sensibilité de QRNA dans des ensembles de séquences simulées avec un G+C% équilibré et en partie sur des ensembles d'alignement des ARNr et ARNt. De ce fait, l'ensemble de paramètres proposé par défaut peut être biaisé dans le cadre de son utilisation pour la détection des ARNnc des familles autres que ARNr ou ARNt, et surtout dans le cas de leur utilisation dans les génomes ayant un G+C% plus élevé.

Son utilisation est limitée en longueur de séquence, la complexité de l'algorithme étant dominée par la complexité dans le cadre du calcul de score de l'état ARN, en $O(n^3)$, où n est la longueur de l'alignement. Dans le cas de l'analyse d'alignements plus longs, ces alignements sont découpés et analysés par tranche.

Je présenterai maintenant l'outil de détection des ARNnc auquel j'ai contribué durant ma thèse.

6.2 Motivation du développement d'un nouvel outil

Au moment du démarrage de ce projet, lorsque nous voulions utiliser une approche comparative pour la détection des nouveaux ARNnc dans un organisme, le choix pouvait se porter soit sur les outils prenant en compte la conservation en séquence et en structure soit sur sRNAPredict, prenant en compte la conservation en séquence. Or, la prédiction dans plus de deux génomes se révèle difficile car l'utilisation de QRNA et de sRNAPredict étaient limitées aux alignements deux-à-deux tandis que RNAz et d'autres outils du même type exige des alignements muliples de bonne qualité, impossibles à générer dans le cas de l'analyse des génomes entiers.

Or, lorsque les séquences de plus de deux génomes sont disponibles, la prise en compte simultanée de leur conservation renforce les prédictions.

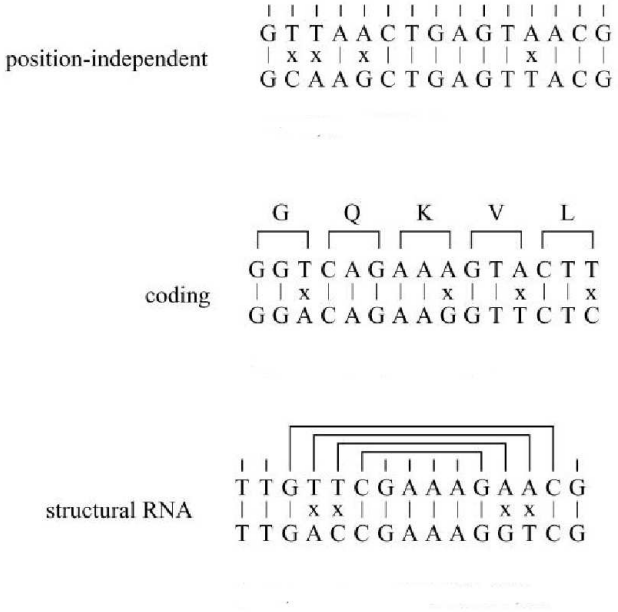


FIG. 6.2 – Trois alignements deux-à-deux ayant la même composition et le même nombre de mismatch peuvent être classés selon la nature de ceux-ci : l'indépendance des bases (en haut), les séquences codantes (milieu) et les ARNnc structurés (en bas). Le calcul de probabilité dans chacun des modèles est indiqué au-dessous des alignements correspondants : position par position, dans le cas de l'indépendances des bases (c'est-à-dire ni codant ni ARNnc), par codon dans le cas des séquences codantes et comme deux appariements dans le cas des ARNnc. Schéma emprunté de Rivas *et al.* (2001).

6.3 Existant au début de la thèse

En 2005, une approche exploitant QRNA et l'étendant à l'analyse de plusieurs génomes a été dévéloppée dans l'équipe, par Céline Noirot (Noirot (2005)). L'outil correspondant, nommé RNAsim, permet d'inférer, à partir des résultats d'un outil de détection des ARNnc deux-à-deux, en passant par une représentation par graphe, les prédictions d'ARNnc dans n génomes (fig. 6.3).

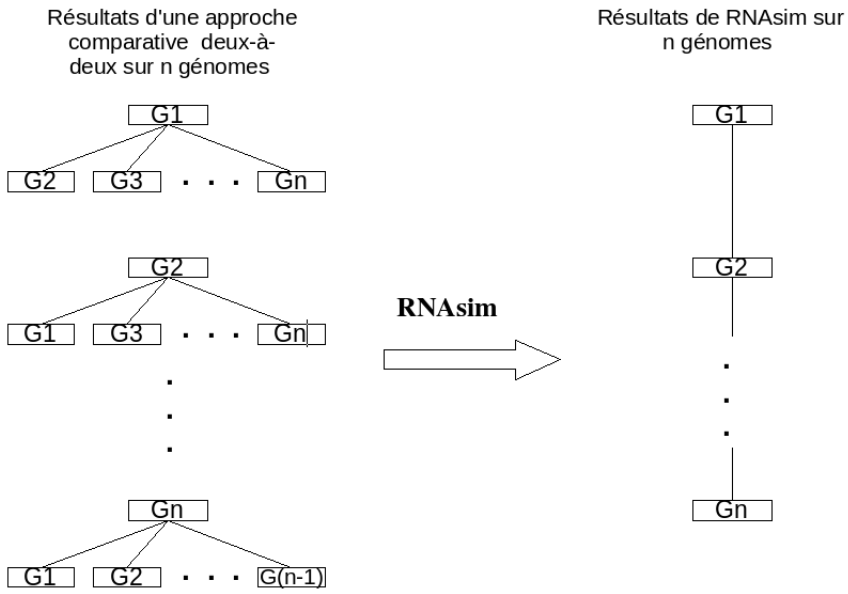


FIG. 6.3 – L'objectif de RNAsim : inférer les prédictions ARNnc dans n génomes notés (G1,G2, ..., Gn) à partir des prédictions deux-à-deux.

Le choix de l'outil de détection des ARNnc deux-à-deux s'est porté sur QRNA, qui a été utilisé avec succès dans plusieurs génomes bactériens (Rivas *et al.* (2001), Silvaggi *et al.* (2006a), del Val *et al.* (2007)).

Ce travail a donné lieu à une première version de RNAsim. RNAsim est préseté en détail dans le chapitre suivant.

Chapitre 7

RNAsim

La disponibilité des séquences de deux nouvelles souches séquencées, IPO1609 et Molk2, de *R. solanacearum* rendait l'approche RNAsim particulièrement adaptée dans le cadre de la thèse.

Une alternative à l'utilisation de RNAsim était, d'une part l'utilisation de sRNAPredict2, dont la deuxième version développée (Livny *et al.* (2006)) propose une approche infèrant les prédictions dans plus de deux génomes et d'autre part l'utilisation de RNAz, avec l'utilisation des nouveaux outils d'alignements multiples des génomes entiers (voir la section 6). Néanmoins, ces deux approches ne tenaient pas compte de la totalité des éléments que nous voulions inclure dans notre analyse, tous utiles dans une approche d'analyse minutieuse des résultats de conservation : des indications sur la "synthénie" des prédictions dans les différents génomes (conservation du contexte génomique des prédictions), les indications sur la conservation dans la totalité des génomes considérés, la présence des copies multiples d'une région conservée.

Il aurait été donc nécessaire de modifier les outils existants de façon considérable afin de les adapter à nos besoins. RNAsim s'est imposé comme une meilleure solution, étant donné l'accessibilité du code et l'aide à la prise en main de la problématique et de l'outil, disponible du fait qu'il a été développé dans le laboratoire.

Pour ces raisons, nous avons décidé de poursuivre le développement de RNAsim dans le cadre de ma thèse pour ensuite l'appliquer au génome de *R. solanacearum*.

La correction de certaines erreurs a d'abord été nécessaire. Ensuite, nous avons apporté un certain nombre d'améliorations, consistant dans :

1. l'introduction d'une nouvelle étape dans la chaîne de traitement des données ;

2. l'intégration de nouvelles fonctionalités facilitant l'analyse des sorties.

Dans ce chapitre je présenterai le principe de fonctionnement de RNAsim ainsi que les grandes lignes de sa mise en oeuvre et de ses fonctionnalités.

7.1 Modélisation du problème

L'usage des graphes et de leurs propriétés est courante en bioinformatique. Ils sont utilisés pour l'analyse d'expression des gènes (Brors (2005)), la modélisation de réseaux de régulation géniques (Huber *et al.* (2007)), la comparaison etla modélisation des structures secondaires des ARNnc (Le *et al.* (1989)) mais aussi, à un méta-niveau, pour la représentation des connaissances biologiques (Ashburner *et al.* (2000)).

La méthode présentée ici s'appuie sur une modélisation par un graphe afin de prédire des ARNnc communs à n génomes à partir de résultats de comparaison deux-à-deux. Nous commençerons sa présentation en rappelant les définitions et les notions nécessaires à sa mise en oeuvre. Ensuite nous présenterons la traduction des objets biologiques manipulés, à savoir les séquences appartenant aux différents organismes, en langage des graphes.

7.1.1 Notions de base

7.1.1.1 Notions sur les graphes

Les définitions exposées ici sont reprises du livre Graph Theory de Diestel (1997). Nous commençons par rappeler la définition d'un graphe.

Définition Un graphe G est défini par un couple $G = (V, E)$ où V représente l'ensemble des sommets du graphe et E est un sous-ensemble de $[V]^2$. Les éléments de E sont appelés "arêtes". Le graphe G est *non-vide* si V n'est pas vide. Le nombre d'éléments de G est noté par $|G|$ et est appelé *l'ordre* du graphe G.

Si pour $u, v \in V$, $(u, v) \in E$, alors les sommets u et v sont dits *adjacents* ou *reliés par une arête*. La liste de tous les sommets adjacents à un sommet $v \in V$ est appelée sa *liste d'adjacence*. Un sommet du graphe G est dit *isolé* s'il n'est adjacent à aucun sommet de G.

Soit maintenant $G' = (V', E')$. Le graphe G' est un *sous-graphe* de G si et seulement si $V' \subseteq V$ et $E' \subseteq E$. Le sous-graphe $G' = (V', E')$ est dit *induit* par G si $E' = [V']^2 \cap E$.

Une chaîne dans le graphe G est un sous-graphe de G, $P = (V_P, E_P)$, tel que $V_P = \{x_0, x_1, \ldots, x_k\}$, $E_P = \{\{x_0, x_1\}, \{x_1, x_2\}, \ldots, \{x_{k-1}, x_k\}\}$ et que $x_i, 0 \leq i \leq k$ sont tous différents. On dit que x_0 et x_k sont connectés par le chemin P ou que P connecte x_0 et x_k.

Un graphe $G = (V, E)$ non-vide est dit *connexe* si pour tout $u, v \in V$ il existe un chaîne dans G qui les connecte.

Définition d'une composantes connexe Soit $G = (V, E)$ un graphe et soit $G' = (V', E')$ un sous-graphe induit de G. Si G' est connexe et si pour tout sous-graphe induit de G, $G'' = (V'', E'')$ tel que $V' \subset V''$, G'' n'est pas connexe, alors G' est *une composante connexe* de G.
L'illustration d'un graphe et de ses composantes connexes sont données dans la figure 7.1.

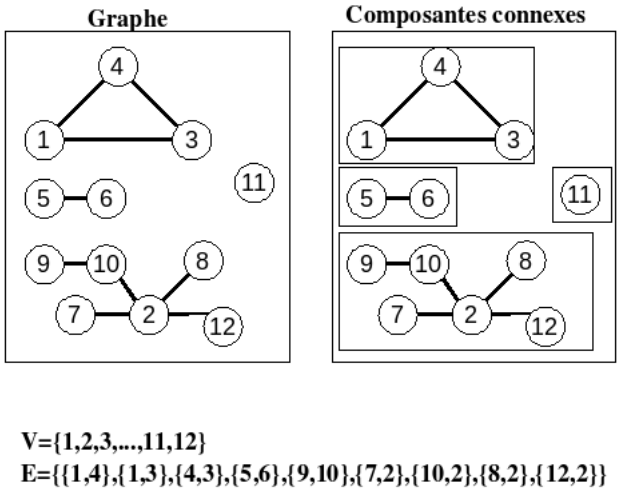


FIG. 7.1 – À gauche : l'illustration d'un graphe. V est l'ensemble des sommets et E l'ensemble des arêtes. À droite : les composantes connexes de ce graphe sont encadrées.

7.1.1.2 L'algorithme de recherche des composantes connexes

La recherche des composantes connexes au sein d'un graphe peut s'effectuer à l'aide d'un algorithme dit de Depth-First-Search (DFS) (voir, par exemple le livre "Algorithms and Theory of Computation Handbook", Atallah (1998)).

L'algorithme initialise d'abord tous les sommets du graphe comme non visités. Le parcours du graphe commence par un sommet arbitraire. Pour un sommet donné non visité, un numéro de composante connexe est attribué, et sa liste d'adjacence est parcourue récursivement par la procédure DFS. Tous les sommets visités ainsi se voient attribuer le même numéro de composante connexe. Lorsque toute la liste d'adjacence est parcourue récursivement la procédure DFS se termine. Ainsi, toute la composante connexe de l'élément à partir duquel DFS a été lancé est visitée. Ensuite, le graphe est parcouru pour les sommets non-visités restants et DFS est lancé à partir de ces sommets. Lorsque il ne reste plus de sommets non-visités dans le graphe l'algorithme se termine. L'algorithme de DFS est décrit en pseudo-code dans les Algorithmes 1 et 2.

La complexité temporelle est $O(|V| + |E|)$ et la complexité spatiale est $O(h)$, où h est la longueur du plus long chemin dans le graphe.

7.1.2 Formalisation du problème

Je présenterai d'abord la formalisation du passage d'un ensemble d'alignement à une représentation par graphe qui a été réalisée dans le cadre de stage de C. Noirot. Ensuite, dans la partie 7.1.2.2, je présenterai ma contribution à une modélisation plus élaborée.

7.1.2.1 Formalisation dans le cadre de la théorie des graphes

Nous supposons à présent que nous disposons d'un ensemble d'alignements deux-à-deux entre les séquences issues de n génomes. Nous appelerons cet ensemble d'alignements *un réseau d'alignements.*

Représentation par graphe d'un réseau d'alignements Si, dans un réseau d'alignements R, un alignement existe entre une séquence A de l'organisme $O(A)$ appartenant au réplicon $C(A)$, aux coordonnées génomiques $x(A)$ et $y(A)$, se trouvant sur le brin $S(A)$ et une séquence B de l'organisme $O(B)$ appartenant au réplicon $C(B)$,

Algorithm 1 DFS (G) : première partie : algorithme parcours les arêtes non-visités

```
Initialisation :
i=0 /* indice utilisé pour le parcours des sommets */
c=0 /*l'indicateur de la composante connexe*/
for (i=1 ;i<|G| ;i++) do
   /* v(i) est le i-ième sommet du graphe*/
   visité(v(i))=faux
   fini(v(i))=faux
   père(v(i))=null
end for

for tous les sommets v de G faire do
   if non visité(v(i)) then
      c=c+1
      DSF(v,c)
   end if
end for
```

Algorithm 2 DFS (v,c) : deuxième partie, parcours récursif de la liste d'adjacence d'un élément non-visité

```
DSF(v,c)
visité(v)=vrai
composante(v)=c
for tout w dans adj(v) /*adj(v) : la liste d'adjacence du sommet v*/ do
   if non visité(w) then
      père(w)=v
      DSF(v,c)
   end if
end for
```

aux coordonnées génomiques $x(B)$ et $y(B)$, se trouvant sur le brin $S(B)$, alors il peut être représenté par le couple (A, B), où $A = (O(A), C(A), x(A), y(A), S(A))$ et $Y = (O(B), C(B), x(B), y(B), S(B))$. Nous dirons alors que A et B sont les séquences appartenant au réseau d'alignements R et que A et B sont alignés dans R.

Nous pouvons, à présent, introduire le graphe associé à un réseau d'alignements R. *Le graphe d'alignement* G_R associé au réseau d'alignement R est défini comme le graphe ayant comme ensemble de sommets V_R, l'ensemble de toutes les séquences du réseau d'alignements R. L'ensemble des arêtes, E_R, est le sous-ensemble de $V_R \times V_R$ tel que $(X, Y) \in E_R$ si et seulement si un alignement existe entre X et Y dans R.

Dans un graphe ainsi défini, la recherche des composantes connexes permet d'identifier des groupes de séquences qui présentent une similarité entre elles : soit directement parce qu'elles s'alignent dans le réseau d'alignement, soit indirectement à travers un ou plusieurs autres alignements.

Le passage des alignements issus des comparaison de n génomes à un graphe était formalisé et implémenté dans la première version de RNAsim. La figure 7.2 présente de façon schématisée, le passage de l'ensemble des alignements deux-à-deux à un graphe représentant le réseau d'alignements.

Quelques remarques

- Notons que, dans cette représentation, un sommet du graphe G_R peut être adjacent à plusieurs autres sommets du graphe. Pour que ceci arrive, il suffit qu'une même séquence du réseau d'alignement R soit alignée avec d'autres séquences de R.
- Un autre constat est qu'aucun élément d'un graphe défini ainsi ne peut être un élément isolé. En effet, les sommets correspondent aux séquences d'un réseau d'alignement, et celui ci contient uniquement les séquences se trouvant dans un alignement.
- Quant à l'utilisation de l'algorithme de recherche des composantes connexes, notons que son utilisation fait apparaître une notion de transitivité entre les alignements. En effet, lorsque deux sommets de trouvent dans la même composante connexe il ne sont pas nécessairement reliés par une arrête, mais l'existence d'un lien, même indirect, entre eux, apparaîtra du fait de l'appartenance à la même composante connexe (les cas des séquences A2 et B2 qui sont liées entre elles

indirectement, par les alignements avec C2 et se trouvent dans la même composante connexe, fig. 7.2). Ceci a un sens biologique : nous pouvons, par exemple, imaginer que les organismes A et C, et B et C sont plus proches phylogénétiquement que A et B ne le sont entre eux. Leurs séquences respectives ont donc pu diverger entre elles au point de ne pas pouvoir être retrouvées par une recherche d'homologie. Néanmoins, leur lien peut être retrouvé en passant par l'homologie avec l'organisme B. Comme alternative à cette modélisation, la représentation des séquences reliées entre elles aurait pu être faite à l'aide des cliques, où tous les sommets du sous-graphe sont reliés entre eux. Ceci aurait permis de retenir

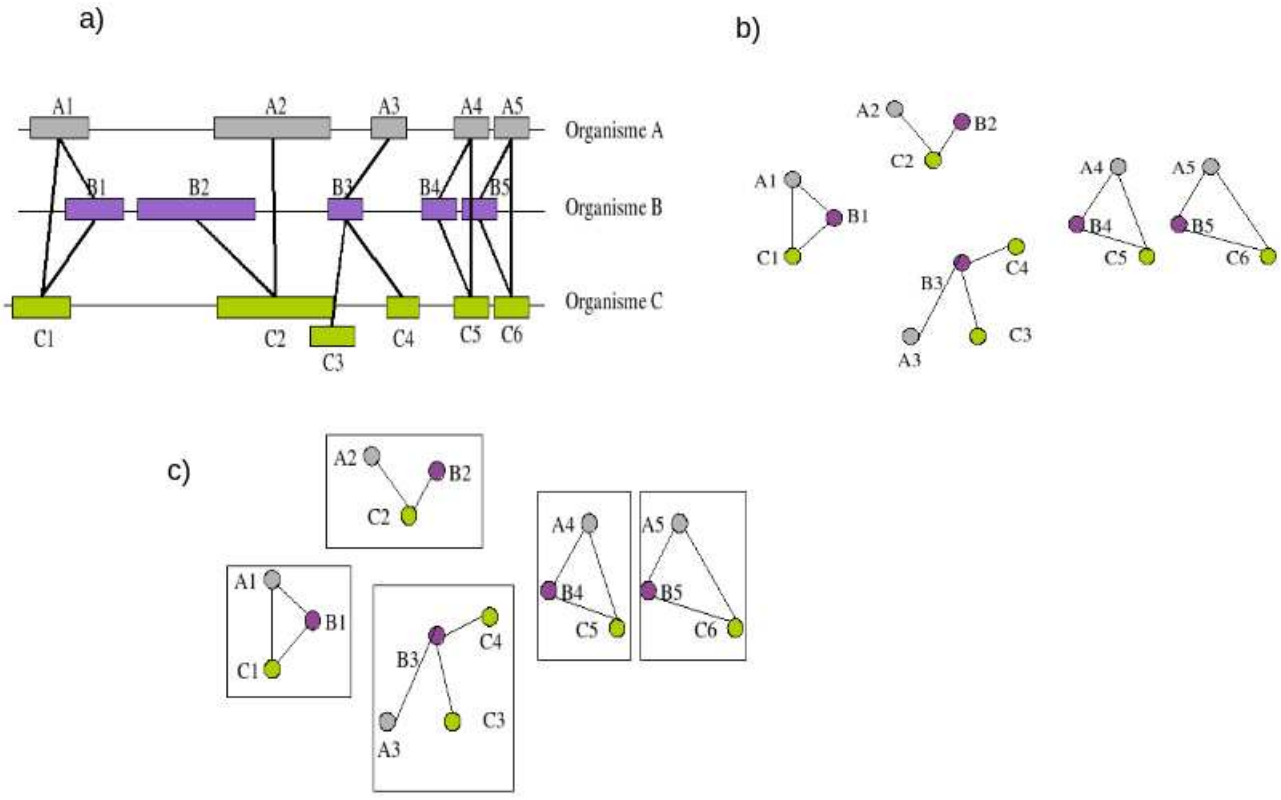


FIG. 7.2 – La représentation schématisée de passage d'un ensemble d'alignements entre les séquences de trois organismes au graphe associé. a) Les rectangles représentent des régions impliquées dans un alignement. Les régions C2 et C3, dans l'organisme C, se chevauchent. b) Le graphe associé. c) Les composantes connexes du graphe associé sont encadrées.

les séquences conservées dans la totalité des génomes considérés mais aurait, en même temps éliminer la notion de transitivité décrite précedemment.
Nous en sommes restés à la modélisation par composantes connexes afin de permettre l'apparition des homologies éloignées dont la prise en compte nous a semblé particulièrement adaptée dans le cadre de la recherche des ARNnc, divergeant rapidement en séquence.

- Notons enfin que cette formalisation fait l'abstraction des scores de pertinence des alignements donnés habituellement par les outils d'alignement deux-à-deux. Cette information pourrait être prise en compte, par exemple, en associant des poids aux arêtes du graphe d'alignement.

7.1.2.2 Deuxième formalisation

Motivation Dans l'exemple présenté sur la figure 7.2 a) les séquences C2 et C3, de l'organisme C, se chevauchent. Cette information est perdue dans le graphe associé, tel qu'il est défini dans 7.1.2.1 (figure 7.2 b) et c)). Or, un tel chevauchement peut indiquer la présence d'une sous-région conservée, correspondant à la région de chevauchement, qui serait conservée dans toutes ces séquences.

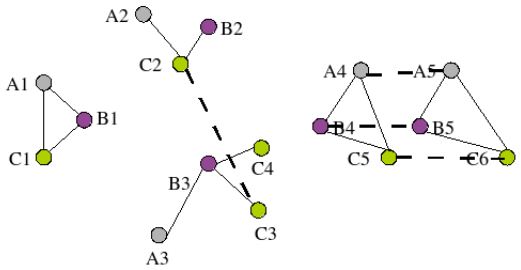


FIG. 7.3 – Représentation du graphe d'alignements avec la prise en compte du voisinage : la graphe correspondant au réseau d'alignement 7.2a). L'arête $(C2, C3)$ est rajoutée car C2 et C3 sont chevauchants.

Pour cette raison, dans la deuxième version de RNAsim, nous avons introduit un autre type d'arêtes dans le graphe d'alignement, entre les sommets correspondant des deux séquences du même réplicon d'un organisme, qui se chevauchent entre elles (l'arête entre les sommets C2 et C3, fig. 7.3).

Un autre cas de figure peut se présenter lorsque les séquences du réseau d'alignement appartenant au même génome et au même replicon sont situées à une courte distance l'une de l'autre (le cas des séquences A4 et A5, B4 et B5 ou bien C5 et C6, sur la fig. 7.2). Ces séquences, situées à proximité immédiate peuvent, en réalité, appartenir à un même élément dont deux régions conservées sont espacées par une courte région variable. Il peut être judicieux d'indiquer leur lien "spatial" par l'attribution d'une arête entre de tels sommets (fig. 7.3) et éviter ainsi qu'ils soient considérés séparément.

Par conséquent, nous modéliserons, dans le graphe d'alignements, ces deux cas de figure en les englobant dans une notion de *voisinage*, où le chevauchement entre deux séquences sera vu comme le cas spécial de la relation du voisinage. Des séquences proches engendreront des sommets voisins qui, eux, seront reliés par une arête dans le graphe d'alignements.

Dans un deuxième temps, nous proposons une opération de "fusion" des sommets (ou prédictions) voisins afin de réduire le nombre de sommets dans le graphe.

Ces nouvelles connections entre les sommets d'un graphe d'alignements ont été définies dans la deuxième modélisation à laquelle j'ai contribué.

Formalisation *Voisinage*

Soit $v \in \mathcal{Z}$. Nous dirons que deux sommets $X1$ et $X2$ sont *sommets v-voisins* si et seulement si $O(X1) = O(X2)$ et $C(X1) = C(X2)$ et

$$\big(x(X_1) - y(X_2) \leq v \text{ et } x(X_1) \geq x(X_2)\big) \text{ ou } \big(x(X_2) - y(X_1) \leq v \text{ et } x(X_2) \geq x(X_1)\big)$$

Lorsque $X1$ et $X2$ sont v-voisins et $v < 0$ on dira que $X1$ et $X2$ se chevauchent.

Pour un v fixé, nous définissons les arêtes associées à la relation de v-voisinage, comme les arêtes entre tous les $X, Y \in V$ tels que X et Y sont v voisins.

Nous définissons maintenant l'opération de "fusion", qui transforme tous les sommets v-voisins en un seul sommet.

Fusion

Soit H une composante connexe du graphe d'alignements G_R associé au réseau d'alignements R et soient $V_v(X), X \in H$ tous les sommets appartenant à H qui sont des v voisins de X. L'opération de fusion crée d'abord un nouveau sommet X' tel que $O(X') = O(X)$, $C(X') = C(X)$, $B(X') = B(X)$ et que

$$x(X') = \min(x(V_v(X))) \text{ et } y(X') = \max(y(V_v(X)))$$

et remplace ensuite $V_v(X)$ dans G_R par X', en remplaçant toute arête $\{Y, X_v\} \in E$, pour tout $X_v \in V_v(X)$ par une arête $\{Y, X'\}$.

7.1.2.3 Conclusion

Dans cette partie j'ai présenté la représentation d'un réseau d'alignements par un graphe et l'utilisation de cette représentation pour inférer la conservation dans n génomes à partir des alignements deux-à-deux.

La modélisation a été élaborée dans la première version de RNAsim, représentant les séquences alignées par les sommets du graphe et la relation d'homologie entre deux séquences, par une arête dans le graphe entre les sommets correspondants.

En me fondant sur cette modélisation, j'ai rajouté la prise en compte du voisinage des deux séquences en introduisant les arêtes entre les sommets correspondant aux séquences du même génome dont les coordonnées génomiques sont proches. Ce nouvel élément ainsi que l'opération "fusion" de tous les sommets voisins en un seul sommet ont permis de réduire considérablement le nombre de sommets dans le graphe (voir la section 7.2.2.4).

Dans la suite du manuscrit, je présente la mise en oeuvre pratique de la méthode présentée dans RNAsim.

7.2 Implémentation de RNAsim

L'outil *RNAsim* met en pratique la méthode présentée. Il prend en entrée des alignements définis par un couple de séquences avec réplicons et organismes auquels elles appartiennent, ainsi que leurs coordonnées génomiques. Nous ne précisons donc pas le type d'alignement ni l'appartenance des séquences à une catégorie spécifique.

En effet, la méthode présentée peut être appliquée à tout type d'alignement deux-à-deux (les alignements du type Blast, Fasta, Yass ou autres, voir Brown (2008) pour une revue partielle des outils d'outils d'alignements deux-à-deux) tandis que les séquences alignées peuvent, *a priori*, appartenir à tout groupe des séquences génomiques (séquences codantes, intergéniques, régions 5' UTR etc.).

Dans son implémentation pratique pour la recherche des ARNnc, RNAsim se focalise sur les séquences appartenant aux régions intergéniques "vides" (ne contenant pas de séquences codantes annotées sur aucun des deux brins).

Dans le but de restreindre le nombre d'alignements considérés aux plus pertinents d'un point de vu de la recherche des ARNnc, l'application de différents filtres sur les séquences et sur les alignements peut être envisagée.

Dans la partie qui suit, nous présentons d'abord les filtres intégrés dans RNAsim et ensuite les différentes étapes permettant d'arriver aux candidats ARNnc dans plusieurs génomes à partir des séquences génomiques annotées.

7.2.1 Filtres sur les séquences et les alignements

Longueur des régions intergéniques à considérer

Un premier filtre intégré dans RNAsim est la taille minimale des régions intergéniques considérées pour effectuer les recherches de similarités avec Blast. Les régions intergéniques courtes, classiquement de taille inférieure à 100nt, sont nombreuses dans les génomes bactériens. La plupart des ARNnc découverts à ce jour se trouvant dans des régions intergéniques plus longues (Hershberg *et al.* (2003)), l'élimination des régions intergéniques au-dessous d'une certaine longueur n'élimine probablement pas ou peu de séquences contenant de vrais ARNnc. De plus, ces courtes séquences, localisées souvent dans les opérons, étant en général, très conservées, apportent du "bruit" supplémentaire aux résultats.

Filtre sur des régions intergéniques spécifiques

Dans certains cas, il peut être souhaitable d'exclure certaines régions spécifiques de l'analyse. Ceci est possible dans RNAsim, sous condition de connaître les coordonnées génomiques de ces régions et de les soummetre au bon format.

Les filtres sur les alignements Blast

Il est possible de fixer la longueur minimale, la E-value et le taux d'identité des alignements Blast à considérer dans la construction du graphe. Cette option a été créée afin de permettre l'élimination des alignements courts ou à faible pourcentage d'identité dont la signification biologique est discutable et qui encombrent les résultats.

Ces filtres sont actuellement utilisés dans RNAsim à travers leur implémentation dans QRNA (voir le filtre suivant).

Filtre QRNA

QRNA analyse un alignement deux-à-deux afin d'identifier l'existence de mutations compensatoires au sein des séquences alignées et par conséquent l'existence de struc-

ture secondaire conservée (voir section 6.1.0.1). Ce filtre permet donc de restreindre les alignements à ceux qui sont potentiellement structurées et d'en diminuer considérablement le nombre.

En revanche, il présente l'inconvénient de l'augmentation considérable du temps d'exécution de RNAsim ainsi que la nécessité d'introduire une nouvelle étape dans RNAsim permettant d'effectuer les assemblages des séquences découpées par QRNA (voir 6.1.0.1 et 8.7).

7.2.2 Etapes de RNAsim

Nous présentons ici les principales étapes de RNAsim en suivant un exemple simple de comparaison de trois génomes.

7.2.2.1 Les entrées RNAsim

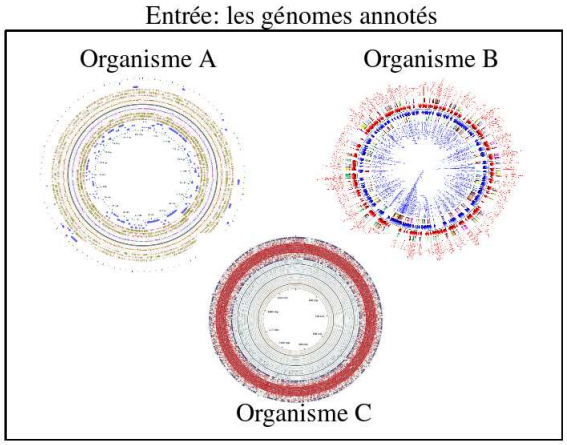


FIG. 7.4 – Entrées RNAsim : les séquences annotées de trois génomes A, B et C (représentées ici sous la forme circulaire). A partir de l'annotation de chacun des génomes, les régions intergéniques sont extraites et comparées deux-à-deux à l'aide de WU-Blast.

Au départ nous devons disposer de plusieurs séquences génomiques d'organismes suffisamment proches pour être susceptibles de contenir des régions conservées codant pour des ARNnc. Ces séquences doivent disposer d'une annotation des gènes et notamment de leurs positions afin que les régions intergéniques puissent être extraites (schématisé dans la figure 7.4 dans le cas de comparaisons entre trois génomes). Si,

de plus, cetaines régions sont à exclure de l'analyse, nous devons disposer de leurs coordonnées génomiques.

7.2.2.2 Préparation des données

Les régions intergéniques sont extraites de n organismes et pour chaque couple d'organismes ($n(n-1)/2$ en tout), les comparaisons entre leurs régions intergéniques respectives sont faites à l'aide de WU-Blast (Gish (1996)).

Sur les alignements Blast obtenus, le filtre de longueur d'alignement minimal, de taux d'identité et de la E-value sont appliqués (si de tels filtres sont définis par l'utilisateur). Le filtre QRNA est appliqué ensuite.

A partir des alignements retenus par QRNA (les alignements étiquetés par QRNA comme ARN), un certain nombre d'opération sont effectuées sur les alignements afin de "recoller" les alignements découpés par QRNA et fusionner les alignements redondants.

7.2.2.3 Passage de l'alignement au graphe et construction des composantes connexes

Dorénavant nous supposerons que le paramètre de voisinage est v fixé à 20.

Le graphe d'alignements est construit à partir des alignements retenus (fig. 7.5a), à gauche). Notons que les sommets B2 et B3, A2 et A3, et, C2 et C3, se trouvent à une distance inférieure à 20nt et par conséquent les sommets correspondants sont reliés par les arêtes, représentées en pointillé, indiquant qu'il s'agit des arêtes associées aux sommets voisins, pour $v = 20$.

Les composantes connexes de ce graphe sont alors calculées (fig. 7.5b).

7.2.2.4 Fusion

L'étape de la fusion permet de réduire le nombre d'éléments dans une composante connexe en "fusionnant" tous les éléments voisins (fig. 7.5c)), démarche décrite dans la section 7.1.2.2).

Notons qu'en procédant ainsi, le nombre d'éléments des composantes connexes peut être réduit considérablement, spécialement dans le cas des composantes connexes de grande taille (fig. 7.6). Par exemple, sur un jeu de données que nous avons utilisé, la composante connexe comportant 884 éléments initialement a été réduite à 44 éléments,

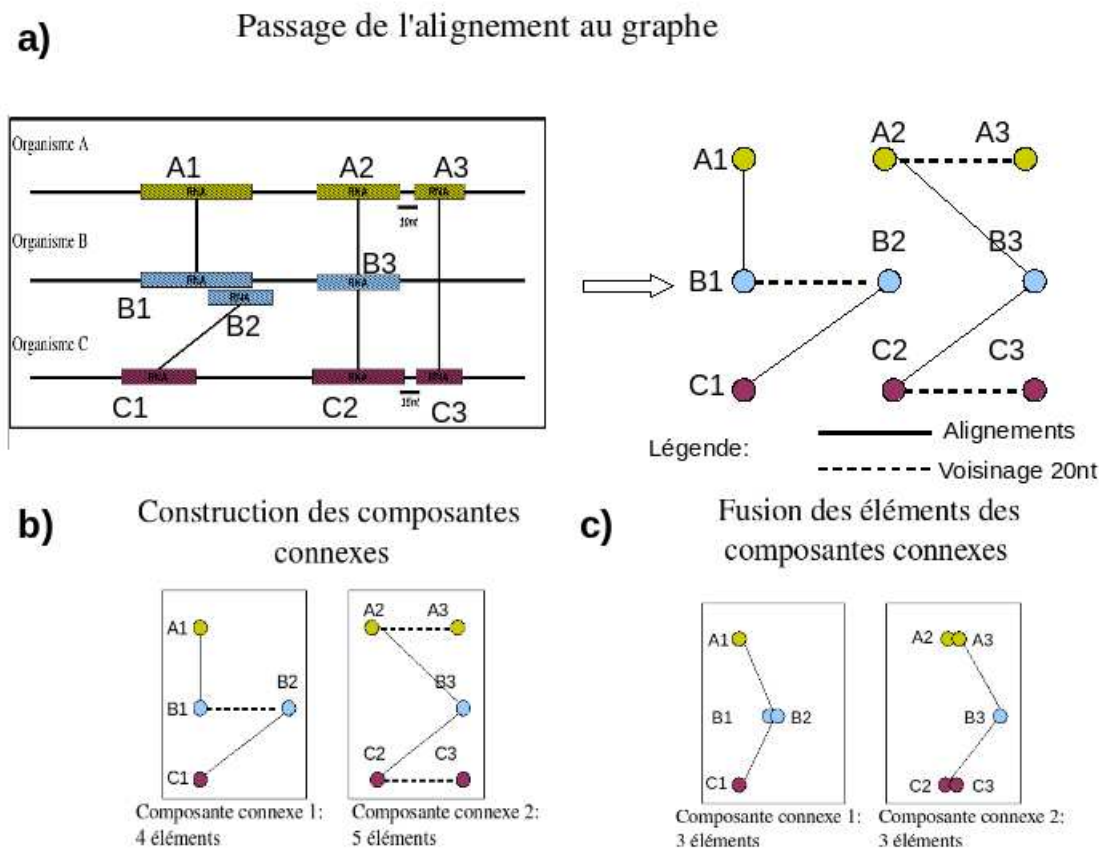


FIG. 7.5 – (a)Passage des alignements retenus (indiqués en rectangles RNA) au graphe d'alignements : les régions as"alignant correspondent au sommets. Les arêtes relient les sommets correspondant aux séquences présentes dans un même alignement (indiquées par les traits pleins) et aux sommets voisins pour un voisinage fixé à $v = 20$ (indiqués par les traits en pointillés). (b) Les composantes connexes du graphe d'alignement. Chacune des deux composantes connexes est encadrée. (c) Fusion : les sommets voisins sont fusionnés en un sommet.

ce procédé réduisant donc sa taille d'un facteur 20. Notons que l'étape de fusion a été intégrée plus tard et après l'analyse des premiers résultats obtenus sans cette étape. En effet, le nombre d'éléments dans certaines composantes connexes était trop élevé pour effectuer une analyse efficace (dans les cas extrêmes une composante connexe comprend plus de la moitié des sommets du graphe d'alignements). L'étape de fusion a donc permis, (comme le montre la fig. 7.6) de réduire la taille des composante connexes tout en gardant la cohérence des résultats.

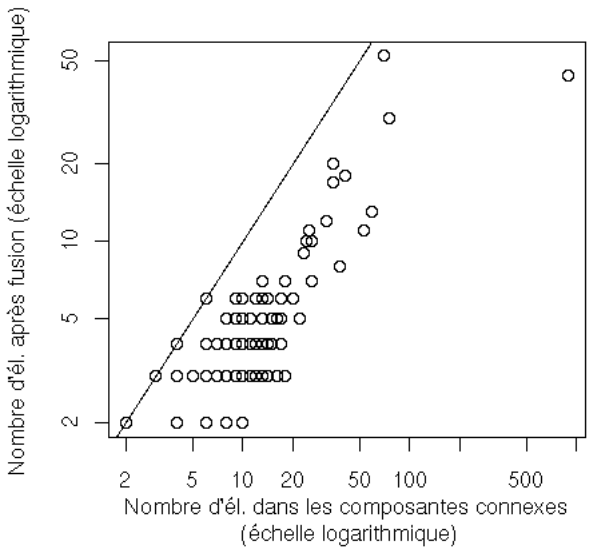


FIG. 7.6 – Réduction du nombre d'éléments dans les composantes connexes suite à l'étape de fusion, présentée en échelle logarithmique. Les données de trois souches de *R. solanacearum* (section 8.1). La droite $x = y$ indique les composantes connexes où le nombre d'éléments n'a pas changé après la fusion.

7.2.3 Temps d'execution

A partir du moment où les données sont préparées (voir section 7.2.2.2) le temps d'exécution de RNAsim est essentiellement dû au temps d'exécutions des comparaisons deux-à-deux de toutes les régions intergéniques des $n(n-1)/2$ organismes à l'aide de Blast aux analyse QRNA correspondantes.

7.3 Sorties et autres fonctionalités

Nous décrirons ici la forme sous-laquelle les sorties RNAsim sont présentées, ainsi que les post-traitements effectués sur ces sorties efin d'optimiser et de faciliter l'analyse des résultats.

7.3.1 Sorties

Les résultats de RNAsim sont disponibles via une page web (http ://cat.toulouse.inra.fr /~ akozomara/cgi-bin/RNAsim/RNAsim.html). Ils sont présentés sous la forme d'un tableau, regroupés en composantes connexes et triés en rang croissant selon la taille des composantes (un aperçu de la présentation d'une composante connexe est donné sur la figure 7.7 a).

Les éléments de chaque composante connexe sont présentés avec leurs coordonnées génomiques, l'indication du brin d'ADN sur lequel la séquence a été identifiée et leur longueur (fig. 7.7). Le G+C% de chaque élément est calculé : ceci a permis d'avoir un aperçu de leur composition et permet de voir si cette composition s'écarte du G+C% moyen génomique (voir section 10.1). Ce critère est également utile pour la prédiction des ARNnc compte tenu du fait de l'étude du biais en composition des ARNnc réalisée dans la première partie de la thèse (section 4.2.2.3). Le tableau donne également un accès aux fichiers des séquences au format FASTA et GFF (General-Finding Format ou General Feature Format). Ce dernier format permet l'intégration directe des résultats issus de RNAsim dans l'outil de visualisation des génomes Apollo (Lewis *et al.* (2002)) et son extension ApolloRNA (Cros *et al.* (2007)).

7.3.2 Post-traitement des sorties RNAsim

Différents outils ont été appliqués de façon systématique sur les données issues de RNAsim afin de faciliter leur analyse. Ces outils sont de 3 types, à savoir i) alignement des éléments des composantes connexes, ii) présentation du contexte génomique autour des séquences candidates et iii) prédiction de la structure secondaire de ces séquences. A chaque fois deux outils, mettant en oeuvre des principes différents, sont utilisés pour compléter l'analyse des différentes composantes connexes.

Ainsi, pour chaque composante connexe issue de RNAsim, les éléments suivants sont

(a) Présentation d'une composante connexe

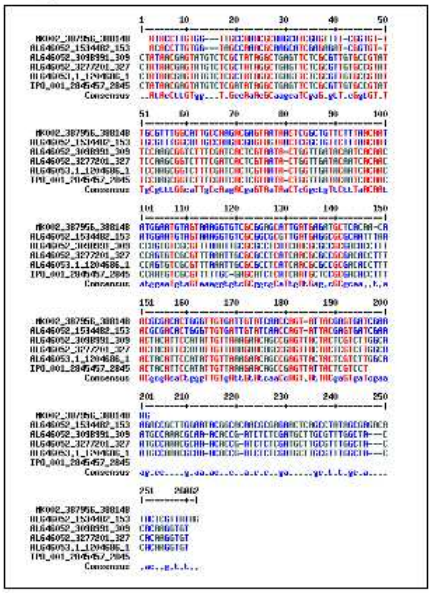
Taille de la composante:6

Fichier des gènes voisins: 6.export (e)

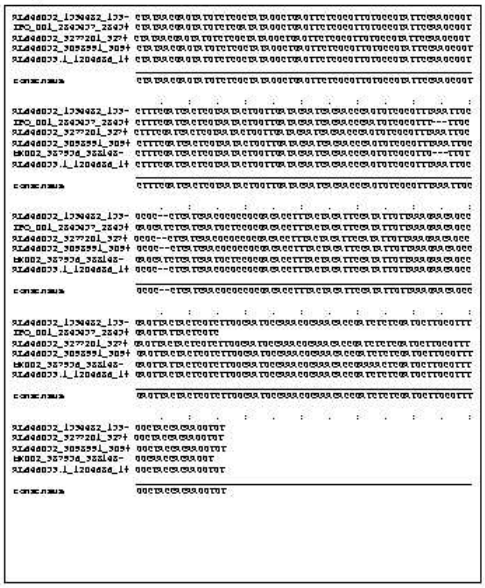
Organisme	Accession	begin	end	brin	Longueur	G+C%	Fasta	Img	GFF
Align.multiple (b)	Assemblage (c)								
MOLK	MK002	387956	388148	+	192	0.442708	Fasta	Img	GFF
Ralstonia	RSc	1534482	1534735	+	253	0.478261	Fasta	Img	GFF
Ralstonia	RSc	3098991	3099244	-	253	0.482213	Fasta	Img	GFF
Ralstonia	RSc	3277201	3277454	-	253	0.482213	Fasta	Img	GFF
IPO	IPO001	2845457	2845650	+	193	0.445596	Fasta	Img	GFF
Ralstonia	RSp	1204686	1204939	-	253	0.482213	Fasta	Img	GFF

(d)

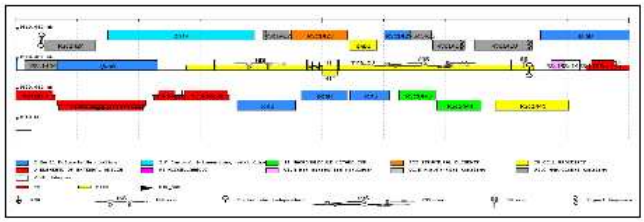
(b) Alignement multiple des séquences d'une composante connexe



(c) Assemblage des séquences d'une composante connexe



(d) Contexte génomique des éléments de la composante connexe



(e) Tableau de synthénie:les éléments adjacents aux éléments de la composante connexe RNAsim sont présentés

Organisme	Contig	Distance à gauche	El. à gauche	Num. accession	Brin	Description	Distance à droite	El. à droite	Brin	Description
MOLK	MK002	9880	-	MK00693	+	response regulator protein				
Ralstonia	RSc	1	RNA	-	+	Ala tRNA	1	RNA	+	23S ribosomal
Ralstonia	RSc	1	RNA	-	-	23S ribosomal	1	RNA	-	Ala tRNA
Ralstonia	RSc	1	RNA	-	-	23S ribosomal	1	RNA	-	Ile tRNA
IPO	IPO001									
Ralstonia	RSp	1	RNA	-	-	RNA-23S ribosomal	1	RNA	-	Ala tRNA

FIG. 7.7 – Présentation des sorties RNAsim : (a) présentation d'une composante connexe à 6 éléments avec les noms d'organismes, les positions, les longueurs des éléments, le brin , le G+C% des éléments, leurs séquence FASTA, la structure secondaire prédite par RNAfold et le fichier GFF correspondant ; (b) l'alignement multiple obtenu à partir des séquences des éléments dans la composante connexe ; (c) l'assemblage ; (d) le contexte génomique des éléments est disponible via iANT ; (e) le tableau de synthénie présentant les éléments génomiques adjacents aux éléments de la composante connexe.

pré-calculés et disponibles (*via* la page web) :

1. *Alignement des éléments des composantes connexes*
   L'alignement multiple des séquences présentes dans une même composante connexe à l'aide de MultAlin (Corpet (1988)) (fig. 7.7b, élargie sur fig.7.8) permet de mettre en évidence le degré de similarité entre séquences d'une même composante connexe. Cependant, l'alignement multiple ainsi généré est un alignement global qui aligne toutes les séquences sur toute leur longueur. Or, lorsque les segments chevauchants ont été intégrés dans la composante connexe au cours d'une des étapes de RNAsim, aligner de telles séquences par un outil d'alignement global n'a pas toujours de sens. C' est pourquoi nous avons également utilisé l'outil d'assemblage de séquences CAP3 (Huang et Madan (1999)) (fig. 7.7c élargie sur fig.7.8) qui est adapté à de telles situations, permettant d'aligner uniquement les séquences présentant localement de fortes similarités.
2. *Visualisation du contexte génomique*
   Lorsque nous analysons les résultats de RNAsim, la visualisation du contexte génomique, c'est-à-dire la position de la séquence candidate par rapport aux gènes qui l'entourent, peut être informative : certaines composantes connexes peuvent ainsi être caractérisées rapidement, comme par exemple les séquences ITS ou les extrémités de séquences d'insertion (pour définition voir section 8.6). De plus, cette fonctionnalité est utile lorsque nous nous intéressons à une fonction ou à une région particulière du génome. Pour faciliter cette tâche, le contexte génomique de chaque élément est disponible *via* iANT (Thebault *et al.* (2000)) (fig. 7.7d élargie dans fig. 7.9). La notion de synténie, introduite au moyen d'un tableau de gènes adjacents pour chaque séquence d'une même composante connexe (fig. 7.7e élargie dans fig. 7.9), nous a été utile lors de l'analyse des résultats (section 10).
3. *Calcul de la structure secondaire prédite*
   Une structure secondaire stable est caractéristique de nombreux ARNnc et d'autres éléments de régulation structurés. C'est donc un élément important dans la prédiction de ces éléments. Pour cette raison, la structure secondaire prédite à l'aide de RNAfold (Zuker (2003), Hofacker (2003)), est calculée pour tout élément de longueur inférieure à 200nt (fig. 7.10). La limite en longueur a été imposée parce que les calculs deviennent trop coûteux quand ce seuil est dépassé mais aussi et surtout parce que la qualité des prédictions diminue avec la longueur des sé-

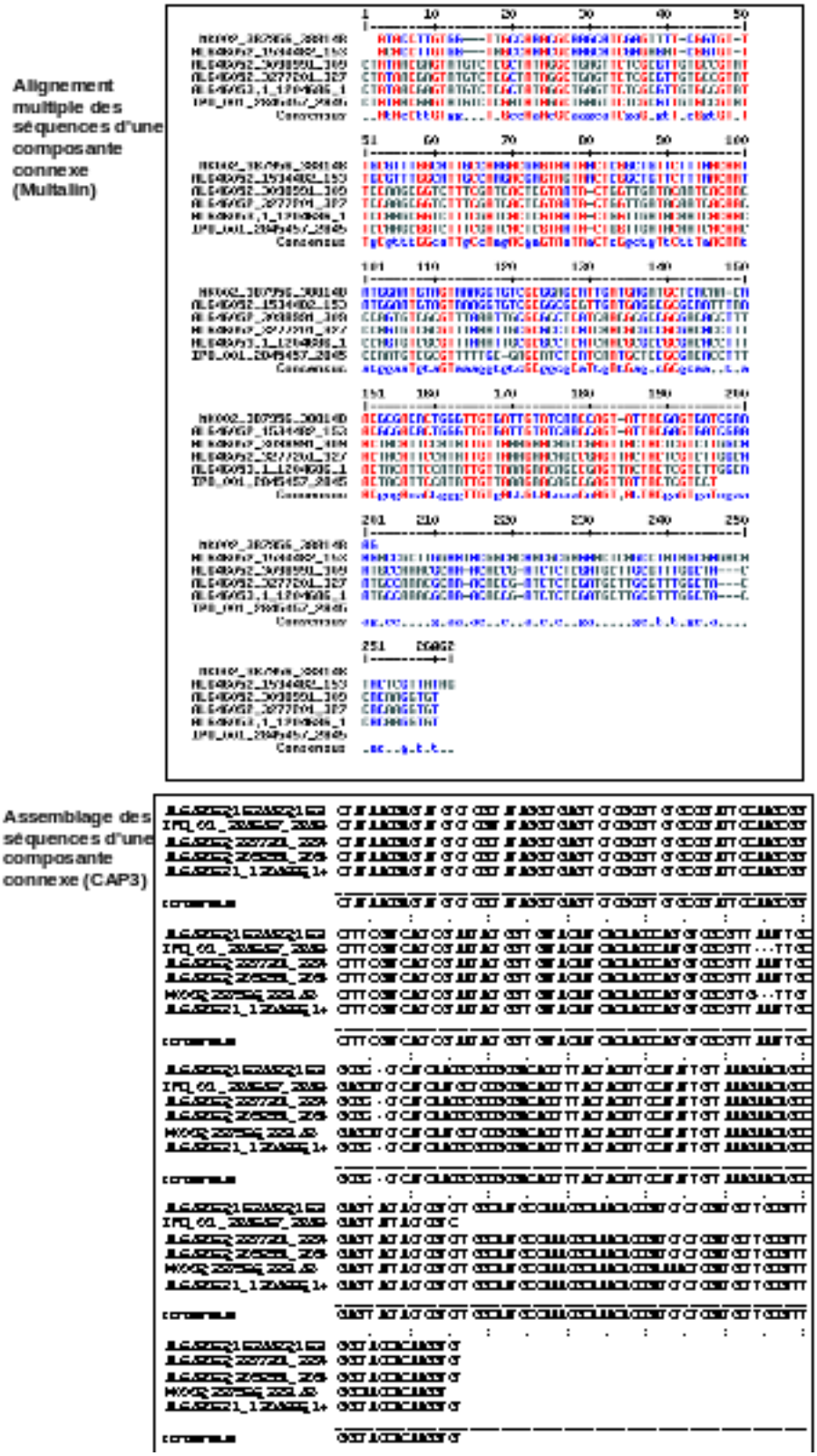


FIG. 7.8 – Les alignements multiples et l'assemblage des séquences de la composante connexe présentée sur la figure 7.7a.

quences (Doshi *et al.* (2004)). Il convient ici de rappeler que Rnafold n'utilise pas l'information comparative contenue dans plusieurs séquences. C'est pourquoi nous avons également utilisé CARNAC (Touzet et Perriquet (2004)), un outil rapide permettant de prédire la structure secondaire commune à un ensemble de séquences sur la base de la recherche des motifs en tige-boucle communs présentant des mutations compensatoires. Nous avons volontairement limité l'utilisation de CARNAC à des composantes connexes de moins de 6 éléments en raison de la difficulté de l'analyse de résultats dans les cas des composantes connexes de taille supérieure à 5.

Analysons à présent l'exemple présenté sur la figure 7.7. Les informations disponibles dans la sortie RNAsim seules montrent immédiatement que les séquences des éléments de la composante connexe sont très homologues (l'alignement multiple, figure 7.8), et pour la plupart presque identiques (pour 5 de 6 séquences l'assemblage est possible, figure 7.8). Ensuite, en inspectant le contexte et le tableau de synthénie (figure 7.9),

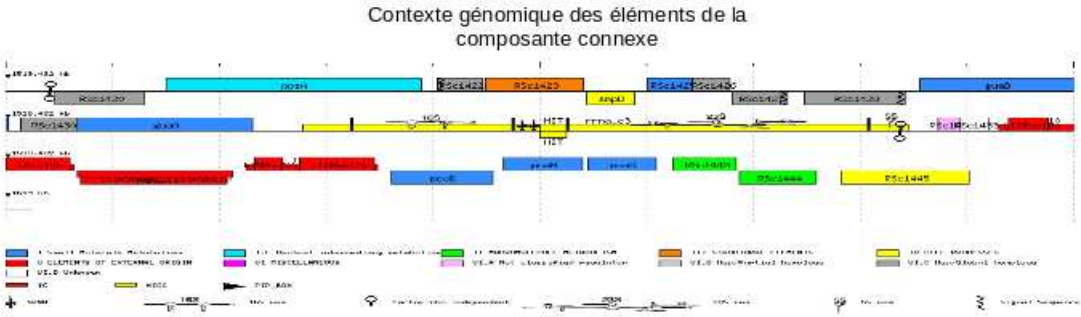


Tableau de synthénie:les éléments adjacents aux éléments de la composante connexe RNAsim sont présentés

Organisme	Contig	Distance à gauche	El. à gauche	Num. accession	Brin	Description	Distance à droite	El. à droite	Brin	Description
MOLK	MK002	9890	-	MK00693	+	response regulator protein				
Ralstonia	RSc	1	RNA	-	+	Ala tRNA	1	RNA	+	23S ribosomal
Ralstonia	RSc	1	RNA	-	-	23S ribosomal	1	RNA	-	Ala tRNA
Ralstonia	RSc	1	RNA	-	-	23S ribosomal	1	RNA	-	Ile tRNA
IPO	IPO001									
Ralstonia	RSp	1	RNA	-	-	RNA-23S ribosomal	1	RNA	-	Ala tRNA

FIG. 7.9 – La visualisation du contexte génomique de la composante connexe de la figure 7.7a.

nous pouvons conclure que les éléments sont quasiment systématiquement entourés par des ARNt et des ARNr. La structure secondaire prédite (figure 7.10) montre une structure secondaire présentant une longue tige-boucle. L'ensemble de ces éléments suggère qu'il s'agit de l'élément ITS (Goertzen *et al.* (2003)) conservé au sein des opérons ribosomiques.

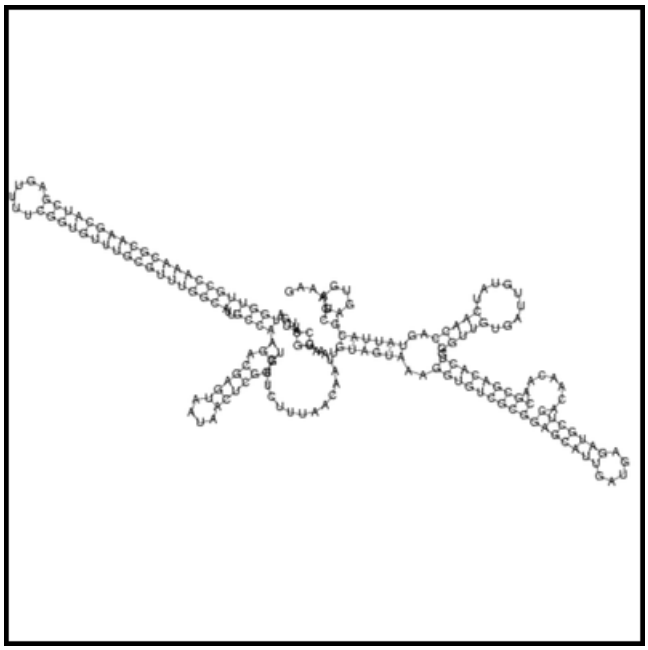


FIG. 7.10 – Prédiction de la structure secondaire par RNAfold du premier élément de la composante connexe de la figure 7.7a.

Chapitre 8

Application au génome de *R. solanacearum*

Nous présentons ici les résultats de l'utilisation de RNAsim sur le génomes de *R. solanacearum* souche GMI1000 avec un certain nombre d'autres génomes.

Dans cette partie, la synthèse des résultats sera présentée tandis que l'analyse détaillée d'une partie des résultats sera exposée dans la partie suivante du manuscript.

8.1 Organismes comparés

Notre objectif scientifique étant la prédiction des ARNnc chez *R. solanacearum* au moyen de RNAsim, nous avons constitué 4 jeux de données qui sont présentés dans le tableau 8.1.

Les ARNnc divergent rapidement en séquence au cours de l'évolution entre différents organismes. Leur prédiction, sur la base d'une conservation en séquences, est donc facilitée lorsque les séquences de plusieurs souches du même organisme sont disponibles. C'est pourquoi nous avons exploité, en plus de la séquence de la souche GMI1000 publiée (Guidot *et al.* (2007)), les séquences de deux nouvelles souches Molk2 et IPO1609 générées dans le groupe de Christian Boucher, en collaboration avec le Génoscope. Ces deux souches se différencient de la souche GMI1000 par leur spécificité d'hôte ainsi que par leur appartenance à un clade distinct de celui de la souche GMI1000.

Ensuite *R. solanacearum* a été comparé à *E. coli*. Chez cet organisme, on connaît 74 familles d'ARNnc, en dehors des ARNt et ARNr (source Rfam) et il est, à ce jour,

l'organisme le plus exploré d'un point de vue des ARNnc. Bien que phylogénétiquement éloignée de *R. solanacearum*, une comparaison entre les deux organismes pourrait permettre d'identifier d'éventuels ARNnc conservés.

Par ailleurs, compte tenu de l'intérêt de nos partenaires pour l'analyse du pouvoir pathogène, nous avons souhaité comparer *R. solanacearum* à un autre organisme phytopathogène *Xanthomonas campestris pv. campestris*. Bien que phylogénétiquement distant de *R. solanacearum*, cet organisme a été choisi car, comme chez *R. solanacearum*, son pouvoir pathogène est dépendant d'un système de sécrétion de type III et des effecteurs afférents dont certains sont très conservés entre les deux bactéries (Christian Boucher, communication personnelle). Par ailleurs, il a été montré que des régulateurs de transcription de l'expression de ces gènes sont conservés entre les deux organismes, laissant présager qu'une convergence pourrait également exister en termes d'ARN régulateurs qui serait impliqués dans le contrôle de la pathogénie.

Enfin, il était aussi important de comparer le génome de la souche GMI1000 de *R. solanacearum* à lui-même, dans la mesure où certains ARNnc peuvent être présents en plusieurs copies dans un génome (par exemple, les ARNnc PrrF1 et PrrF2 chez *P. aeruginosa* ont été découverts en tant que répétitions, Wilderman *et al.* (2004)). Ce procédé pourrait notamment permettre d'identifier les ARNnc souche-spécifiques, présents en plusieurs copies dans un génome.

Toutes les séquences publiques utilisées pour cette étude ont été obtenues à partir du NCBI
(ftp ://ftp.ncbi.nih.gov/genomes/Bacteria/ en Novembre 2007). Les séquences génomiques des souches IPO1609 et Molk2 de *R. solanacearum* ont été fournies par nos partenaires biologistes.

8.1.1 Constitution de l'ensemble des séquences utilisées

Dans tous les génomes utilisés, toutes les régions intergéniques de longueur inférieure à 80nt ont été éliminées, ainsi que les séquences des ARNt et ARNr. A titre d'exemple, dans le chromosome de *R. solanacearum*GMI1000, 1463 régions sur 3062 régions intergéniques sont de longueur inférieure à 80nt.

Nous avons inclus, dans les séquences à comparer, les ORF courtes (longueur inférieure à 300nt). La raison en est que l'annotation du génome de *R. solanacearum*

Jeux de données	Organismes	But recherché
1	• 3 souches de *R. solanacearum*	Recherche des ARNnc dans le génome de *R. solanacearum*
2	•*R. solanacearum* GMI1000 •*X. campestris pv. campestris str. 8004*	Recherche des ARNnc impliqués dans la pathogénie
3	•*R. solanacearum* GMI1000 •*E. coli* K12	Recherche des ARNnc communs à *E. coli* et *R. solanacearum*
4	•*R. solanacearum* GMI1000	Recherche des ARNnc, souche spécifiques, répétés dans le génome

TAB. 8.1 – Résumé des jeux données utilisés dans l'application de RNAsim sur les séquences génomiques.

contient un certain nombre d'ORF de petite taille pour lesquelles des gènes homologues ne sont pas décrits dans d'autres organismes et qui pourraient de ce fait avoir été prédites en tant que régions codantes abusivement. Elles sont au nombre de 494.

8.2 Paramètres initiaux

Pour la comparaison des régions intergéniques des différents organismes nous avons utilisé WU-Blast (version 2.2.6) avec le schéma de score par défaut. La longueur minimale du mot d'ancrage W a été fixée à 8 (sa taille par défaut étant de 11). Ce choix a été réalisé pour permettre aux séquences de moins de 11 nucléotides consécutifs identiques d'être prises en compte : en effet, les ARNnc conservés ne doivent pas obligatoirement contenir un si long enchaînement de nucléotides identiques. La diminution de la longueur de mot d'ancrage augmentant considérablement le temps de calcul, la valeur 8 a été choisie comme étant un compromis entre une vitesse réaliste d'exécution et la sensibilité de Blast. La E-value maximale a été fixée à 0.01.

Pour QRNA, les valeurs par défaut ont été utilisée.

La taille de la fenêtre glissante a été fixée à 150nt, valeur maximale. La valeur de

pas a été également fixée à 150.

Le paramètre de chevauchement dans le calcul des composantes connexes a été fixé à 1 : ainsi toutes les séquences se chevauchant sont mises dans la même composante connexe. Dans un premier temps, les séquences courtes ont été laissées dans le graphe.

L'application de RNAsim ainsi paramétré a donné lieu à des résultats difficiles à interpréter. Pour chacun des jeux de données utilisés une composante connexe « géante » était obtenue. Par exemple, dans le cas de la comparaison des trois souches de R. solanacearum, 8180 éléments sont obtenus avant l'étape de fusion. 5160 de ces éléments appartiennent à une seule composante connexe alors que les autres éléments se répartissent dans 979 composantes connexes restantes. Une situation similaire se produit dans le cas des autres jeux de données. Après l'étape de fusion le nombre d'éléments a été réduit à 4628 et la taille de la grande composante connexe à 2040 éléments.

Même si, d'un point de vue de l'étude comparative des génomes, ces résultats sont intéressants, ils restent néanmoins difficiles à analyser dans le cadre de la détection des ARNnc. Les questions de "comment identifier les liens existants entre ces 2040 éléments" et de "savoir si cette composante connexe avait un sens biologique" se sont posées. Cette dernière question est abordée plus amplement dans la section 8.7. On peu déjà dire que certains des alignements n'ont pas de pertinence biologique. Pour un nombre important des alignements à taux d'identité entre 50% et 70% il s'agissait des longs segments de G ou C alignés, résultant de la comparaison de deux séquences G+C riches et pour lesquelles la fréquence d'apparition aléatoire de telles structures est importante.

8.3 Restriction selon la taux d'identité des alignements

Nous avons effectué des tests en utilisant différents jeux de paramètres afin de contourner l'écueil de formation d'une grande composante connexe. Quelques pistes explorées sont : la contamination par les alignements courts, les paramètres de WU-Blast et notamment les pénalités de gap ainsi que le taux d'identité des alignements trop peu élevés pour éviter les alignements peu pertinents. Ce dernier critère a permis de réduire la taille des composantes connexes à une taille raisonnable et de rendre aux connexions entre les différentes séquences un sens biologique. Les autres pistes sont restées sans effet. Nous avons donc fixé à 70% le seuil inférieur de pourcentage d'identité en dessous duquel les alignements ont été considérés comme non-significatifs. Cette so-

lution a permis de diminuer considérablement le nombre d'arêtes à manipuler dans le calcul des composantes connexes. Avec la contrainte de 70% sur le pourcentage d'identité, la taille de la plus grande composante connexe est de 51 éléments. De plus, et contrairement à la composante « géante », cette composante est facilement expliquée par ses éléments (section 8.4).

8.4 Résultats

Nous présenterons d'abord une vue globale des résultats à travers l'effet des filtres du taux d'identité et de QRNA. Ensuite, nous présenterons une synthèse des résultats pour chacun des jeux de donnés définis.

8.4.1 Analyse de l'effet des filtres

Un premier filtre, préalable à l'analyse comparative, a consisté à enlever les régions intergéniques de longueur inférieure à 80nt et les régions correspondantes aux ARNt et ARNr, laissant 2332 séquences, pour les chromosome et mégaplasmide confondus de la souche GMI1000, sur lesquelles l'analyse comparative a été effectuée.

Le nombre de régions retenues dans chacune des phases préliminaires à la construction des composantes connexes est donné dans le tableau 8.2.

Un nombre important de régions est conservé entre la souche GMI1000 et les deux autres souches. Ceci est attendu, étant donné qu'il s'agit des génomes des différentes souches de la même bactérie, donc des génomes présentant une même organisation génétique. En revanche, le nombre de régions conservées est moins important dans le cas de la comparaison avec *X. campestris*. Dans le cas de la comparaison avec le génome de *E. coli* la conservation est rare. Ces résultats sont cohérents avec la distribution phylogénique des organismes utilisés, les trois souches de la béta-proteobactérie *R. solanacearum* étant très proches et plus éloignées des génomes des gamma-protéobactéries *X. campestris* et *E. coli*. Néanmoins, nous aurions pu nous attendre à ce que le génome de *X. campestris* soit plus proche de *R. solanacearum* que du génome de *E. coli* : *X. campestris*, tout comme *R. solanacearum* est une bactérie phytopathogène. De plus, le génome de *X. campestri*, comme celui de *R. solanacearum* est riche en G+C. Notons que le filtre de 70% d'identité accentue davantage cette différence : le nombre

GMI1000	Organismes comparés	Blast(Eval=0.01)	ID%>70	QRNA
2332	IPO1609	1640	1406	812
	Molk2	1840	1501	865
	GMI1000	580	209	91
	X.c. 8004	568	38	4
	E.c. K12	54	9	5

TAB. 8.2 – Le filtre effectué sur les IGR dans les étapes préalables à la construction du graphe, pour les différents génomes comparés avec le génome de *R. solanacearum* GMI1000 : IPO1609, souche Molk2, *X. campestris pv. campestris str. 8004 (XC 8004)* et *E. coli K12 (EC K12)*. Colonne Blast : le nombre d'IGR de *R. solanacearum* GMI100 présentant des homologies WU-Blast avec les IGR des autres génomes utilisés pour comparaison. Colonne Blast ID>70% : le nombre d'IGR présentant des homologies à pourcentage d'identité supérieur à 70%. Colonne QRNA : le nombre d'IGR présentant des homologies étiquetées comme ARN par QRNA. Dans le cas de la comparaison GMI1000-GMI1000 uniquement les régions non-identiques ont été comptées.

d'alignements retenus ne varie pas dramatiquement dans le cas des trois souches de *R. solanacearum* tandis que dans le cas de *X. campestris* et d'*E. coli* le nombre de régions retenues diminue pour atteindre quelques unités.

L'étape QRNA filtre de façon importante les régions intergéniques retenues (dans le cas des 3 souches de *R. solanacearum* le nombre de régions est réduit d'un facteur 2). De ces résultats nous pouvons conclure que QRNA est un filtre utile pour réduire le nombre de régions intergéniques à considérer. Néanmoins, la pertinence de ce filtre reste à établir (voir section 8.7).

8.4.2 Résultats de comparaison de trois souches de *R. solanacearum*

4628 éléments, tous génomes confondus sont issus de la comparaison des trois souches de *R. solanacearum*. Ces éléments sont repartis en 1708 composantes connexes. La classe la plus peuplée, contenant 876 composantes, correspond aux composantes contenant deux éléments et la plus grande des composantes contient 52 éléments.

Le grand nombre de composantes connexes à 2 éléments indique que, dans un nombre

important de cas, la correspondance existe seulement dans deux des trois génomes considérés (fig. 8.1). Les composantes connexes à 2 éléments correspondent essentiellement aux régions conservées entre les souches IPO1609 et Molk2 et absentes de la souche GMI1000, cette dernière étant plus éloignée phylogénétiquement (Fegan et Prior (2005) et Christian Boucher, communication personnelle). Dans la suite de l'analyse, toutes les composantes connexes ayant des éléments dans seulement deux des trois génomes comparés n'ont pas été expertisées. Leur conservation dans seulement deux des trois souches est un argument en défaveur d'une conservation fonctionelle.

Un grand nombre de composantes connexes ont été analysées en détail et les résultats les plus intéressants issus de cette analyse seront présentés dans la section 10. Toutefois, à l'aide des tableaux de synténie et à l'examen du contexte génomique, la nature de la majorité des composantes connexes de grande taille a pu être identifiée immédiatement. Les composantes connexes à 19 et 51 éléments correspondent à des structures palindromiques présentes aux extrémités de séquences d'insertion. La composante à 29 éléments correspond à un terminateur de transcription. Le cas de la composante connexe à 41 éléments est différent : ses éléments correspondent à une longue répétition quasi-palindromique que nous avons identifié dans le génome de *R. solanacearum*. Son

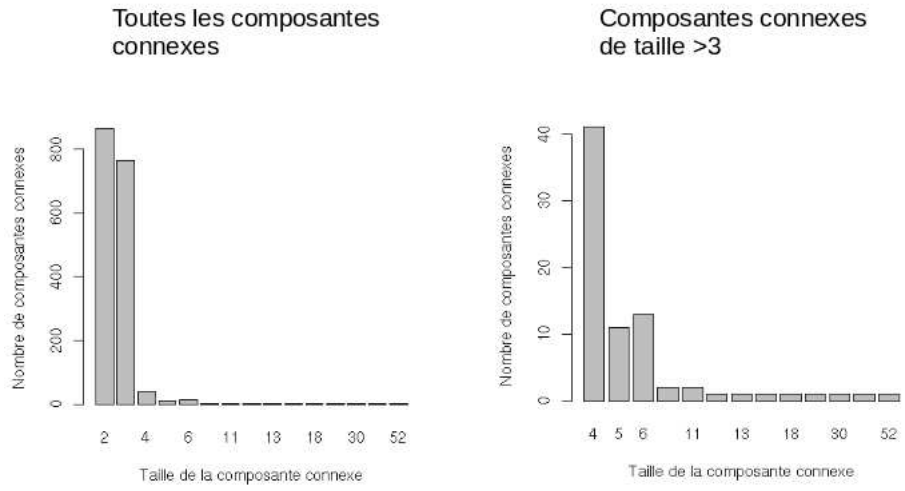


FIG. 8.1 – La répartition du nombre de composantes connexes en fonction de leur taille. A gauche, la totalité des composantes connexes, à droite les composantes connexes de taille supérieure à 3.

analyse ne sera pas présentée dans le cadre de la thèse. Dans une première analyse des résultats de RNAsim ayant pour but la recherche des ARNnc toutes ces composantes peuvent être éliminées comme correspondant à des éléments structurés autres que ceux recherchés. Au terme de cette première analyse, il reste dans le génome de *R. solanacearum* 1105 candidats pour des ARNnc à expertiser individuellement.

8.4.3 Résultats de comparaison GMI1000-GMI1000

La comparaison intra-génomique a donné lieu à 39 composantes connexes. 28 correspondent à des composantes à deux éléments. La plus grande composante connexe contient 11 éléments. La majorité de ces composantes se situe au sein d'une région de 31 kb dupliquée en tandem sur le mégaplasmide de *R. solanacearum* GMI1000 (Salanoubat et al. 2002b) et ont donc peu de chance de correspondre à des ARNnc. La majorité des 11 composantes connexes restantes d'ordre supérieur correspond aux séquences d'insertion et à des séquences présentes aux sein des opérons codant pour les protéines ribosomiques. La composante connexe à 8 éléments correspond à la répétition quasi-palindromique présente dans la comparaison entre trois souches de *R. solanacearum*. Bien que nous n'ayons pas eu l'occasion d'analyser en détail l'ensemble de ces résultats, nous avons pu identifier au sein de ces candidats la présence d'un élément de régulation potentiel qui sera présenté en détail dans le chapitre Candidats (les élements en amont des gènes *popF1* et *popF2*, section 10.3.1).

8.5 Autres résultats

Les comparaisons entre *R. solanacearum* GMI1000 et *X. campestris* ainsi qu'avec *E. coli* génèrent un faible nombre d'éléments (voir tableau 8.2). L'analyse des résultats de comparaison avec *X. campestris* montre que la plupart des régions conservées correspondent aux ORF courtes conservées entre deux organismes. Cette analyse suggère que d'éventuels ARNnc communs entre ces deux organismes ne sont probablement pas conservés en séquence.

Quand à la comparaison avec *E.coli*, un élément conservé est adjacent, dans les deux organismes, aux séquences d'insertion qui sont vraisemblablement communes à ces deux génomes (séquence *insC* de *E. coli* et séquence IsRso10a de *R. solanacearum*).

Deux autres éléments correspondent, dans les deux génomes, à des séquences dans les régions 3'UTR des gènes où des terminateurs rho-indépendants ont été prédits suggérant qu'il s'agit, en fait, des terminateurs conservés entre les deux organismes. Néanmoins, la possibilité qu'il s'agisse des ARNnc n'est pas définitivement exclue.

8.6 Eléments de validation de RNAsim

L'analyse préliminaire des résultats a permis d'apporter quelques éléments de validation de RNAsim. En effet, nous avons pu retrouver un certain nombre d'ARNnc connus et aussi plusieurs éléments génomiques connus pour être structurés et conservés dans les différents organismes. Nous les présentons ici.

8.6.1 Les ARNnc connus retrouvés

9 ARNnc *R. solanacearum* autres que les ARNr, ARNt, et tmRNA sont annotés à ce jour (ref partie Introduction). Dans notre analyse, nous avons retrouvé 8 de ces 9 éléments. Le candidat manquant correspond à un des 2 exemplaires de riboswitch *yybP-ykoY*, le second exemplaire ayant été éliminé dans l'étape QRNA. Ces résultats seront discutés plus amplement dans la section 8.7.

8.6.2 D'autres éléments structurés retrouvés

RNAsim a permis de retrouver plusieurs éléments génomiques conservés, appartenant aux régions intergéniques et connus pour être structurés. On peut citer les éléments ITS et les séquences d'insertion.

Les éléments ITS

Les éléments ITS (Internal Transcribed Spacer) sont des éléments génomiques se trouvant dans l'espace inter-ARN dans les opérons ribosomiques. Ils ont un rôle important dans le positionnement des sous-unités du ribosome et dirigent leur propre excision du transcript. Les éléments ITS sont hautement variables relativement aux séquences d'ARNr mais leur structure secondaire est souvent conservée même entre des organismes phylogénétiquement éloignés (Goertzen et al. 2003). Le chromosome de *R. solanacearum* GMI1000 contient 3 opérons ribosomiques et le mégaplasmide en contient un. A la

suite de l'analyse par QRNA, l'ensemble des éléments ITS ont été regroupés dans une même composante connexe.

Les bordures de séquences d'insertion IS

Plusieurs composantes connexes correspondent au regroupement d'extrémités de séquences d'insertion (IS). Or pour la majorité des IS, il a été montré que la séquence codante de la transposase est flanquée par des courtes répétitions inversées (les séquences IR) allant de 10 à 40nt et susceptibles de former des structures en tige-boucle ou des structures cruciformes (Cozzuto *et al.* (2008)). Ce sont bien ces régions conservées et structurées qui ont été identifiées au cours de notre analyse.

8.7 RNAsim : discussion et perspectives

RNAsim est un outil développé dans le but de recherche des ARN non-codant à l'aide de l'approche comparative utilisant plusieurs génomes. Il regroupe et synthétise les résultats de comparaisons deux-à-deux de plusieurs génomes de façon cohérente et applique, sur ces résultats un ensemble d'outils facilitant leurs analyse.

Néanmoins, RNAsim peut être vu comme un outil d'exploration des génomes d'un point de vue de la conservation des régions intergéniques. En effet, des éléments conservés structurés autres que les ARNnc ont pu être mis en évidence à l'aide de RNAsim (voir section 8.6).

Les résultats obtenus dans la comparaison entre 3 souches de *R. solanacearum* sont encore trop nombreux pour être repris tels quels et nécessitent d'être expertisés individuellement avant d'enterprendre une validation par des méthodes de biologie. Nous présentons ce travail dans la partie suivante de cette thèse (chapitre 10).

RNAsim

Dans le travail présenté le graphe d'alignements a été construit de façon directe : un alignement donnait une arête entre les sommets correspondants aux régions impliquées. Notre but ayant été de développer un outil rapidement opérationnel donnant des résultats cohérents afin de générer les candidats d'ARNnc dans le génome de *R. solanacearum*, nous ne nous sommes pas attardés à réaliser une modélisation plus sophistiquée du réseau d'alignements. Par conséquent, toutes les composantes connexes sorties de RNAsim ont la même pertinence. Or, les outils sous-jacents, Blast et QRNA possèdent chacun un système d'évaluation de pertinence des résultats générés. D'un point de vue

de la modélisation du réseau d'alignements, les perspectives de RNAsim pourrait s'inscrire dans l'introduction d'un score de pertinence des alignements considérés, par le biais d'attribution d'un poids aux arêtes. L'introduction d'un tel score permettrait une classification des prédictions selon leur pertinence. L'introduction de la notion de score permettrait également de définir un filtre sélectionnant les composantes connexes se trouvant dans un intervalle spécifique, adaptée à la distance phylogénétique des génomes comparés. Un tel filtre a été, en quelque sorte, simulé lorsque les alignements ont été restreints à ceux dépassant 70%. Néanmoins, ce filtre qui semblait adapté à la comparaison des trois souches de *R. solanacearum* a donné peu d'éléments de conservation dans le cas des organismes plus éloignés comme *X. campestris* ou *E. coli*. Néanmoins, la difficulté à combiner les scores lors de la fusion d'alignements reste à étudier.

Du point de vue de l'implémentation des fonctionalités manquantes de RNAsim, nous avons pu constater le besoin d'accéder aux différents alignements associés à une composante connexe. Actuellement, leur recherche doit être faite manuellement.

Du point de vu technique, RNAsim peut être considérablement amélioré notamment en rendant automatique l'exécution des différentes étapes et la gestion des options et des filtres.

Enfin, lors de l'analyse des résultats de RNAsim nous avons pu constater que le filtre QRNA n'a pas été toujours adapté : certains alignements ont été rejetés en raison de leur trop grande homologie. Ceci est justifié du point de vue de l'approche QRNA car les mutations compensatoires n'ont pas pu être identifiées dans de telles séquences. Néanmoins, le rejet des tels alignements situés à l'intérieur des régions intergéniques ne nous semble pas justifié d'un point de vue de la recherche des ARNnc dans des organismes différents. En effet, de telles conservations au sein des régions intergéniques longues pourraient mettre en évidence l'importance de ces régions pour la biologie de l'organisme et pourraient indiquer la présence d'ARNnc. L'absence de mutations compensatoires peut être lié à l'absence de la structuration dans les ARNnc potentiels ainsi qu'à l'absence de divergence suffisante dans les génomes très proches (par exemple, les différentes souches d'un même organisme). Un deuxième inconvénient de QRNA, lié à son utilisation dans les génomes G+C% riches est que l'estimation de ses paramètres par défaut a été effectuée sur des génomes simulés, avec un G+C% autour de 50%. De ce fait, le nombre de prédictions dans les génomes G+C% riches est probablement surestimé (Rivas et Eddy (2001)). Au regard cette analyse, une recherche basée uniquement sur la

similarité en séquences est à envisager. Le filtre de QRNA pourrait, par exemple, être remplacé par une recherche en séquence plus stringeante générant moins de candidats.

Analyse des résultats

Lors des analyses intermédiaires des résultats de RNAsim nous avons observé la formation systématique d'une composante connexe "géante". La possibilité qu'il s'agisse d'un phénomène spécifique au génome de *R. solanacearum* ou des génomes G+C% riches a été écartée après que le même phénomène ait été observé lors de l'application de RNAsim sur le génome de *Staphylococcus aureus*, un génome A+T% riche.

Ce problème a été résolu en fixant un seuil de taux d'identité plus élevé pour les alignements retenus, la question des raisons de sa formation au sein du réseau d'alignements persiste. Nous pouvons nous poser la question de la structure de la composante connexe géante : s'agit-il d'un sous-graphe au sein duquel tous les sommets sont connectés entre eux ou plutôt d'un sous-graphe composé de "clusters" fortement connectés avec les arêtes connectants ces "'clusters" entre eux ? Pour explorer la structure de la composante connexe, des outils d'analyse de graphes complexes, tels que HMETIS (Caldwell *et al.* (2000)) pourraient être utilisés.

Dans les discussions sur cette question, un autre point de vue nous a été suggéré avec l'observation que la formation des composantes connexes géantes était une propriété des graphes aléatoires. Les graphes aléatoires sont les graphes dont les arêtes entre les sommets du graphe sont attribuées de façon aléatoire, selon une loi de probabilité. Une hypothèse à explorer, afin d'expliquer le phénomène de la formation des composantes connexes géantes au sein des graphes d'alignements, serait le lien entre les graphes d'alignements et les graphes aléatoires (pour la définition d'un graphe aléatoire voir, par exemple, Reka et Albert-Laszlo (2002) et pour les conditions de formation de la composante connexe "géante" voir, par exemple Newman (2003)).

Après l'analyse des candidats générés par RNAsim dans *R. solanacearum* nous pouvons nous poser la question de savoir si notre choix des jeux de données a été judicieux. En effet, même si les trois souches séquencées de *R. solanacearum* présentaient une occasion privilégiée pour détecter des ARNnc dans cet organisme, du fait de leur proximité phylogénétique, un grand nombre de candidats a été généré. D'un autre côté, les recherches complémentaires dans les organismes plus éloignés n'ont pas donné lieu à assez de conservations pour que les éventuels ARNnc communs soient détectés. Dans l'analyse détaillée des candidats nous avons pu constater une conservation presque sys-

tématique des candidats prometteurs trouvés chez *R. solanacearum* dans le génome de *Ralstonia pickettii*, un autre génome du groupe des *Ralstonia*, séquencé depuis peu. Par conséquent, l'utilisation de ce génome, ensemble avec les trois souches de *R. solanacearum* aurait probablement permis une réduction importante du nombre de candidats à analyser.

Quatrième partie

L'analyse des résultats de RNAsim pour identification des ARNnc candidats dans le génome de *Ralstonia solanacearum*

Introduction

Dans cette partie nous présentons les candidats les plus prommeteurs issus de l'analyse des résultats de comparaison des tois souches de *R. solanacearum* à l'aide de RNAsim. Sur ces candidats nous proposons d'entreprendre des expériences visant la validation par des méthodes biologiques.

Les candidats les plus prometteurs ont été sélectionnés d'après un ensemble de critères défini à partir des travaux sur la recherche systématique des ARNnc. Un état de l'art, prenant en compte les travaux publiés jusqu'en 2008, est présenté.

Ensuite nous présentons un exemple répresentatif de l'analyse d'un candidat prometteur, les autres candidats retenus étant présentés dans une liste commune indiquant les principaux éléments prédictifs qui ont été utilisés pour leur séléction. Au cours de l'analyse, certaines prédictions RNAsim ont été écartées en tant que les ARNnc mais les régions ainsi identifiées présentent des caractéristiques qui traduisent très probablement la conservation des fonctions biologiques. Certains de ces éléments sont présentés dans la suite des candidats ARNnc sélectionnés pour valisation.

Chapitre 9

Recherche d'ARNnc bactériens à l'échelle génomique, à l'aide des approches bioinformatiques

Hormis les ARNr et les ARNt, 12 familles d'ARNnc étaient connues jusqu'en 2001 dans le génome de *Escherichia coli* (Chen *et al.* (2002)). Ces ARNnc, ensemble avec l'ARN III de *Staphylococcus aureus* et une multitude des ARNnc chez les eucaryotes, représentaient l'ensemble d'ARNnc connus à l'époque (Wassarman *et al.* (1999)). Ils ont été, pour la plupart, découverts par hasard : soit du fait de leur abondance dans le milieu cellulaire (les ARN SRP, tmRNA, 6S RNA, RnaseP RNA et Spot42) soit lors d'études des protéines (OxyS, CsrB, GcvB) ou bien par surexpression de certains fragments du génome (MicF, DicF, DsrA, RprA) (Wassarman *et al.* (2001)). L'expansion du domaine de recherche des ARNnc couplée aux séquençage d'un nombre croissant de génomes (le séquençage du génome de *E. coli* a été terminée en 1997), a engendré l'apparition d'études portant spécifiquement sur la recherche des ARNnc dans les génomes. Étant données la quantité et la complexité des données génomiques désormais disponibles, certaines de ces études se sont appuyées sur des méthodes bioinformatiques, rapides et de faible coût afin d'orienter les approches expérimentales.

Les études consacrées à la recherche systématique des ARNnc dans un génome donné ont d'abord été effectuées chez *E. coli* (pour une revue des premiers travaux voir Hershberg *et al.* (2003)). Les études portant sur d'autres organismes bactériens ont alors suivi. A ce jour, 23 génomes différents ont été examinés, de façon systématique, afin d'y

identifier des ARNnc (Livny *et al.* (2008)). Un grand nombre de ces études comportait

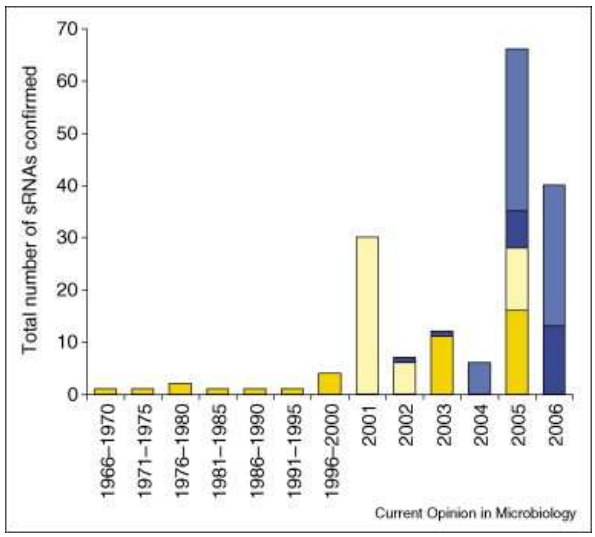


FIG. 9.1 – Le nombre d'ARNnc identifiés tous les ans depuis 1966. Jaune foncé : les ARN chez *E. coli* identifiés par les approches expérimentales. Jaune clair : les ARNnc chez *E. coli* identifiés par les approches bioinformatiquse. Bleu foncé, bleu clair même chose, chez les autres bactéries. Repris de Livny et Waldor (2007).

une approche bioinformatique et que, chez *E. coli* ou d'autres organismes, la majorité de nouvelles familles d'ARNnc découvertes l'ont été à l'aide de ces approches (figure 9.1).

Ces études suivent, en général, le schéma selon lequel une stratégie bioinformatique est définie pour la sélection des régions dans le génome contenant des ARNnc potentiels, appelés ARNnc candidats. Les candidats obtenus ainsi sont alors soumis à une validation biologique consistant à vérifier expérimentalement la transcription des régions désignées. Cependant, la validation d'un candidat ARNnc ne permet pas de répondre aux questions essentielles pour la caractérisation de ces nouveaux ARNnc :

1. Dans quel processus biologique les ARN identifiés sont-ils impliqués ?
2. Quelles est leur structure secondaire ?
3. Quel est leur mode d'action ?

Ainsi, pour un nombre important d'ARNnc connus aujourd'hui aucun rôle n'a été attribué et pour ce faire les études supplémentaires doivent être effectuées. L'exemple de RyhB, un ARNnc découvert lors d'une recherche systématique dans le génome de *E. coli*

montre le cheminement de la découverte à la caractérisation d'un ARNnc (Wassarman *et al.* (2001) ; Massé et Gottesman (2002) ; Masse *et al.* (2005) ; Jacques *et al.* (2006) ; Pré *et al.* (2007)).

Dans ce chapitre nous présentons une revue des études recherchant systématiquement des ARNnc dans des génomes bactériens, d'un point de vue des stratégies bioinformatiques de sélection des candidats. Nous présenterons d'abord celles portant sur le génome d'*E.coli* et ensuite celles portant sur d'autres génomes bactériens. La synthèse de ces stratégies nous permettra de définir notre propre stratégie de choix de candidats dans le cas de la recherche des ARNnc de *R. solanacearum*.

9.0.1 Recherches systématiques dans le génome d'*E. coli*

Les premiers travaux de recherche systématique d'ARNnc dans le génome d'*E. coli* sont parus entre 2001 et 2002 (Wassarman *et al.* (2001), Argaman *et al.* (2001), Rivas *et al.* (2001), Chen *et al.* (2002), Tjaden *et al.* (2002)) et parmi eux, un seul (Tjaden *et al.* (2002)) ne comporte pas d'étape bioinformatique utilisée dans la sélection des ARNnc candidats.

Par la suite, d'autres recherches des ARNnc chez *E. coli* ont été menées, en se basant exclusivement sur des approches biologiques, telles que l'utilisation de puces à ADN façonnées pour les régions intergénique ou le clonage aléatoire des petits ARN (approche dite *RNomics*) (Zhang *et al.* (2003), Vogel *et al.* (2003), Vogel et Sharma (2005)).

Parmi les études comportant une approche bioinformatique, Wassarman *et al.* (2001), Argaman *et al.* (2001) et Chen *et al.* (2002) peuvent être qualifiées d'essentiellement biologiques. Elles s'appuient sur des prédictions produites à l'aide d'outils bioinformatiques. Ces études ont instauré un standard dans le choix des stratégies bioinformatiques de recherche systématique des ARNnc.

Argaman *et al.* (2001) et Wassarman *et al.* (2001) font le constat que les ARNnc connus se trouvent majoritairement dans les IGR "vides" (comportant aucun gène annoté sur aucun des deux brins) et qu'ils sont conservés, en séquence, dans les génomes proches. En effet, dans la plupart des cas, la conservation des ARN entre le génome de *E. coli* et le génome de *Salmonella* était supérieure à 85% tandis que la conservation en séquence dans les régions codant pour des protéines homologues est souvent inférieure à

70% Argaman *et al.* (2001)). Par conséquent, les deux études définissent des stratégies globalement similaires consistant à rechercher les zones de conservation dans les régions intergéniques "vides".

Dans le cas de Argaman *et al.* (2001), la stratégie suivante a été appliquée :

1. identification des régions intergéniques « vides »
2. recherche des séquences promotrices du facteur sigma 70 et des terminateurs rho-indépendants dans ces régions
3. sélection des régions intergéniques « vides » contenant un promoteur et un terminateur prédits espacés de 50 à 400 paires de bases
4. comparaison des séquences issues du point 3 aux séquences génomiques des autres génomes séquencés et sélection de celles présentant une homologie significative (E-value inférieure à 0.001) avec d'autres organismes. Une similarité a été jugée pertinente si (1) la conservation couvrait plus de 70% de la région prédite, (2) l'homologie couvrait de 50% à 70% de la région prédite et le promoteur *ou* le terminateur était prédit avec un score élevé (3) l'homologie couvrait 30% à 50% de la région prédite et le promoteur *et* le terminateur étaient prédits avec un score élevé.
5. exclusion des régions appartenant aux régions 5' et 3' UTR des gènes voisins

Cette stratégie est assez restrictive en raison de l'exigence de la présence d'un promoteur σ^{70} et d'un terminateur rho-indépendant, et exclut donc tous les ARNnc qui seraient transcrits à l'aide d'un autre facteur que σ^{70} ou dont la terminaison de transcription ne se ferait pas par le mécanisme rho-indépendant. En conséquence, un nombre relativement petit de candidats a été retenu (24) dont 14 ont été validés, indiquant une bonne spécificité de la stratégie choisie.

Soulignons que, concernant la conservation de séquences dans les régions identifiées deux schéma d'organisation ont été observés : i) conservation de l'ensemble de la région intergénique et des gènes adjacents ainsi que de leur région UTR (synthénie) ; ii) conservation limitée à la zone codante de l'ARNnc indépendamment du contexte génomique. La majorité des ARNnc validés lors de cette étude (8 sur 10 des ARN connus et 10 sur 14 ARN validés) présentent le deuxième schéma de conservation.

Wassarman *et al.* (2001) appliquent une stratégie similaire, les principales différences étant la limitation aux IGR de longueur supérieure à 180nt, la prise en compte du

degré de conservation en plus du score d'alignement Blast et le fait que la présence des promoteurs et des terminateurs de transcription prédits n'étaient pas obligatoire pour retenir un candidat. Cette stratégie, moins restrictive, a généré 60 candidats dont 17 ont été validés comme des ARNnc.

Contrairement aux études évoquées ci-dessus, Chen *et al.* (2002) ne s'appuient pas sur une approche comparative et basent leur choix de candidats sur la prédiction des promoteurs et des terminateurs de transcription (espacés de 45 à 350nt). De ce fait, cette stratégie n'est pas dépendante de la comparaison avec d'autres génomes et peut donner des candidats génome-spécifiques. Elle présente néanmoins l'inconvénient majeur d'être difficilement généralisable du fait des faibles connaissances concernant la structure des séquences promotrices dans les organismes autres que *E. coli*. Dans cette étude, 8 ARNnc candidats ont été testés et 7 ont été validés (le critère sur lequel ces candidats ont été choisis n'a pas été donné).

A la différence des études citées précédemment, essentiellement biologiques, Rivas *et al.* (2001) présentent un modèle formel pour la prédiction des ARNnc implementé dans le logiciel QRNA. La méthode proposée est décrite plus amplement dans la section 6 et ici nous présenterons les résultats dont certains ont été validés expérimentalement.

Rivas *et al.* (2001) ont utilisé quatre génomes proches pour effectuer une recherche de régions intergéniques conservées en séquence et en structure à l'aide de QRNA. 556 candidats ont été générés dont 49 ont été testés expérimentalement donnant 11 candidats ARNnc validés. Etant donnée que l'étude expérimentale est restée limitée (une seule condition de culture ayant été testée), d'autres candidats issus de cette étude pourraient se révéler transcrits sous d'autres conditions de culture.

D'une façon générale, la recherche de la conservation entre les IGR des organismes phylogénétiquement proches a été le critère le plus utilisé dans les études de recherche systématique des ARNnc chez *E. coli* (tableau 9.1). L'ensemble de ces études a permis de détecter et de valider 45 nouveaux ARN non-codant chez *E. coli* (Hershberg *et al.* (2003)).

Bien que les approches utilisées dans Wassarman *et al.* (2001), Argaman *et al.* (2001) et Rivas *et al.* (2001) soient similaires, le nombre de candidats communs à plusieurs approches est faible : 1 ARNnc trouvé dans Wassarman *et al.* (2001) est retrouvé dans une autre étude, 4 dans le cas Argaman *et al.* (2001) et 9 dans le cas de Rivas *et al.* (2001). Ceci suggère que cette méthode est particulièrement sensible aux paramètres

Étude	Critères utilisés	Nb. candidats	Nb. cand. testés	Nb. cand. validés	Nb. d'ARN connus retrouvés
Argaman *et al.* (2001)	– les régions intergéniques « vides » – recherche des promoteurs et des terminateurs espacés de 50 à 400 nucléotides – recherche des séquences homologues dans d'autres organismes avec un critère sur le pourcentage de recouvrement des régions homologues	24	24	14	4/10
Wassarman *et al.* (2001)	– les régions intergéniques de longueur supérieure à 180nt – classification de ces régions intergéniques en fonction de leur conservation avec 2 organismes proches (les régions immédiatement en amont ou en aval des gènes étant rejetées) – intégration d'autres informations disponibles en rapport avec les régions choisies (présence d'éléments de régulation, promoteurs, terminateurs etc.)	60	60	17	x
Chen *et al.* (2002)	– les régions intergéniques « vides » – présence des promoteurs σ70 dépendants à 45nt-350nt d'un terminateur rho-indépendant prédit	227	8	7	33
Rivas *et al.* (2001)	– Les régions intergéniques « vides » de longueur supérieure à 50nt – Conservation en séquence (filtre BLAST) et en structure(QRNA)	556	49	11	4/4

TAB. 9.1 – Tableau récapitulatif des critères utilisés dans les recherches systématiques des ARNnc dans le génome de *E. coli*. Pour chaque étude le nombre de candidats générés, testés et validés est donné ainsi que le nombre d'ARNnc connus qui ont été retrouvés (vrais positifs). Dans le cas de Wassarman *et al.* (2001) les ARNnc qui auraient pu servir de contrôle ont été retirés. Dans le cas de Chen *et al.* (2002) les ARNt ont été pris en compte pour évaluer le nombre de vrais positifs. Dans le cas de Rivas *et al.* (2001) seulement 4 ARNnc connus ont servi de contrôle.

utilisés pour cette recherche, tels que les paramètres de BLAST, la longueur minimale des régions intergéniques, le degré de conservation etc.

Sur 45 ARNnc validés dans les études présentées, 5 ont été étudiés et une fonction leur a été associée (l'implication dans la régulation du métabolisme du fer pour RyhB Massé et Gottesman (2002), régulation négative des protéines de membrane externe pour sraD (Johansen *et al.* (2006)), implication dans le stockage du carbone pour csrC (Weilbacher *et al.* (2003)). Les autres restent à caractériser. En plus des ARNnc aujourd hui validés, les travaux de recherche décrits ont généré environ 1000 candidats non-redondants, dont la validité reste à tester (Hershberg *et al.* (2003)).

9.0.2 Recherches systématiques dans d'autres organismes bactériens

A ce jour, 23 génomes différents ont été étudiés de façon systématique afin de prédire des ARN non-codants (Livny *et al.* (2008)). Certaines de ces études ont été conduites par des approches purement expérimentales, notamment les études sur les génomes d'*Aquifex aeolicus* (Willkomm *et al.* (2005)), *Bacillus subtilis* (Silvaggi *et al.* (2006b)) etc. Néanmoins, la majorité de ces approches s'est appuyée sur des recherches bioinformatiques préalables (figure 9.1).

Plusieurs travaux utilisant des outils bioinformatiques ont été menés afin d'identifier une famille d'ARNs particuliers dans un génome donné (nous pouvons citer la recherche des homologues des ARN CsrB et CsrC régulant la protéine CsrA de *E. coli* dans *Vibrio fischeri* (Kulkarni *et al.* (2006)) ou la prédiction de l'homologue de l'ARNnc RyhB de *E. coli* dans le génome de *P. aeruginosa* en s'appuyant sur la prédiction des boîtes Fur (Wilderman *et al.* (2004))). Ici nous nous intéressons plus spécifiquement aux travaux dont l'objectif est de rechercher les ARNnc de façon systématique dans le génome entier d'un organisme à l'aide d'une approche bioinformatique (tableau 9.2).

Le point commun entre ces études est l'utilisation d'une approche comparative effectuée entre les génomes proches.

Organisme	Groupe	Gram	G+C%	Approche	Nb. candidats	Nb. candidats testés	Nb. cand. validés	Référence
Staphylococcus aureus	Firmicutes	+	33	Comparative/Composition	191	191	12	Pichon et Felden (2005)
Prochlorococcus	Cyanobactérie	+	31.3	Comparative/Cons. structure	x	18	7	Axmann *et al.* (2005)
Synechococcus	Cyanobactérie	+	55	Comparative/Cons. structure	x	18	7	Axmann *et al.* (2005)
Vibrio cholerae	Gamma	-	47	Comparative/Terminateurs (sRNA-Predict)	32	9	5	Livny *et al.* (2005)
Pseudomonas aeruginosa	Gamma	-	66	Comparative/Terminateurs (sRNA-Predict)	38	34	17	Livny *et al.* (2006)
*Bacillus subtilis**	Firmicutes	+	43	Comparative/Synthénie (QRNA)	70	12	8	Silvaggi *et al.* (2006b)
Sinorhizobium meliloti	Alpha	-	62	Comparative (QRNA/ RNAz)	32	32	7	del Val *et al.* (2007)
				Comparative	67	67	17	Ulvé *et al.* (2007)
				Promoteur/Terminateur/Comparative (sRNAPredict)	17	17	8	Valverde *et al.* (2008)
Borrelia burgorferi	Spirocheates	-	28	Comparative	x	x	2	Ostberg *et al.* (2004)
Streptomyces coelior/S.avermitilis	Actinobactéries	+	72.9	Comparative/Recherche des terminateurs	37	32	9	Pánek *et al.* (2008)
Burkholderia cenocepacia	Béta	-	67	Comparative (QRNA)	3441	213	4	Coenye *et al.* (2007)

TAB. 9.2 – Tableau récapitulatif des études de recherche systématique des ARNnc effectuées sur les génomes autres que *E. coli*. La lettre x signifie : le nombre de candidats de départ n'a pas été communiqué dans l'étude. * signifie :les ARNnc de *B. subtilis* identifiés dans les conditions de sporulation.

Chapitre 10

Expertise des candidats ARNnc de *R. solanacearum* en vue de leur validation biologique

Une fois les résultats de RNAsim générés, il a été nécessaire de passer par une étape d'analyse fine de ces résultats afin de pouvoir sélectionner le sous-ensemble de candidats le plus prometteurs. Ces candidats auront une présomption suffisamment forte permettant d'envisager leur mise en évidence biologique.

10.1 Critères pour le choix de candidats

La sélection des candidats a été faite en suivant une « stratégie » qui était fondée sur un certain nombre de critères aux motivations bioinformatiques et biologiques que nous appelerons les éléments prédictifs. Ces éléments prédictifs ont été en grande partie empruntés aux études présentées précédemment dont nous, à savoir :

1. Conservation partielle d'IGR Les ARNnc sont souvent conservés en séquence entre les organismes proches. Ainsi la conservation d'une région ne contenant pas de gènes peut indiquer la présence d'un ARNnc.

Lorsqu'une IGR est conservée en séquence, nous pouvons distinguer plusieurs schémas de conservation :

– Conservation de l'IGR entière

– Conservation partielle de l'intérieur de l'IGR
– Conservation des régions en amont ou en aval des gènes adjacents à l'IGR

Au cours de notre analyse, nous avons pu constater que la conservation partielle de l'IGR est particulièrement indicative, écartant la possibilité d'une conservation de l'IGR en raison de la présence des séquences régulatrices des gènes adjacents, ce qui peut être le cas dans les deux autres types de conservation. L'argument en faveur de la présence d'un ARNnc dans la zone conservée est d'autant plus fort que la conservation n'est pas restreinte aux génomes d'une espèce bactérienne. Ainsi, nous distinguerons la conservation au sein de trois souches de *R. solanacearum* de la conservation plus large que nous appellerons le *noyau de conservation*.

2. Conservation de la structure secondaire Les membres d'une même famille d'ARNnc sont souvent conservés en structure, même si leurs séquences primaires ont divergés au cours de l'évolution. L'existence de mutations compensatoires, dans une structure secondaire prédite dans des séquences homologues, constitue donc un argument fort pour renforcer l'existence d'un élément fonctionnel structuré.

3. Présence des terminateurs rho-indépendants prédits Comme nous l'avons déjà indiqué (section 1.1), il est démontré que la transcription chez les bactéries, peut être terminée de deux façons : à l'aide des protéines rho et à l'aide des terminateurs intrinsèques dits rho-indépendants. Contrairement aux premiers, les terminateurs rho-indépendants sont bien caractérisés, se trouvant dans les parties 3' des unités de transcription et ayant une structure secondaire stable en tige-boucle suivie par une queue poly-T (Ermolaeva *et al.* (2000), voir figure 10.1).

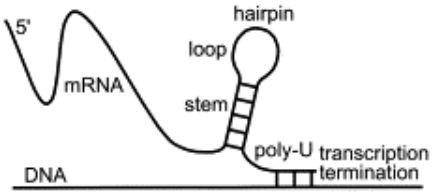


FIG. 10.1 – Modèle de structure d'un terminateur rho-indépendant (Ermolaeva *et al.* (2000)).

Nous avons donc considéré la présence et la conservation d'un terminateur orphelin rho-indépendant prédit en aval d'un candidat ARNnc comme un élément prédictif fort.

Néanmoins, il n'a pas été rendu indispensable car une proportion considérable d'ARNnc ne présentent pas de terminateur (Wassarman *et al.* (1999)).

4. Synténie La plupart des ARNnc agissant en *trans* ne sont pas associés aux gènes adjacents (les gènes adjacents ne constituent pas leurs cibles et ils ne sont pas impliqués dans les voies de régulation de ces gènes) et pour un ARNnc donné, les gènes adjacents sont généralement conservés entre génomes proches (par exemple, deux souches du même organisme). Pour plus de détails, on consultera une étude sur la conservation des gènes adjacents des ARNnc chez *E. coli* menée par Hershberg *et al.* (2003). D'autre part, l'IGR séparant deux gènes peut être conservée en raison de la présence d'éléments régulateurs associés à ces gènes qui ne sont pas des ARNnc, tels que les régions promotrices. Ainsi, la non-conservation de la synténie des gènes adjacents d'un candidat ARNnc indique que cet élément a été conservé indépendamment des gènes l'entourant et qu'il peut potentiellement s'agir d'une unité transcriptionnelle indépendante. Pour cette raison nous avons utilisé la non-conservation de la synthénie présente comme un élément prédictif.

5. La composition en G+C Cet élément prédictif a été considéré pour les raisons suivantes :

- *Pertinence de l'alignement Blast* La méthode que nous proposons est, entre autres, basée sur la recherche des zones conservées dans les IGR données par les alignements Blast. Il est donc important de s'assurer de la pertinence de ces alignements. Or, lorsque les séquences G+C riches sont alignées entre elles, les scores ainsi qu'un pourcentage d'identité élevés peuvent être obtenus en raison de la présence d'un grand nombre de nucléotides G et C dans les deux séquences alignées (illustré dans la figure 10.2). Etant donné que nous avons analysé les résultats de comparaison entre les souches de *R. solanacearum*, qui sont toutes G+C riches, un grand nombre d'alignement se trouvait dans cette catégorie.
- *Fiabilité de la structure secondaire prédite* RNafold introduit un biais lié au G+C% lorsque la structure secondaire d'une séquence est calculée. Ce biais est dû à la plus grande stabilité des appariements G-C par rapport aux appariements A-U ou G-U, favorisant les appariements G-C dans les structure secondaires prédites à l'aide du modèle thermodynamique.

– *Rupture de contexte* Une zone A+T riche, au sein d'un génome G+C riche témoigne d'une pression sélective potentiellement exercée sur cette zone agissant contre la tendance d'homogénéisation en composition dans le génome et en conséquence d'un rôle régulateur possible.

Pour les raisons évoquées, le G+C% a constitué un critère prédictif : une attention particulière a été portée aux candidats ayant un G+C% inférieur à la moyenne génomique.

Chacun des critères prédictifs énumérés ci-dessus s'est vu attribuer un poids en fonction de sa valeur prédictive, évaluée *ad hoc* (tableau 10.1). Ce poids a ensuite été utilisé dans le calcul d'un score calculé pour les candidats examinés.

En plus des éléments prédictifs bioinformatiques, lors du choix des candidats à tester expérimentalement, les partenaires biologistes ont été guidés par les critères permettant de présumer un lien de l'ARNnc candidat et la pathogénie. Ces critères sont la proximité des gènes de pathogénie, étant donné que, chez *R. solanacearum*, les gènes impliqués dans la pathogénie sont souvent co-localisés dans le génome, sous forme d'îlots fonctionnels. Certaines des régions ACUR correspondent à ces îlots de pathogénie.

10.1.1 Stratégie de recherche des ARNnc

RNAsim a généré 1105 candidats (801 sur le chromosome et 304 sur le mégaplasmide). Même si un nombre important de candidats peut être éliminé immédiatement de l'analyse grâce aux fonctionalités associées à RNAsim (voir section 7.3) (ceux correspondant aux régions d'insertion, aux régions ayant été dupliquées dans le génome)

```
Query:    115 ACAGTCCCGCTCCGGCTACGCGTCGCAGCGGCAAGCG-TGGATTCGTCTCGATTTCAACT 173
              AC G  CCGC CCGGC  C CG C   GCG C   CG TG    CG CT G    CA C
Sbjct: 370113 ACCG--CCGCCCCGGCCGC-CGCCATCGCGCCGTCCGATGCCGGCGGCTTGCCCGCATCG 370169

Query:    174 TCCGGACGAATCG-GCCGGTGGGCGCAGGGCATC-TCCCGCGCAAGCCTGATCGAGCCGT 231
               CCG  CG A C  GC GGTGG C CAG   A   T CCG GC  GCC G  C  GCCG
Sbjct: 370170 CCCGCGCGGACCCTGCTGGTGGCC-CAGCAGACGGTGCCG-GCC-GCC-GCGCC-GCCGC 370224
```

FIG. 10.2 – Aperçu d'un alignement de deux régions G+C riches. Le taux d'identité de l'alignement est de 60% et la E-vleur de $2.2e^{-05}$. Cet alignement est composé des "stretchs" de G et C, présents en grand nombre dans les deux séquences en raison de leur compostion riche en G+C. Il ne reflète pas une similarité biologiquement pertinente entre deux séquences.

Eléments prédictifs **bioinformatiques**	Poids
Conservation partielle de l'IGR au sein de trois souches de *R. solanacearum*	1
Noyau de conservation (conservation dans la banque nr)	1
Conservation de la structure secondaire, présence des mutations compensatoires	1
Présence et conservation des terminateurs rho-indépendants prédits	1
Synthénie	0.5
Régions A+T riche	0.5
Eléments de considération **biologiques**	
Appartenance à une région ACUR	
Proximité des gènes de pathogénie	

TAB. 10.1 – Les critères utilisés pour le choix des candidats ARNnc avec la valeur prédictive ("le poids") correspondante qui leur a été associée et les éléments de considération ayant guidé le choix des candidats du point de vue biologique.

Réplicon	Nb. candidats à 3 él.	Nb. IGR>500nt	Nb. IGR>500nt contenant cand. à 3 él.
Chromosome	801	146	28
Mégaplasmide	304	112	47

TAB. 10.2 – Rapartition entre les deux réplicons de *R. solanacearum* des candidats RNAsim dans les grandes IGR (>500nt).

l'examen systématique des candidats restants n'a pas été envisageable dans un premier temps.

Comme les candidats les plus nombreux sont issus des composantes connexes à trois éléments (773 sur 1105 candidats, voir section 8.4.2), nous avons restreint l'analyse à ceux de ces candidats qui appartiennent aux IGR de longueur supérieure à 500nt, appelées grandes IGR (tableau 10.2). Les génomes bactériens sont compacts et la conservation des zones dans les grandes IGR peut indiquer l'existence d'éléments fonctionnels conservés dans ces régions. Ce constat argumente le choix de l'examen prioritaire de ces régions.

Les composantes connexes à plus de 3 éléments ont été examinées systématiquement.

Le "score" de chaque candidat a été calculé comme la somme des contributions de chacun des éléments prédictifs énumérés dans la section 10.1, selon le schéma présenté dans la figure 10.3.

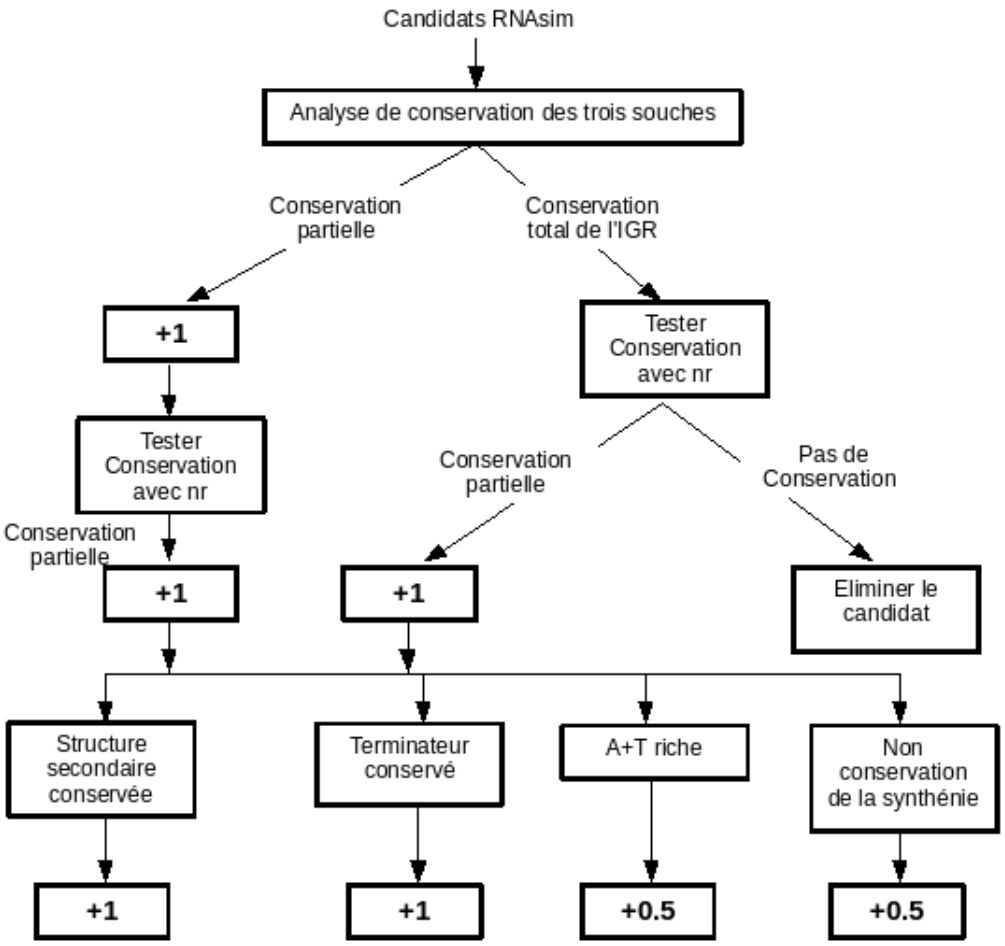


FIG. 10.3 – Stratégie de recherche des candidats pour validation biologique à partir des résultats RNAsim.

Les candidats au meilleur score seront présentés dans la suite.

10.2 Candidats ARNnc

Plus de 300 candidats ont été analysés et les plus prometteurs, selon le schéma de choix de candidats présentés mais aussi selon le choix fait par les biologistes, ont été retenus pour validation.

Dans la suite, nous donnons la présentation détaillée du premier candidat suivie par la liste des autres candidats retenus.

10.2.1 Candidat 1

Cet élément a été identifié par RNAsim du fait de sa conservation entre les trois souches de *R. solanacearum*. Le noyau de conservation entre trois souches correspond à une région A+T riche délimitée par des terminateurs prédits conservés (fig. 10.4, trait rouge en pointillé). Les terminateurs de transcription se trouvant sur le brin - , l'analyse de ce candidat sera présentée avec les séquences de ce brin.

La région intergénique correspondante est aussi conservée dans 5 autres génomes des béta-protéobactéries (figure 10.4, conservation dans *R. pickettii*, *R. metallidurans*, deux souches de *R. eutropha* et *C. taiwanensis*), sur une zone plus courte, représentée par les traits pointillés en rose sur la figure. Le noyau de conservation est réduit à la zone délimitée par les deux terminateurs prédits (terminateurs 2 et 3, figure10.4)

et cette zone varie en longueur de 67 à 90 nucléotides. Toutes les zones conservées se trouvent au milieu de

grandes régions intergéniques des génomes correspondants et les terminateurs prédits y sont conservés en structure (voir figure 10.4).

Ces constats sont également visibles dans l'alignement multiple des zones conservées dans tous les génomes considérés (figure 10.5). Le bloc de conservation apparent entre les positions 55 à 135 environ correspond à la zone conservée dans les huit génomes et les zones encadrées correspondent aux terminateurs prédits 2 et 3 (voir figure 10.4).

L'ensemble de ces éléments suggère que la région identifiée serait conservée possiblement en tant qu'un ARNnc.

Toutefois, l'alignement multiple de la totalité des séquences conservées ne met pas en évidence la structure de la conservation entre les organismes les plus proches, qui sont les trois souches de *R. solanacearum* et *R. pickettii* et dont la conservation est plus large. Celle-ci est révélée par l'alignement multiple des séquences de ces quatre génomes (figure 10.6). Dans cet alignement multiple nous pouvons distinguer deux blocs (les blocs encadrés dans la figure 10.6) : le premier correspondant à l'élément conservé dans les 8 génomes, mis en évidence dans l'alignement multiple de la totalité des séquences et un autre, conservé uniquement dans les génomes de *R. solanacearum* et *R. pickettii*. Ce deuxième bloc se termine par une queue poly-T correspondant au terminateur prédit 1. Malgré l'absence de la conservation du deuxième bloc dans les 4 autres génomes sa bonne conservation à l'intérieur des régions intergéniques et la présence d'un terminateur prédit conservé pourraient indiquer qu'il constitue une autre

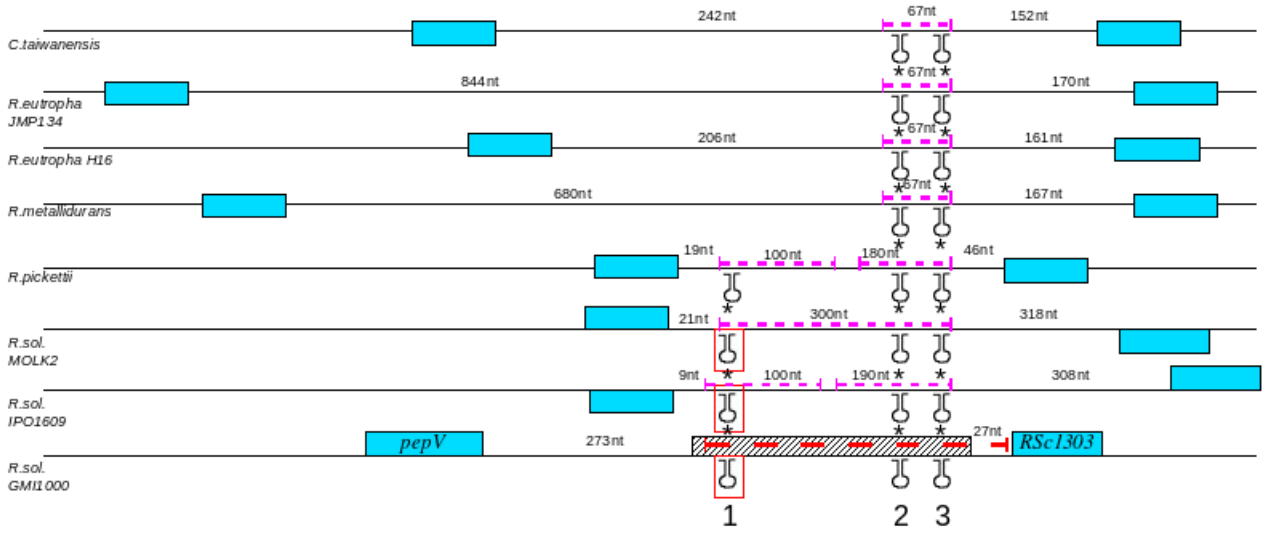


FIG. 10.4 – Schéma de conservation du Candidat 1. La prédiction RNAsim est représentée par la zone hachurée. Les gènes sont représentés par les rectangles bleus. Les gènes sur le brin + se trouvent au dessus et les gènes sur le brin - , au-dessous de la ligne de référence. Les gènes pour lesquels nous ne connaissons pas le brin sont situés au milieu de la ligne. Les proportions de la taille des gènes et des régions intergéniques ne sont pas strictement respectées. Les terminateurs prédits sont annotés par les signes en tige-boucle, portant les numéros 1, 2 et 3. Les terminateurs présentant des mutations compensatoires par rapport à la séquence de *R. solanacearum* GMI1000 sont annotés par une étoile et ceux déjà annotés dans iANT sont encadrés en rouge. La région A+T riche est annotée par le trait rouge pointillé et les régions conservées par les traits roses pointillés. La longueur des régions conservées ainsi que la distance entre les gènes voisins sont marqués au-dessus des régions correspondantes. Les abréviations : *R.pickettii* - *Ralstonia pickettii 12J*, *R.metallidurans* - *Ralstonia metallidurans CH34*, *R.eutropha* - *Ralstonia eutropha* et *C.taiwanensis* - *C.taiwanensis LMG19424*.

unité de transcription codant pour un ARNnc.

L'analyse, par RNAz, de l'alignement multiple de chacun de ces blocs predit l'existence d'une structure secondaire conservée au sein de chacun des deux blocs. Ces structures sont présentées dans la figure 10.7.

Notons que les structures prédites du bloc 1 et du bloc 2 présentent des ressemblances : elles contiennent, toutes les deux, deux structures en tige-boucle et une région variable les séparant. De plus, si nous "ouvrons" la base de la tige-boucle 2 du bloc 2 de façon à avoir une tige-boucle plus stable dans laquelle la boucle interne a disparu, nous obtenons alors deux structures très voisines.

Le Candidat 1 serait un membre de la famille suhB *de Rfam*

Au cours de notre travail de thèse une nouvelle prédiction d'ARNnc dans le génome de *R. solanacearum* a été répertoriée dans Rfam, correspondant à une partie du bloc 2 du Candidat 1. Il s'agit d'un membre de la famille des ARNnc putatifs *suhB*, à fonction inconnue, prédite par approche comparative dans les alpha-protéobactéries (Corbino *et al.* (2005)).

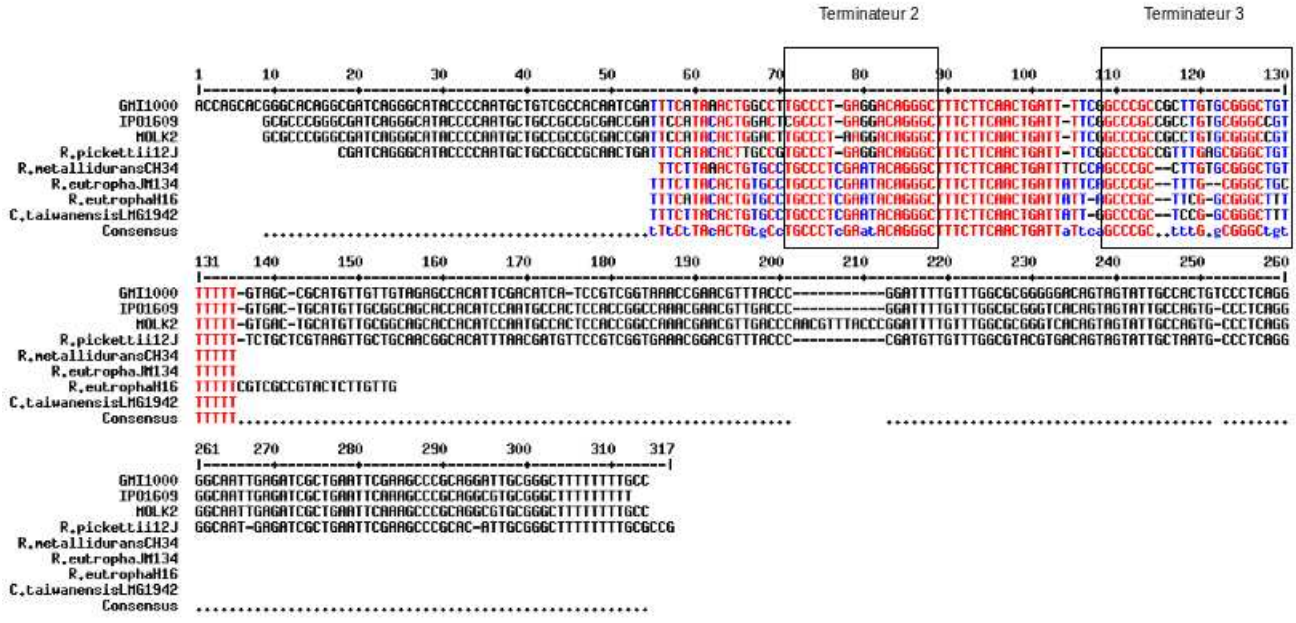


FIG. 10.5 – L'alignement multiple (par Multalin) des régions conservées. Les zones correspondant aux terminateurs prédits 2 et 3 de la figure 10.4 sont encadrés.

A ce jour, Rfam recense 31 membres de la famille *suhB* et *R. solanacearum* est actuellement la seule béta-protéobactérie chez laquelle un membre de cette famille ait été prédit. L'alignement générique en structure proposé pour cette famille d'éléments est présenté dans la partie 1 de la figure 10.8.

Cet élément est caractérisé par deux tiges-boucles conservées en structure mais pas en séquence. Les zones reliant ces deux éléments sont hautement variables, en séquence et en longueur. Structure secondaire prédite dans Rfam est donnée dans la figure 10.9 a).

L'occurrence de *suhB* dans la souche GMI1000 de *R. solanacearum* correspond au bloc 2 de conservation. En conséquence, nous avons comparé les autres éléments de ce bloc à l'alignement structural *suhB* (partie encadrée 2 de la figure 10.8). Les séquences du bloc 2 sont en adéquation avec cet alignement, possédant les deux éléments de structure secondaire. En particulier, la deuxième tige-boucle prédite correspond, dans le bloc 2, aux séquences du terminateur 1 (figure 10.4) et ne constitue donc pas, selon Rfam, un terminateur de transcription proprement dit.

Au vue de ces résultats, nous pouvons conclure que les membres de la famille des

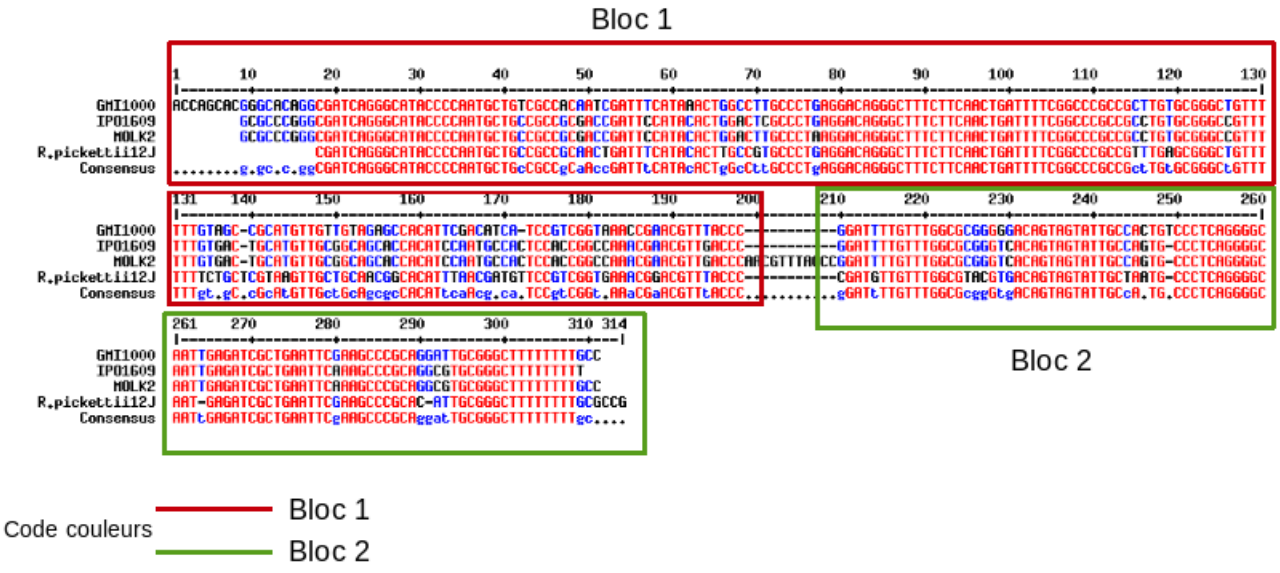


FIG. 10.6 – Alignement multiple des séquences conservées entre les génomes de trois souches de *R. solanacearum* et de *R. pickettii*. Deux blocs de conservation sont apparents, encadrés en bordeaux et vert.

ARNnc putatifs *suhB* sont également présents dans les souches IPO1609 et Mol2 de *R. solanacearum* ainsi que dans le génome de *R. pickettii*.

Quant aux éléments du bloc 1, la similarité de leur structure commune prédite avec celle du bloc 2 était visible dans la figure 10.7. Leur structure est également en adéquation avec la structure prédite de *suhB* (figure 10.9 a) et b)). Pour cette raison les éléments du bloc 1 ont été comparés avec l'alignement structural de *suhB* (figure10.8, partie encadrée 3). Ces séquences semblent en adéquation avec l'alignement structural, malgré les différences importantes en séquence avec le bloc 2 présentant, quant à lui, de façon certaine des occurrences de *suhB*. De ce fait, le bloc 1 pourrait en comporter une autre occurrence.

Cela signifierait que, dans chacune des trois souches de *R. solanacearum* et dans le génome de *R. pickettii*, deux occurrences de *suhB*, se trouvant dans la même région intergénique, pourraient être présentes. Nous aborderons ce point plus amplement dans la discussion.

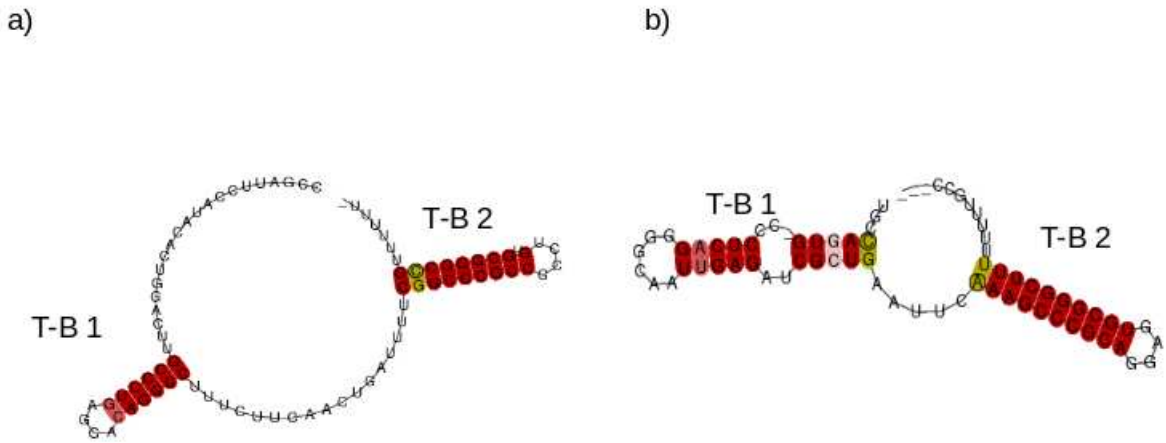


FIG. 10.7 – La structure secondaire commune aux séquences de chacun des blocs identifiés dans la figure 10.6), prédite par RNAz. a) La structure secondaire prédite pour le bloc 1 b) La structure secondaire prédite pour le bloc 2. Les nucléotides en rouge correspondent aux regions sans mésappariements dans l'alignement multiple. La couleur rouge pâlit avec le nombre de mésapariements grandissant. Les nucléotides en vert désignent la présence des mutations compensatoires au niveau du nucléotide encerclé. Dans chacune des figures les tiges-boucles prédites sont notées par T-B 1 et T-B 2.

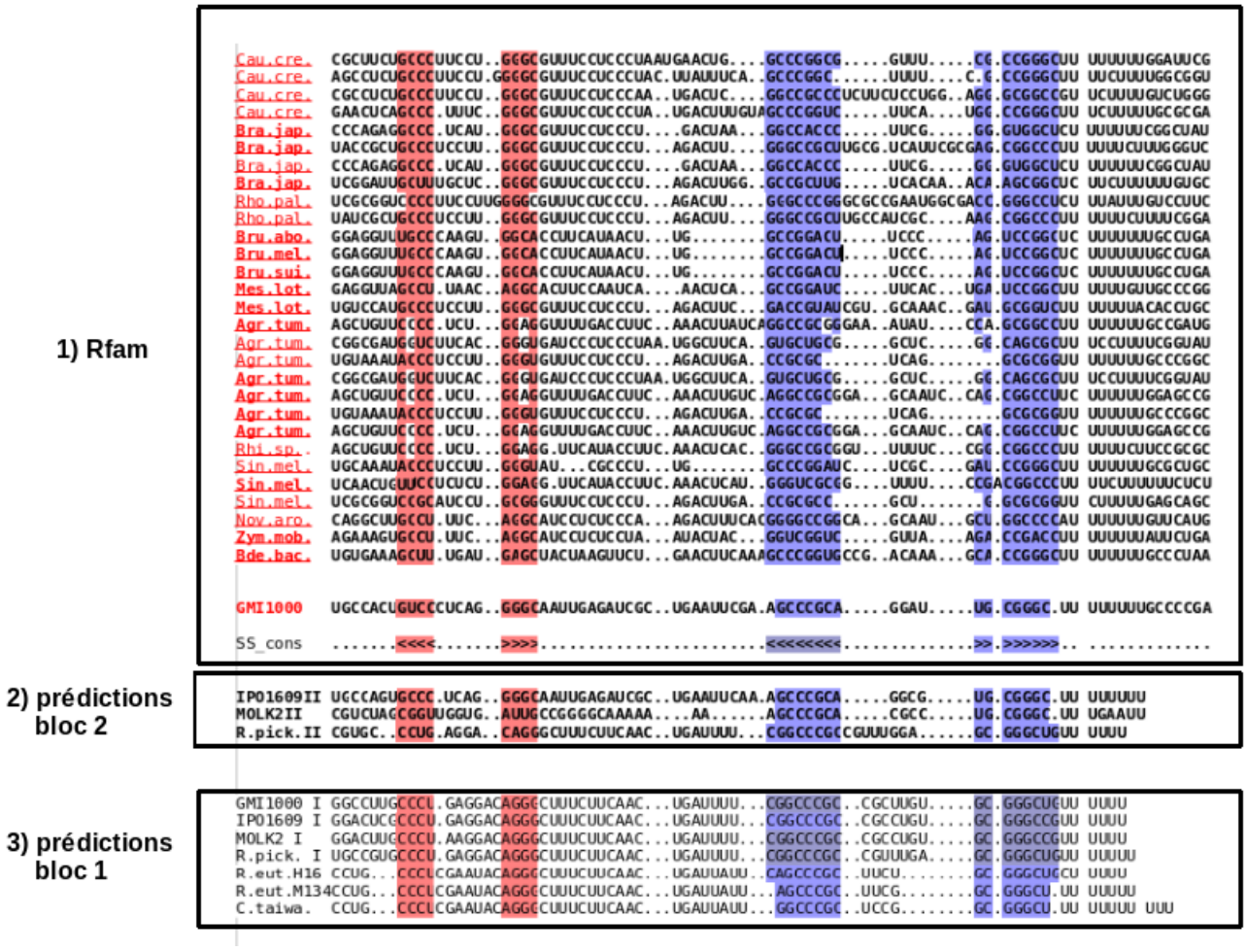


FIG. 10.8 – L'alignement structural Rfam. Partie encadré 1) L'alignement structural Rfam de l'ARNnc prédit *suhB*. Chacune de couleurs présente un des éléments de structure secondaire conservés : en rouge une tige de longueur 4 et en bleu une hélice de longueur 7 à 8 nt. La structure secondaire est représentée dans la ligne SS_cons, < > représentant un appariement de deux bases. La légende des noms des organismes dans l'alignement Rfam est donnée dans l'annexe. L'alignement de la séquence de *R. solanacearum* GMI1000 dans Rfam, appartenant au bloc 2, est présentée à part en bas. Les organismes présents dans l'alignement : *Caulobacter crescentus CB15* (Cau. cre.), *Bradyrhizobium japonicum symbiotic gene region* (Bra. jap.), *Rhodopseudomonas palustris CGA009* (Rho. pal.), *Brucella abortus biovar 1 str. 9-941* (Bru. abo.), *Brucella melitensis 16M* (Bru. mel.), *Brucella suis 1330* (Bru. sui.), *Mesorhizobium loti MAFF303099* (Mes. lot.), *Agrobacterium tumefaciens str. C58* (Agr. tum.), *Rhizobium sp. NGR234, cosmid pXB9* (Rhi. sp.), *Sinorhizobium meliloti 1021* (Sin. mel.), *Novosphingobium aromaticivorans* (Nov. aro.), *Zymomonas mobilis subsp. mobilis ZM4* (Zym. mob.), *Bdellovibrio bacteriovorus* (Bde. bact.). Partie encadrée 2) L'alignement des éléments du bloc 2 fait en respectant l'alignement structural de *suhB*. L'appartenance au bloc 2 est signalée avec le numéro II à côté du nom du génome. 3) L'alignement des éléments du bloc 1 fait en respectant l'alignement structural de *suhB*. L'appartenance au bloc 1 est signalée avec le numéro I à côté du nom du génome.

Discussion

Deux éléments conservés ont été identifiés après l'analyse de Candidat 1 de RNAsim (les éléments identifiés avec leurs positions génomiques sont résumés dans le tableau 10.3). Ces deux éléments constituent des blocs de conservation entre quatre génomes : trois souches de *R. solanacearum* et *R. pickettii*.

Un de ces éléments (correspondant au bloc 2 du tableau 10.3) correspond à une occurence de la famille des ARNnc putatifs *suhB* de Rfam, répertoriée récemment dans le géome de *R. solanacearum*. Une occurence de cet élément se trouve dans chacun des génomes de trois souches de *R. solanacearum* et de *R. pickettii* ; celles correspondantes aux souches IPO1609 et Molk2 et à *R. pickettii* ne sont pas répertoriées dans Rfam.

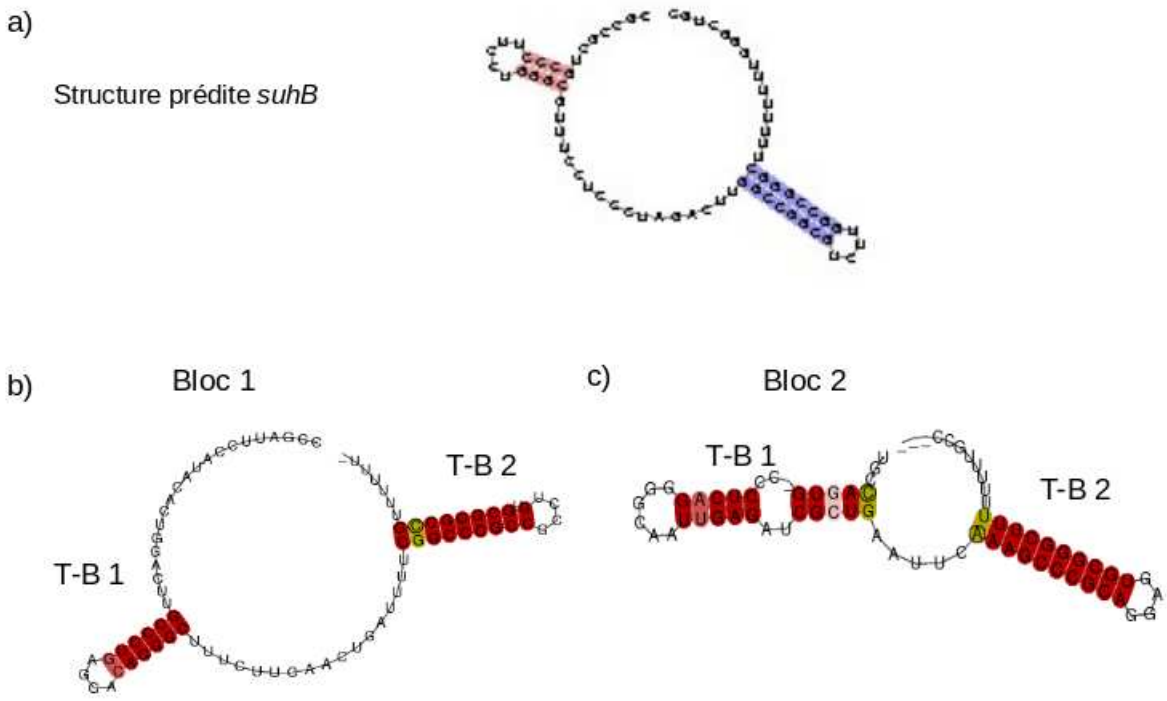


FIG. 10.9 – Les structures secondaires prédites. a) La structure secondaire de l'élément *suhB* prédite dans Rfam. Les couleurs correspondent aux couleurs de l'alignement structural de *suhB* (partie encadrée 1 de la figure 10.8). b), c) correspondent aux structures secondaires communes prédites des blocs 1 et 2, respectivement (voir la figure 10.6).

Bloc	Génome	Brin	Position début	Position fin	Longueur
2	*R. solanacearum* GMI1000	-	1390699	1390629	70
2	*R. solanacearum* IPO1609	+	1657145	1657211	66
2	*R. solanacearum* Molk2	-	181232	181163	69
2	*R. pickettii* 12J	-	1267493	1267436	57
1	*R. solanacearum* GMI1000	-	1390876	1390794	82
1	*R. solanacearum* IPO1609	+	1656957	1657038	81
1	*R. solanacearum* Molk2	-	181420	181339	81
1	*R. pickettii* 12J	-	1267678	1267598	80
1	*Ralstonia metallidurans* CH34	-	2161892	2161815	77
1	*Ralstonia eutropha* H16	-	2479440	2479345	95
1	*Ralstonia eutropha* JMP134	-	2209784	2209708	76
1	*Cupriavidus taiwanensis* LMG19424	-	1926828	1926752	76

TAB. 10.3 – Positions, brin et longueur génomiques des éléments identifiés appartenant aux deux blocs de conservation.

Nous avons ainsi recensé trois nouveaux membres de la famille *suhB*, en plus de 31 membres prédits dans Rfam. La souche GMI1000 est, à ce jour, la seule béta-protéobactérie chez qui l'élément *suhB* a été prédit. Nos résultats montrent qu'il existe dans le génome de *R. pickettii* et suggèrent qu'il pourrait être répandu plus largement parmi les béta-protéobactéries.

La raison pour laquelle les membres de la famille *suhB* des souches Molk2, 1609 de *R. solanacearum* et de *R. pickettii* n'ont pas été répertoriés dans Rfam peut probablement être expliquée par l'absence de ces génomes de la banque génomique ENSEMBL ayant servi à générer les prédictions Rfam.

En plus de la région correspondant au bloc 2, un deuxième élément est conservé au sein de la même IGR (bloc de conservation 1 dans le tableau 10.3). Ce bloc contient une zone de 67nt présentant une forte homologie avec quatre autres génomes du groupe des béta-protéobactéries (*R. metallidurans*, deux souches de *R. eutropha* et *C. taiwanensis*). Sa large conservation, la localisation au milieu des grandes régions intergéniques dans tous ces génomes et la présence des mutations compensatoires maintenant des parties structurées suggèrent qu'il s'agit d'un élément de régulation et potentiellement d'un ARNnc.

Nous nous sommes appuyés sur la similarité en structure des séquences du bloc 1 et du bloc 2 ainsi qu'à l'adéquation avec l'alignement structural de *suhB*, pour émettre l'hypothèse qu'il s'agirait d'une autre occurrence de *suhB* dans la même région intergénique. Plusieurs occurrences de *suhB* sont répertoriées dans la plupart des alpha-

protéobactéries chez lesquelles cet élément a été prédit (par exemple *A. tumefaciens* ou *S. meliloti*, voir figure 10.8). La séquence de ces occurrences au sein du même génome ne sont pas toujours similaires (par exemple les séquences des occurrences dans *Sinorhizobium meliloti*, figure 10.6). D'autre part, il existe des exemples d'ARNnc de la même famille dont les occurrences se trouvent dans la même région intergénique (les ARNnc *rygA* et *rygB* dans le génome *E.coli*). L'ensemble de ces faits indique qu'une deuxième occurrence de *suhB* dans une même région intergénique est tout à fait probable. Cette seconde occurrence aurait été omise dans Rfam en raison de la faible similarité en séquence entre les deux occurrences et avec les autres membres de cette famille préalablement prédits chez d'autres organismes. Néanmoins, l'existence d'une telle occurrence au sein du bloc 1 est une hypothèse qui doit être encore vérifiée.

Notons enfin que l'identification de ce candidat est un premier élément qui permet de valider l'approche méthodologique que nous avons développée pour identifier des ARNnc dans le génome de *R. solanacearum*.

10.2.2 Autres candidats ARNnc

Au terme de l'analyse des résultats de RNAsim et après discussion avec les partenaires biologistes, huit candidats les plus prometteurs ont été sélectionnés. Ils sont représentés sur le tableau récapitulatif 10.4, ensemble avec leurs éléments prédictifs et le score qui leur a été associé.

10.3 Autres éléments régulateurs potentiels

10.3.1 Eléments en amont des gènes *popF1* et *popF2*

Deux régions conservées, comportant une structure en tige-boucle et se trouvant en amont des gènes *popF1* et *popF2* portés par le mégaplasmide de *R. solanacearum*, ont été identifiées dans une composante connexe à deux éléments, lors de la comparaison de la souche GMI1000 contre elle-même.

Ces deux régions, longues de 135nt, partagent 82% d'identité tandis que les régions flanquantes des deux régions intergéniques ne sont pas homologues. La conservation de cette région a déjà observée par Meyer *et al.* (2006).

Candidat	Rép.	Gènes adj.	Orientation	Extr.5'	Extr.3'	Long.	Long. IGR	Eléments prédictifs						Score	Crit. bio.	Cons. 1 org.
								Cons. part.	nr	s.s.	Term.	Non synthénie	Comp.			
Cand 1.1	c	pepV/RSc1303	>< >	1390872	1390777	95	571				1			4		*R. pickettii*
Cand 1.2*	c	pepV/RSc1303	>< >	1390699	1390625	74	571							3		*R. pickettii*
Cand 2	c	RSc3386/RSc3387	>>>	3651833	3652007	174	914				3			2.5		
Cand 3	p	Rsp1268/RSp1269	>< <	1605747	1605582	165	563				2			2		*R. pickettii*
Cand 4	p	Rsc1022/prhG	>>>	1290424	1290573	149	816							1.5		
Cand 5	c	Rsc0973/RSc0974	<><	1021907	1022205	298	517							4		
Cand 6*	p	Rsp0568/RSp0569	>><	709773	709875	102	338							3.5		
Cand 7-cop1	c	Rsc3149/RSc3150	>< >	3400046	3399987	59	1045				?			3		
Cand 7-cop2	c	Rsc3358/dbhB	<>>	3617829	3618014	186	642				?			3		
Cand 7-cop3	c	Rsc3084/RSc3085	<<<	3322751	3322662	89	1069				?			3		
Cand 8	c	Rsc0824/PopP1	>>>	866550	866863	313	544				?			2		*V. eiseniae*

TAB. 10.4 – Tableau récapitulatif des meilleurs candidats RNAsim. Liste des abbréviations : Rép.-réplicon (c-chromosome, p-mégaplasmide) ; Gènes adj.-gènes adjacents ; Orientation se réfère à la l'orientation du candidat et des gènes adjacents : $>$ represénte le brin $+$ et $<$ le brin -. Extr. 5'(3')-extrémité 5'(3') ; Long.-longueur ; Cons. part.-conservation partielle ; nr-conservation dans la banque nr ; s.s.- conservation de la structure secondaire ; Term. - présence de terminateurs orphelins ; Comp.- G+C% plus bas que la moyenne génomique ; Crit. bio.-éléments de considération biologique (pour les huit dernières catégories voir section 10.1). Cons. 1 org. : candidat est conservé dans un organisme de nr. Une case est en gris lorsque le critère correspondant est rempli. "*" : la région correspondante a été répertoriée dans Rfam comme appartenant à la famille putative *suhB* (voir section 10.2.1). "*" : les séquences IS ou les éléments mobiles du génome se trouvent dans la proximité du candidat. Cand 1.1 et Cand 1.2 appartiennent à la même région intergénique. Candidats 7-cop1, 2 et 3 représentent le Candidat 7 qui est répété trois fois dans le génome. Le nombre dans la case Term. représente le nombre de terminateurs orphelins prédits dans l'IGR correspondante tandis que le signe " ?" signifie que le terminateur ne se trouve pas dans le bon sens par rapport au sens de la transcription présumé.

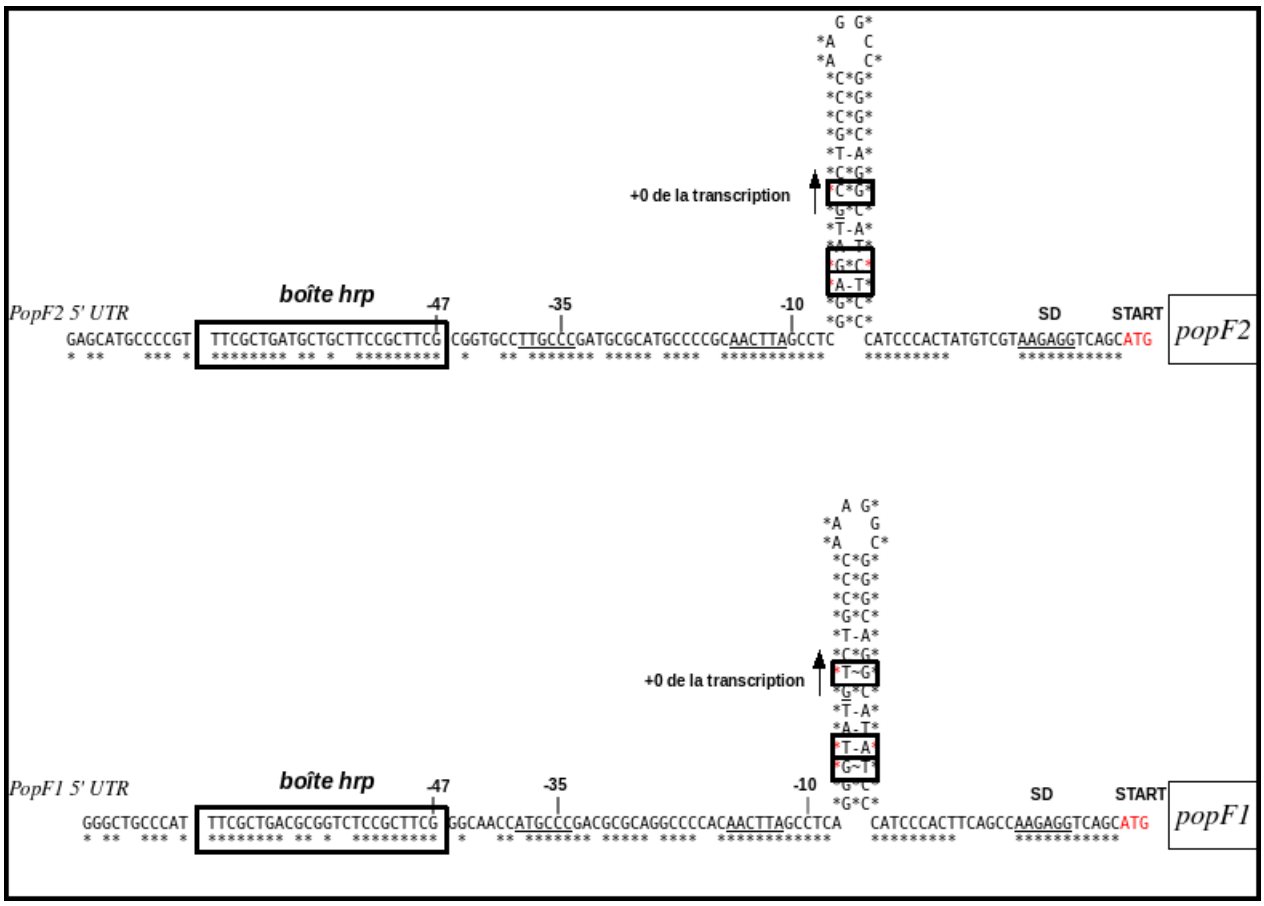


FIG. 10.10 – Structures des deux régions homologues, 135nt en amont des gènes *popF1* et *popF2* correspondant aux éléments identifiés. Les sites comprenant des mutations compensatoires dans les structures potentielles en tige-boucle sont encadrés. Le site d'initiation de la transcription (+0 de la transcription désigné par une flèche) est compris dans la structure en tige-boucle. Les sites putatifs de fixation de la l'ARN polymérase sont soulignés dans la séquence autour des position -10 et -35. La séquence de Shine-Dalgarno potentielle est indiqué par SD et le début de la séquence codante est indiqué par START. La boîte *hrp*, commencant à la position -47 est encadrée. Les positions conservées entre les deux séquences sont indiquées par * (dans le cas des mutations compensatoire * est en rouge).

Les deux éléments conservés contiennent un motif en tige-boucle de 34nt, dont 28nt sont appariés constituant une longue tige de 14nt (fig. 10.10). Une recherche PatScan de ce motif (admettant au plus 1 mésappariement), dans le génome de *R. solanacearum*, montre qu'il s'agit d'un motif qui n'est pas commun puisque les seules occurences trouvées sont celles situées en amont des gènes *popF*. De plus, les structures prédites sont conservées, entre deux tiges potentielles, du fait de l'existence de 4 mutations compensatoires entre leurs séquences, suggérant qu'il ne s'agit pas d'une structure conservée par hasard mais qu'elle pourrait avoir un réel rôle biologique.

Dans les deux cas 27nt séparent les structures potentielles du gène en aval. Ces deux gènes, *popF1* et *popF2* sont paralogues et codent pour un composant de l'appreil de sécrétion de type III probablement impliqué dans la jonction entre le pilus bactérien, des structures protéiques filamenteuses disposées sur la paroi bactérienne et la membrane cytoplasmique de la cellule végetale et sont, de cette façon, impliqués dans la pathogénie de la bactérie (Meyer *et al.* (2006)).

43nt en amont des structures conservées, à la position -47 du site d'initiation de la transcription, se trouvent, dans les deux cas, les boîtes *hrp*, correspondant aux séquences du type $TTCG$-N_{16}-$TTCG$, qui ont été identifiées comme étant un élément régulateur essentiel à l'expression transcriptionnelle des gènes associés au système de secretion de type III (Cunnac *et al.* (2004)).

La tige-boucle potentielle renferme le site d'initiation de la transcription. Ceci suggère que ces éléments pourraient jouer un rôle régulateur dans l'expression des gènes *popF1* et *popF2 via* la séquestration du site d'initiation de la transcription, révélant ainsi l'existence d'un nouveau système de régulation du pouvoir pathogène de la bactérie et en particulier des gènes *popF1* et *popF2*, qui n'est pas entièrement élucidé à ce jour (Meyer *et al.* (2006)).

L'élément est conservé dans d'autres souches de R. solanacearum

Les souches Molk2 et IPO1609 comportent seulement un des gènes de la famille *popF* : le gène *popF2*. Dans les deux souches, la séquence promotrice de ce gène comporte, à l'instar des gènes de la souche GMI1000, un motif en tige-boucle, à un mésappariement près dans la souche IPO1609 (fig. 10.12), situé à la position -26 du gène *popF2* . Un appariement G-C supplémentaire, par rapport aux tiges-boucles de la souche GMI1000, est possible à la base de la tige, renforcant ainsi la structure secondaire prédite.

Nous avons recherché une structure similaire dans une quatrième souche de *R. solanacearum*, la souche UW551, dont la séquence partielle est disponible dans NCBI. Cette souche possède un gène qui est orthologue direct du gène *popF1* de la souche GMI1000 (Meyer *et al.* (2006)).

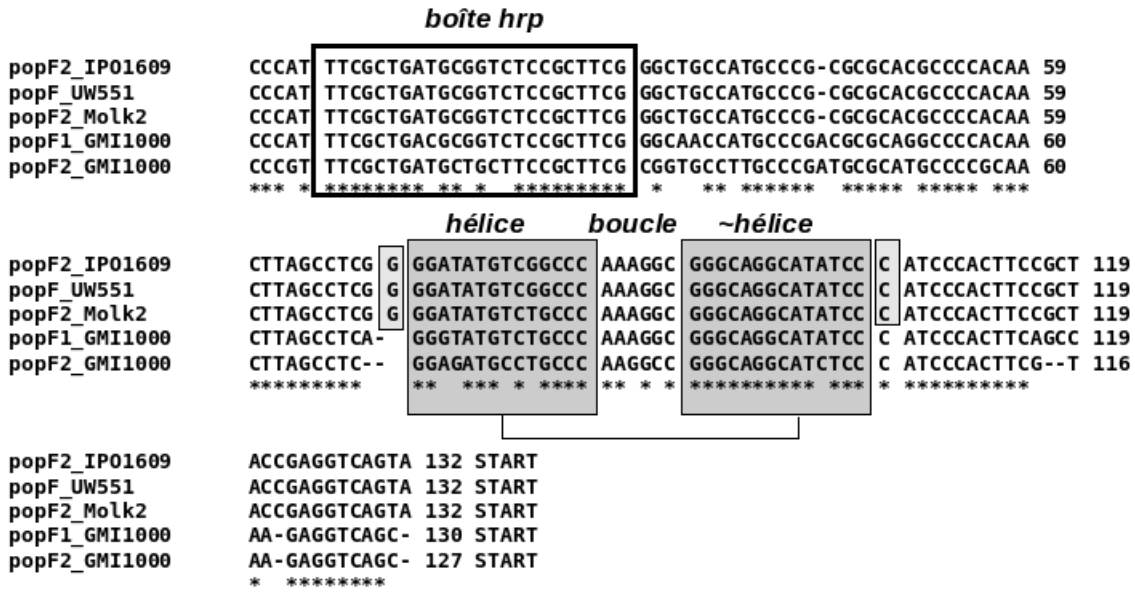


FIG. 10.11 – L'alignement multiple des régions de 130nt en amont des gènes *popF1* et *popF2*, dans la souche GMI1000, IPO1609 et Molk2 et du gène $PopF_{UW551}$ dans la souche UW551. La boîte *hrp* est encadrée. Les deux séquences formant l'hélice de la tige-boucle sont encadrées en gris. Les nucléotides encadrés en gris plus clair indiquent les appariements potentiels supplémentaires dans les souches IPO1609, Molk2 et UW551.

La région promotrice du gène $PopF_{UW551}$ est conservée avec celles des gènes *popF1* et *popF2* de la souche GMI1000 ainsi que le motif en tige-boucle, dont le positionnement par rapport au codon d'initiation est conservé. De même, cette structure est précédée par une boîte *hrp*. L'alignement des régions promotrices est représenté sur la fig. 10.11. La séquence du motif est identique à celle de la souche IPO1609, à deux nucléotides près, situés dans la boucle. Ceci était attendu, étant donné que la souche UW551 est phylogénétiquement très proche de la souche IPO1609. L'appariement des deux nucléotides qui différencient les séquences des motifs en tige-boucle dans ces deux souches prolongent, dans le souche UW551, la tige prédite (fig. 10.12).

Recherche de motifs similaires dans les régions en amont des gènes orthologues de popF1 et popF2

Une recherche Blast sur la banque *nr* n'ayant pas permis d'identifier de nouvelles séquences de ce type dans d'autres organismes, nous avons cherché à savoir si une structure secondaire similaire était présente dans les régions promotrices des gènes orthologues aux gènes *popF1* et *popF2* de *R. solanacearum* GMI1000. De tels gènes ont été identifiés par recherche Blast chez *Sinorhizobium fredii, Rhizobium sp. NGR234, Mesorhisobium loti, Xanthomonas campestris pv. campestris, Xanthomonas oryzae pv. oryzae* et *Xanthomonas campestris pv. vesicatoria*

Une recherche de motif avec PatScan dans les régions en amont de ces gènes n'a pas permis pas non plus de détecter de motif similaire au motif en tige-boucle situé en amont des gènes *popF1* et *popF2* de *R. solanacearum* GMI1000. Une recherche en utilisant le

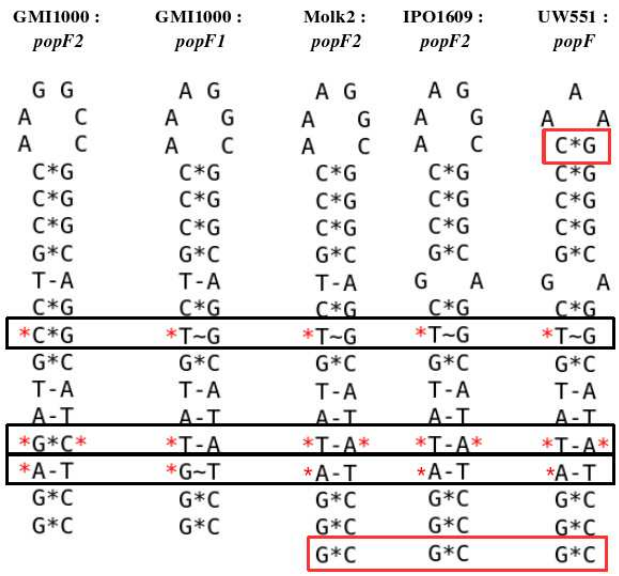


FIG. 10.12 – Comparaison des motifs en tige boucle se trouvant en amont des gènes *popF1* et *popF2* de la souche GMI1000, *popF2* des souches IPO1609 et Molk2 et $PopF_{UW551}$ de la souche UW551. Les appariements A et T sont indiqués par A-T, G et C par G*C et les appariements T et G par T G. Les positions où les changements compensatoires ont eu lieu sont indiqués par * en rouge et ces appariements sont encadrés en noir. En rouge sont encadrés les appariements potentiels supplémentaires chez les souches Molk2, IPO1609 et UW551.

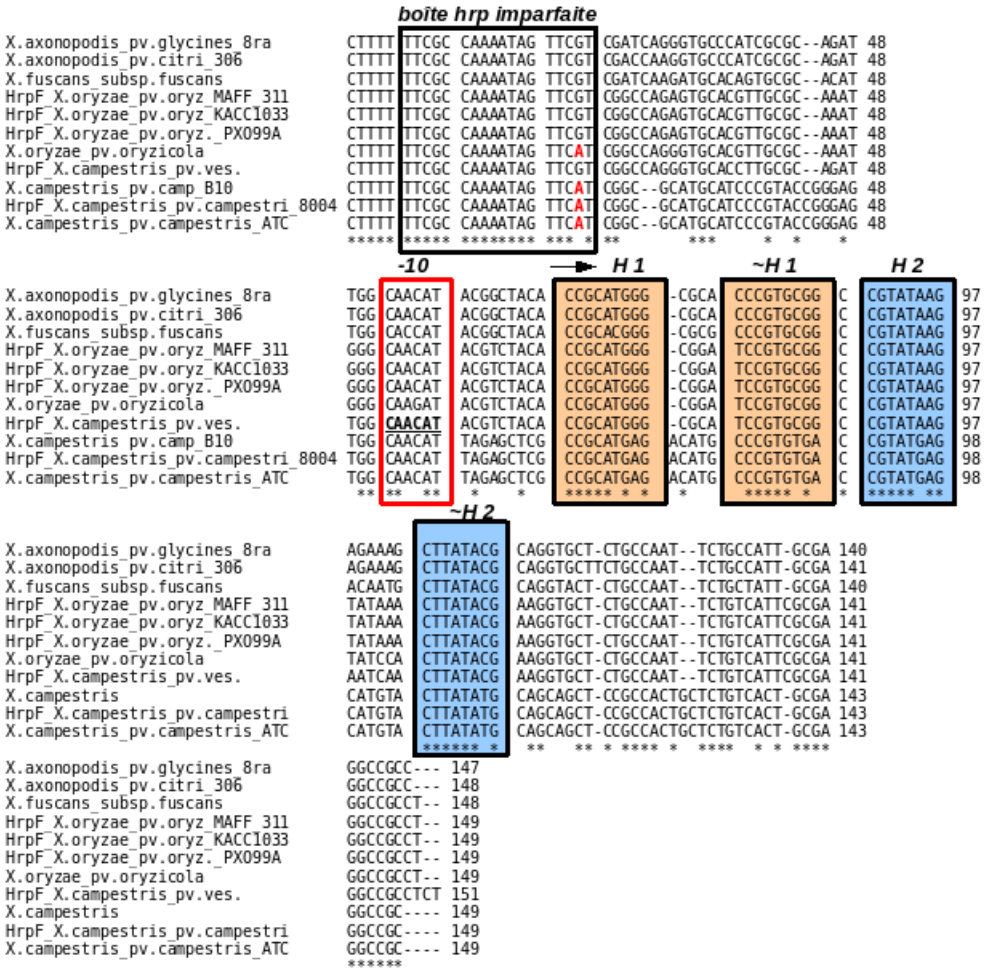


FIG. 10.13 – L'alignement multiple des séquences en amont des gènes HrpF des *Xanthomonas* contenant le motif composé de deux tiges-boucles et des séquences homologues retrouvées dans les génomes non-annotées de la banque *nr*. La boîte *hrp* imperfaite, $TTCGC\text{-}N_8\text{-}TTCGT$, est encadrée en noir. Les nucléotides en rouge, dans cette boîte, ne respectent pas le motif consensus. La position probable du -10 en amont de la position de début de transcription est encadrée en rouge. La séquence de *X.oryzae pv.vesicatoria*, qui a servi comme référence pour déterminer la position -10 de la transcription, est soulignée. La position approximative du début de la transcription est indiquée par une flèche. Les séquences appartenant à la tige du premier motif en tige-boucle sont encadrées en orange et du deuxième en bleu. H1 et ˜ H1 s'appariant pour former la tige de la première, et H2 et ˜ H2 de la deuxième boucle. Les séquences se trouvant entre les parties encadrées représentent les tiges des deux motifs appartenant aux deux boucles.

logiciel RNAfold avec fenêtre glissantes ne permet pas dedétecter de structure similaire mais met en évidence un autre motif intéressant, composé de deux tiges-boucles, présent en amont des gènes *HrpF* de *Xanthomonas oryzae pv. oryzae*, *Xanthomonas campestris pv. vesicatoria*, *Xanthomonas campestris pv. campestris*, orthologues de *popF1* et *popF2* chez *R. solanacearum*.

L'alignement multiple d'environ 150nt en amont de ces gènes aisni que des séquences homologues identifiées dans les génomes non-annotés (*X. oryzae pv. oryzicola, X. axonopodis pv. glycines str 8ra, X. axonopodis pv. citri str. 306, X. fuscans subsp. fuscans strain CFBP4834-R* et *X. campestris pv. campestris str. ATCC 33913*) montre une forte conservation de cette région dans toutes ces organismes (fig. 10.13).

Il a été montré (Koebnik *et al.* (2006) et Tsuge *et al.* (2005)) que les gènes *hrpF* de *X. campestris pv. vesicatoria* et *X.oryzae pv. oryzae* contiennent, dans leur région promotrice, une boîte *hrp* imparfaite, de forme $TTCGC\text{-}N_8\text{-}TTCGT$. Nous avons identifié cette boîte dans les régions comparées (fig. 10.13, encadré en noir), la séquence consensus étant respectée dans toutes les régions sauf, à un nucléotide près, chez *X. campestris* et *X.oryzae pv. oryzicola*.

Le motif en double tige-boucle que nous avons identifié (encadré en orange et bleu, fig. 10.13) est long d'environ 44nt et se situe 35nt en amont du codon d'initiation de la traduction et 43nt en aval de la boîte *hrp* imparfaite. Notons que chez *Xanthomonas* et chez *R. solanacearum* la distance qui sépare la boîte *hrp* du commencement de l'élément structuré est rigoureusement conservée, ce qui renforce l'hypothèse du rôle biologique de ces structures.

D'après Koebnik *et al.* (2006), l'héxamère $CAACAT$, situé 28nt en aval de la boîte *hrp* de *X. campestris pv. vesicatoria*, constitue le -10 du début de la transcription du gène *HrpF*. A partir de cette information, l'ensemble des séquences comparées étant très similaire autour de cette position (encadré en rouge sur la fig. 10.13), nous avons inféré la position hypothétique du début de transcription dans l'ensemble des séquences (indiqué par la flèche sur la figure). Selon cette supposition, le site de l'initiation de la transciption se trouverait au sein de la première tige-boucle potentielle.

Une vue plus détaillée de la région contenant le motif en deux tiges-boucles montre la conservation en séquence des deux tiges prédites (fig. 10.14a), H1˜ H1 et H2 ˜ H2) contrairement à ce qui est observé dans les boucles. Au sein de ces alignements, deux groupes de séquences peuvent être identifiés. Un premier groupe, présentant une conser-

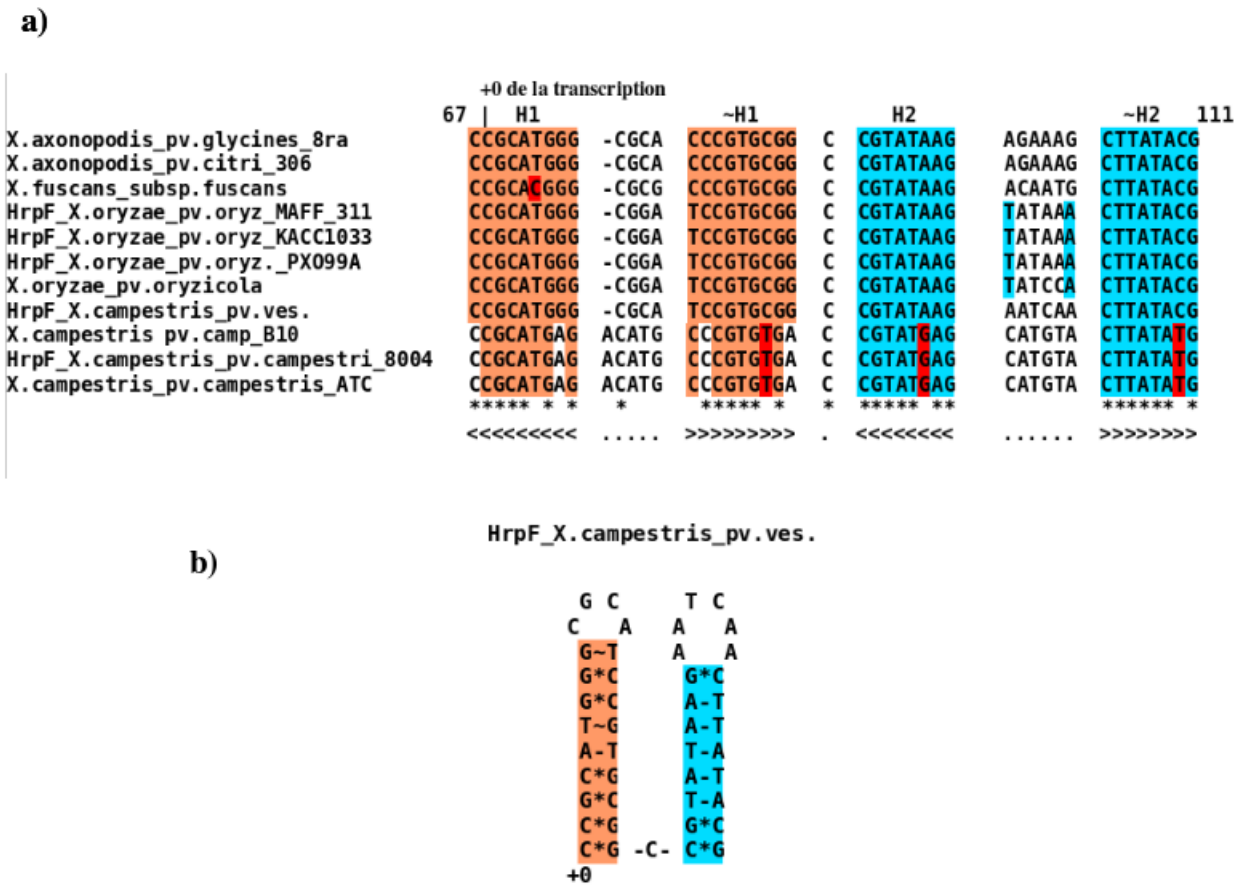


FIG. 10.14 – La structure du motif composé de deux tiges boucles : a) L'alignement multiple des sous-régions de la position 67 à à la position 111 des régions comparées, correspondant au motif identifié. Les régions appartenant à la première tige sont indiqués en orange et celles appartenant à la deuxième sont en bleu. Les positions non appariées ne sont pas colorés. Les sites positions mettant en évidence des mutations compensatoires sont en rouge. Dans la deuxième boucle, les sites colorés en bleu indiquent des appariements supplémentaires potentiels. La conservation à une positions dans l'ensemble de séquneces est indiquée par une * au-dessous de l'alignement à cette position. Au dessous de l'alignement, la structure secondaire consensus est indiquée au format parenthésé. b) La structure secondaire du motif chez *Xanthomonas campestris pv. vesicatoria*, représentative de la plus grande partie des séquences alignées (code couleurs est défini sous a)). Les appariements G et C sont indiqués par une * et les apparements G et T par ~

vation parfaite dans les deux tiges, à une mutation compensatoire près (les sept premières séquences dans l'alignement fig. 10.14a), leur structure secondaire prédite étant en deux tiges-boucles adjacentes 10.14a). Les séquences du deuxième groupe sont identiques entre elles et correspondent aux souches de *X. campestris pv. campestris*. La possibilité d'existence de la première tige-boucle, au sein de ce groupe, est moins probable du fait de présence de deux mésappariements. Paradoxalement, c'est au sein de ce groupe que l'on trouve le plus grande nombre de mutations compensatoires potentielles : une dans la première tige et deux dans la deuxième. Ceci est peut être à rapprocher du fait que la boîte *hrp* est dégénérée chez les souches correspondantes (fig. 10.13).

Même si l'ensemble de ces observations suggère qu'un élément en deux tiges-boucles adjacentes pourrait exister dans les régions *hrpF* et être potentiellement impliqué dans leur régulation, les arguments présentés sont moins forts que dans le cas de la longue tige-boucle située en amont des gènes *popF* de *R. solanacearum*.

Néanmoins, comme le montre figure 10.15, la comparaison de ces motifs identifiés chez Xanthomonas et *R. solanacearum* et de leurs régions adjacentes présentent des similarités importantes : la même distance par rapport à la boîte *hrp* et séquéstration du site d'initiation de la transcription au sein de la structure prédite. S'agissant de gènes orthologues, ceci suggère fortement l'existence d'une ribo-régulation au niveau des séquences promotrices de ces gènes.

10.3.2 Elément CIRCE

Le présent candidat, conservé dans les trois souches de *R. solanacearum* montre une large conservation partielle dans plus de trente génome bactériens de *nr* (une partie représentative de ces conservations est représentée sur la figure 10.16). La zone conservée peut être divisée en deux parties.

La première zone, entre les positions 1 et 50 du candidat (encadrée en noir sur la figure), est conservée dans les trois souches de *R. solanacearum* et présente deux fois dans chacun de ces génomes. Elle est également conservée dans les génomes proches, *Ralstonia piketti 12J* et *Ralstonia eutropha JMP134*.

La deuxième zone présente une conservation plus large. Cette zone contient un motif palindromique de 27nt de long, composé de deux bras conservés, de 9nt de long, capables de former une tige et d'une région variable, de 9nt, séparant les deux bras (les deux

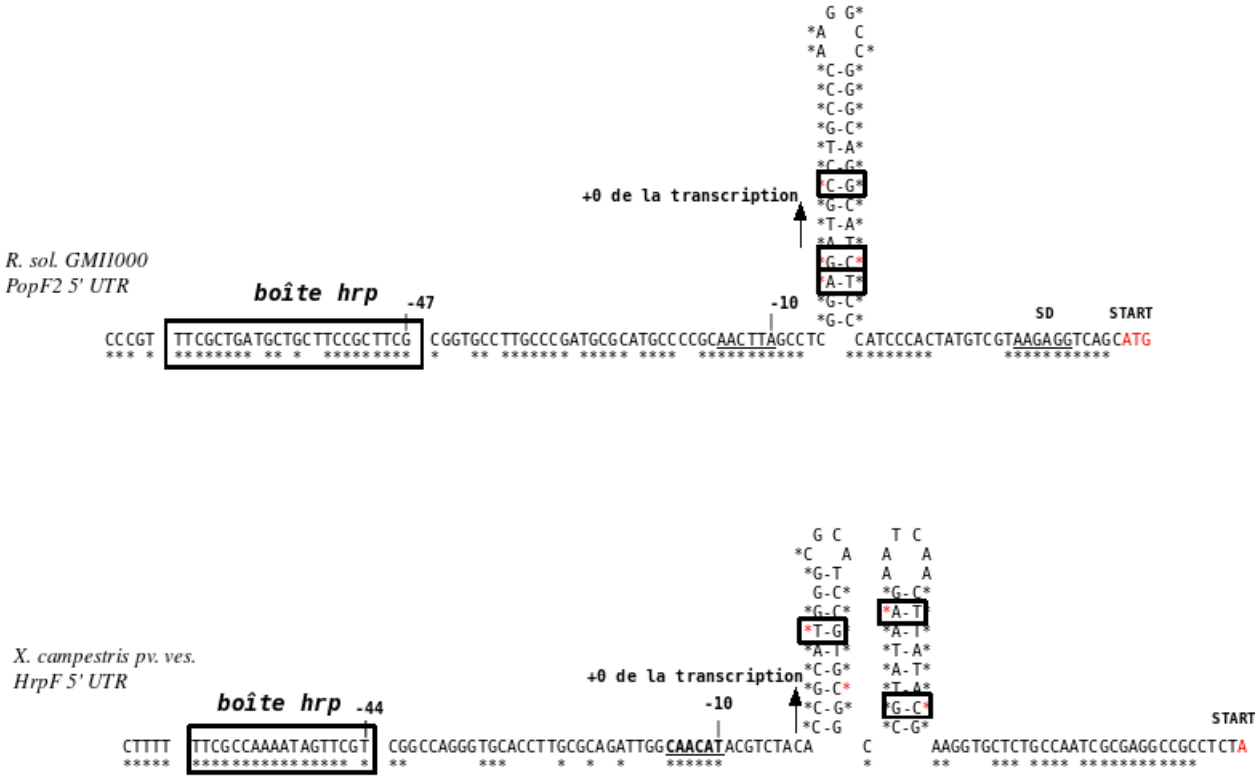


FIG. 10.15 – Les régions d'environ 150nt en amont des gènes *popF1* de *R. solanacearum* et *hrpF* de *X. campestris pv. vesicatoria*. Les positions mettant en évidence des mutations compensatoires dans les structures potentielles en tige-boucle sont encadrés. Le site d'initiation de la transcription (+0 de la transcription désigné par une flèche) est compris dans la structure en tige-boucle. Les sites putatifs de fixation de SZl'ARN polymérase sont soulignés dans la séquence autour des position -10 et -35. La séquence de Shine-Dalgarno potentielle est indiqué par SD dans le cas *R. solanacearum* et le début de la séquence codante est indiqué par START. Les boîte *hrp* pour *R. solanacearum* et la boîte *hrp* imparfaite de *X. campestris pv. vesicatoria* sont encadrés. Dans chacune des séquences, les positions conservées dans les régions homologues ((voir fig. 10.11 et fig. 10.13) sont indiquées par * (dans le cas des mutations compensatoires * est en rouge).

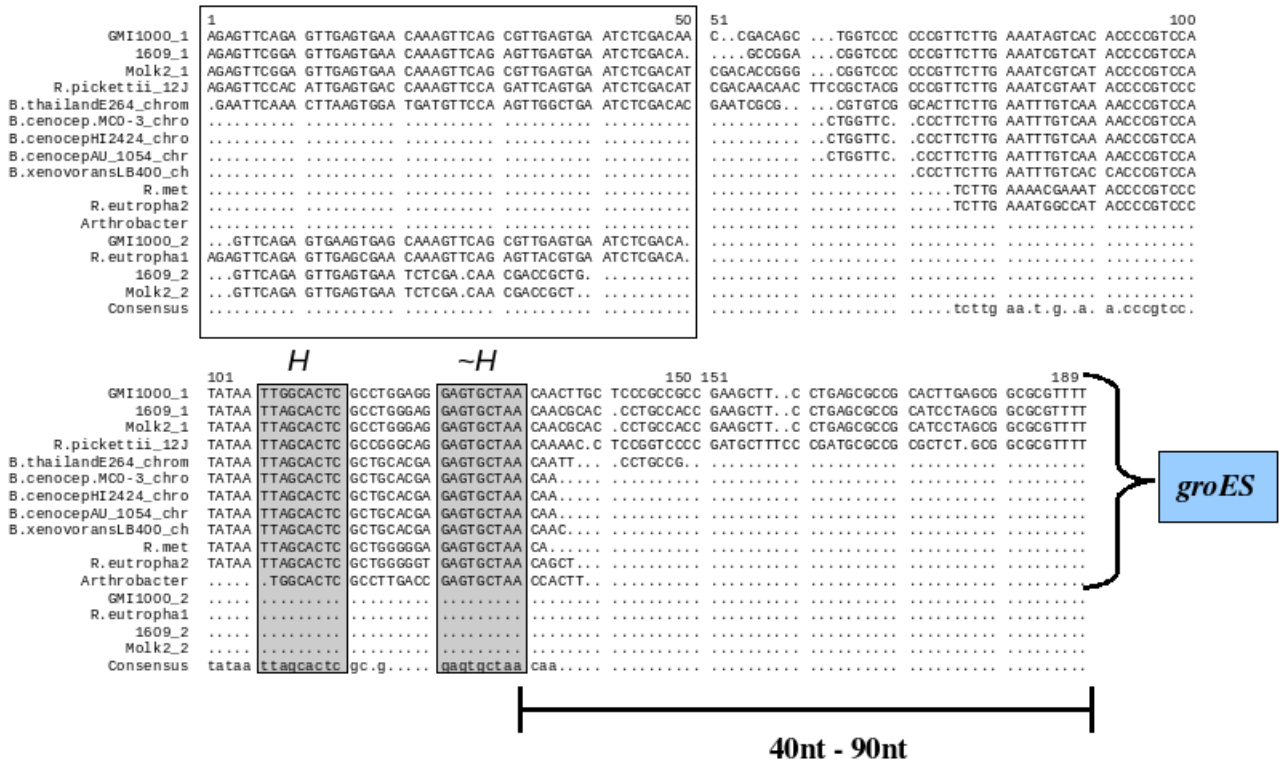


FIG. 10.16 – Alignement multiple du candidat identifié dans *R. solanacearum* et les régions homologues de *nr* (de telles régions ont été identifiées dans plus de trente génomes bactériens, une partie en est présentée dans l'alignement multiple). La région chez *R. solanacearum* GMI1000 (première ligne dans l'alignement multiple) est conservée en deux parties : première partie encadrée en noir, deuxième partie encadrée par deux rectangles gris. La deuxième zone conservée est organisé en palindrome, H et ˜ H, présentant les deux bras du palindrome. De 40 à 90nt en aval de toutes les séquences présentant ce palindrome se trouve le début du gène de la famille *groES*.

bras du palindrome inversé sont désignés par H et ˜ H sur la figure 10.16). Ce motif palindromique peut être décrit sous la forme du motif conensus $TTGGCACTC\text{-}N_9\text{-}GAGTGCTAA$.

Les motifs palindromiques identifiés se trouvent, de façon systématique, dans les régions intergéniques des gènomes correspondants, en amont de l'opéron *groESL* (les gènes *mopB* et *mopA* dans le cas de *R. solanacearum*), impliqués dans l'empêchement de la dénaturation protéique lors d'un choc thermique. Cette observation a suggéré l'existence d'un lien potentiel entre le motif identifié et la régulation des gènes de la famille *groES*.

En effet, un tel lien a été établi entre une structure similaire et le gène *groES* chez *Bacilus subtilis* (Zuber et Schumann (1994)). Chez cette bactérie, le motif palindromique, nommé "CIRCE" (pour "Controlling Inverted Repeat of Chaperone Expression") est impliqué dans la régulation négative de l'expression de l'opéron *groESL* lors du choc thermique. Depuis, l'existence de l'élément CIRCE a été rapportée dans plusieurs génomes bactériens (Mogk *et al.* (1997), Narberhaus (1999)) et la fixation à ce motif d'une protéine, nommée HrcA (pour Heat Regulation at Circe), a été démontréée (Roberts *et al.* (1996)). Le mécanisme de régulation de l'expression de l'opéron *groESL*, nommé

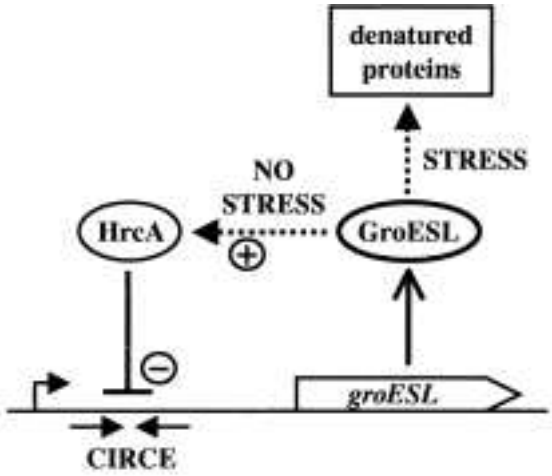


FIG. 10.17 – Le mécanisme de régulation HrcA/CIRCE de l'opéron *groESL* : GroESL agit comme le thérmomètre cellulaire. Les lignes en pointillés présentent les interactions protéine-protéine ; (+) représente l'activation et (-) la répression. La figure est reprise de Narberhaus (1999).

HrcA/CIRCE (Narberhaus (1999), figure 10.17) proposé est le suivant : la protéine HrcA nécessite la protéine GroES pour l'aquisition de sa conformation fonctionnelle. Le niveau basal de la protéine GroES dans la cellule, dans les conditions normales, maintien la protéine HrcA dans sa conformation active qui, à son tour, agit comme répresseur de l'expression de l'opéron *GroESL via* la fixation au motif CIRCE, se trouvant en amont de l'opéron. Suite à un choc thermique, les protéines GroES s'engagent dans le maintien des repliements des protéines dénaturées de la cellule, laissant la protéine HrcA dans sa forme inactive. Le site de fixation CIRCE de la protéine HrcA reste vacant, permettant l'expression continue de l'opéron *groESL*.

Etant donné que le gène *hrcA* existe chez *R. solanacearum*, le motif identifié constitue probablement l'équivalent de l'élément CIRCE dans *R. solanacearum* et est probablement impliqué dans la régulation de l'expression génique en réponse au choc thermique.

En plus de l'occurence de l'élément CIRCE en amont de l'opéron *groESL*, une recherche exhaustive de ce motif dans le génome de *R. solanacearum* a relevé une deuxième occurence, située en amont du gène *rpoH*, codant pour le facteur de transcription σ_{32}. La région en amont du gène *rpoH* est conservée dans plusieurs autres génomes proches (*Ralstonia pickettii*, *Cupriavidus taiwanensis*, *Ralstonia eutropha*, *Burkholderia sp. 383*) et contient l'élément CIRCE.

La présence d'un élément CIRCE a été identifiée en amont du gène *rpoH* de *Geobacter sulfurreducens* (Ueki et Lovley (2007)). Les résultats de cette étude ont suggéré que l'expression de *rpoH* est reprimée par le système HrcA/CIRCE. La présence de l'élément CIRCE en amont du gène *rpoH* de *R. solanacearum* et des génomes proches mentionnés, laisse présager que le même système pourrait contrôler l'expression de *rpoH* dans ces génome.

Discussion

Nous avons identifié un élément constitué d'un palindrome inversé, impliqué probablement dans la régulation génique de la réponse au choc thermique, comme cela a été montré dans d'autres génomes bactériens (tableau 10.5).

Une deuxième occurence du même motif a été identifiée en amont du gène *rpoH*, suggérant que l'expression de ce gène pourrait être également régulée par le même mécanisme.

Notons enfin qu'une autre zone conservée au sein de la même région intergénique (correspondant à l'encadré noir dans la figure 10.16) a été identifiée en deux copies dans

Organisme	Gène ou protéine	Palindrome inversé
B. subtilis	orJ39	TTAGCACTC-N9-GAGTGCTAA
B. subtilis	groESL	TTAGCACTC-N9-GAGTGCTAA
C. acetobutylicum	orfA	TTAGCACTC-N9-GAGTGCTAA
C. acetobutylicum	groESL	TTAGCACTC-N9-GAGTGCTAA
Synechococcus sp.	groESL	TTAGCACTC-N9-GAGTGCTAA
Synechococcus sp.	urf3, urf4	TTAGCACTC-N9-GAGTGCTAA
Synechocystis sp.	cpn-60	TTAGCACTC-N9-GAGTGCTAA
M. tuberculosis	10-kDa AGa	TCAGCACTC-N9-GAGTGCTAc
M. tuberculosis	65-kDa AG	cTtGCACTC-N9-GAGTGCTAA
M. bovis BCG	MPB57	CTAGCACTC-N9-GAGTGCTAg
M. leprae	65-kDa AG	cTgGCACTC-N9-GAGTGCcAg
C. psittaci	hypA, hypB	gTAGCACTt-N9-aAGTGCTAA
R.solanacearum	**mopB**	**TTGGCACTC-N9-GAGTGCTAA**
Consensus		TTAGCACTC-N9-GAGTGCTAA

TAB. 10.5 – Les éléments CIRCE dans différents organismes, emprunté de Wetzstein *et al.* (1992). Complété (en gras) par le palindrome inversé que nous avons identifié dans *R. solanacearum*.

le génome de *R. solanacearum* et des génomes proches. Sa conservation systématique dans les régions intergéniques vides suggère qu'il pourrait également s'agir d'un élément de régulation.

10.4 Discussion

L'analyse des résultats de RNAsim a permis de sélectionner les huit candidats ARNnc les plus prometteurs. Ces candidats seront testés biologiquement et cette vérification constituera la réelle validation de la démarche que nous avons utilisés.

Néanmoins, des éléments de validation ont été apportés par une étude récente (outil SIPHT, Livny *et al.* (2008)), utilisant une approche comparative afin de prédire les ARNnc dans tous les génomes bactériens séquencés à ce jour. Dans *R. solanacearum*, SIPHT prédit 20 candidats sur le chromosome et un sur le plasmide. Trois de ces candidats coïncident avec des candidats que nous avons chois pour validation. Nos

candidats, au regard des candidats SIPHT sont présentés dans le tableau 10.6. En plus des candidats SIPHT, nous y représentons les candidats identifiés par approche *ab initio*. Ceci est fait car la majorité des candidats que nous avons proposés présentent un biais de composition ce qui est également le cas des candidats SIPHT.

En dehors des candidats proposés pour validation, un certain nombre de candidats SIPHT sont présent dans l'ensemble des résultats RNAsim ainsi que dans les résultats HMM. Ceci est représenté sur la figure 10.18.

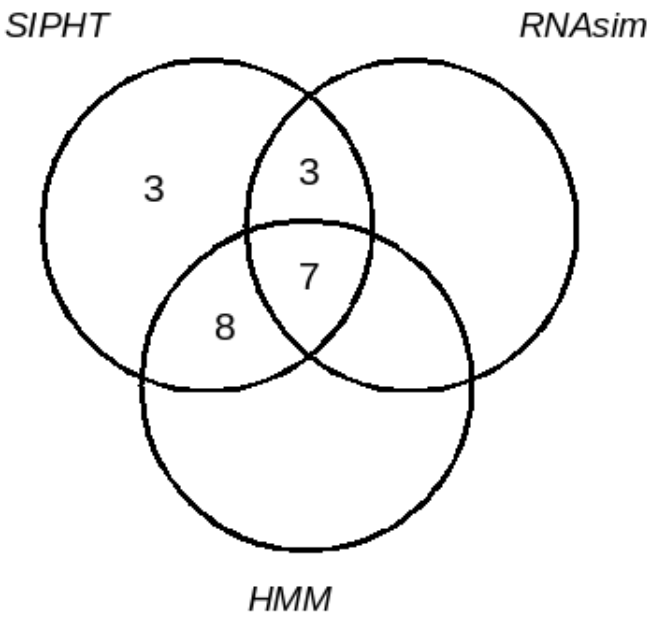


FIG. 10.18 – Diagramme montrant l'intersections entre les prédictions SIPHT (Livny *et al.* (2008)), les prédictions RNAsim et HMM, pour le chromosome et le mégaplasmide de *R. solanacearum*.

Candidat	Rép.	Gènes adj.	Orientation	Extr.5'	Extr.3'	Long.	Long. IGR	SIPHT Extr. 5'	SIPHT Extr. 3'	HMM
Cand 1.1	c	pepV/RSc1303	>< >	1390872	1390777	95	571	1390921	1390794	inclu
Cand 1.2*	c	pepV/RSc1303	>< >	1390699	1390625	74	571			inclu
Cand 2	c	RSc3386/RSc3387	>>>	3651833	3652007	174	914			inclu
Cand 3	p	Rsp1268/RSp1269	>< <	1605747	1605582	165	563			
Cand 4	p	Rsc1022/prhG	>>>	1290424	1290573	149	816			
Cand 5	c	Rsc0973/RSc0974	<><	1021907	1022205	298	517			
Cand 6*	p	Rsp0568/RSp0569	>><	709773	709875	102	338			
Cand 7-cop1	c	Rsc3149/RSc3150	>< >	3400046	3399987	59	1045			
Cand 7-cop2	c	Rsc3358/dbhB	<>>	3617829	3618014	186	642	3617802	3617999	
Cand 7-cop3	c	Rsc3084/RSc3085	<<<	3322751	3322662	89	1069			
Cand 8	c	Rsc0824/PopP1	>>>	866550	866863	313	544	866466	866804	inclu

TAB. 10.6 – Tableau comparatif entre les candidats RNAsim proposés pour validation (section 10.2.2), les prédictions SIPHT (Livny *et al.* (2008)) et HMM.

Cinquième partie

Discussion générale

Chapitre 11

Discussion et perspectives

Chacune des trois parties de la thèse ayant été discutée séparément, ici nous récapitulons les principaux conclusions de ces trois parties, en indiquant des pistes éventuelles à explorer qui ont émergé de ce travail. Parfois les observations et conclusions des différentes parties convergent, dans le même faisceau de preuves, pour appuyer une affirmation commune. Dans cette discussion récapitulative, on essayera de mettre ceci en avant.

Partie *ab initio* Dans un premier temps, nous avons exploré la possibilité de mettre en évidence statistiquement une différence de composition en G+C entre les ARNnc et le reste du génome de *R. solanacearum*, à l'aide des modèles logistiques, une classe de modèles linéaires généralisés. Cette méthodologie a d'abord été appliqué au génome de *Staphylococcus aureus*, permettant de mettre en évidence une différence en composition. Ce travail rentre dans le cadre d'une publication qui sera soumise sous peu.

Le cas de *R. solanacearum* s'étant révélé plus complexe, nous avons utilisé une modélisation markovienne des séquences afin de mettre en évidence une différence de composition entre les ARNnc connus et le reste du génome .

Au cours de cette étude, nous avons également observé que les riboswitch, une classe particulière des ARNnc non-transcrits, situés dans les partie 5' des gènes, présentent une composition différente de l'ensemble des séquences d'ARNnc et que cella là serait plus proche de la composition des régions codantes. Ces résultats suggèreraient qu'une approche par biais de composition ne peut pas être utilisée avec succès pour détecter ces éléments, mais une étude plus globale devrait être faite pour étendre cette affirmation

à l'ensemble des génomes séquencés.

L'utilisation des méthodes de segmentation de génome sur la base de la composition (HMM) a permis de générer un grand nombre de candidats dans le génome de *R. solanacearum*. Néanmoins, cette recherche n'est pas assez spécifique, générant beaucoup de zones incluses dans les régions codantes. Une piste serait de modéliser plus précisément les régions codantes pour rendre le modèle plus prformant et une autre serait d'exclure les régions ACUR, ayant une composition variable et pouvant fausser l'estimation des paramètres des modèles utilisés.

Au cours de ce travail, en vue de la segmentation du génome, nous avons étudié les propriétés de l'algorithme de Viterbi, en concluant que son utilisation n'est pas appropriée lorsque les états du modèle ne sont pas suffisament "éloignés" en termes de la composition.

Enfin, la question essentielle posée par cette étude est : les ARNnc dans les génomes A+T riches étant plutôt de composition en G+C plus élevée, la tendance serait-elle inversée dans les génomes G+C riches ? Est-ce que ceci dépend de l'espèce bactérienne ou du groupe taxonomique ? Pour répondre à ces questions, dont la réponse ici est confinée au génome de *R. solanacearum*, une analyse systématique dans les génomes G+C riches devrait être faite. Une réponse permettrait une avancée dans les méthodes de recherche d'ARNnc *ab initio*.

RNAsim : une approche comparative pour la détection des ARNnc Au cours de la thèse, les séquences de deux nouvelles souches de *R. solanacearum* sont devenues disponibles. Ceci a constitué une occasion privilégiée pour l'utilisation d'une approche comparative de détection des ARNnc dans cet organisme, la conservation des ARNnc étant souvent confinée au sein de la même espèce bactérienne.

En conséquence, nous avons repris RNAsim, un outil existant pour la détection des ARNnc dans n génomes, basé sur la recherche des composantes connexes d'un graphe, les composantes connexes étant construites à partir des réseaux d'alignements issus de comparaison deux-à-deux de n génomes. Nous l'avons amélioré, notamment d'un point de vue de la construction des composantes connexes et de la fusion des données permettant une simplification de l'analyse des résultats, tout en rajoutant les fonctionalités pertinentes pour l'analyse des candidats ARNnc, telles que l'analyse de la conservation des gènes adjacents et l'alignement multiples des élements des composantes

connexes.

Analyse des candidats RNAsim L'application de RNAsim sur les génomes de trois souches de *R. solanacearum* a donné lieu à environ 1000 candidats ARNnc dont un nombre considérable a pu être écarté à l'aide des outils associés à RNAsim, permettant de détecter efficacement les candidats correspondants aux séquences d'insertion ou des terminateurs de transcription conservés.

Les candidats restant nécessitaient une expertise manuelle afin de décider de leur "potentialité" en tant que candidat ARNnc. Cette "potentialité"a été mesurée à l'aide des critères que nous avons définis, ces critères étant liés aux propriétés bioinformatiques des ARNnc, tels que la conservation partielle d'une IGR ou l'existence d'un noyau de conservation ou encore la conservation d'un terminateur de transcription rho-indépendant prédit. Nous avons également observé que la composition en G+C des régions candidates permet de mieux discriminer les candidats fiables, résultant dans l'utilisation d'un critère basé sur la sélection des candidats au G+C% plus bas.

Au terme de l'analyse effectuée sur environ 300 candidats, une première sélection a été discuté avec nos partenaires biologistes permettant de choisir les huit candidats les plus prometteurs en tant que ARNnc mais aussi d'un point de vue de leur possible implication dans la pathogénie.

En plus des candidats ARNnc, plusieurs séquences potentiellement régulatrices possédant une structure secondaire ont pu être mises en évidence. Un exemple d'une telle structure est une tige boucle particulièrement longue, conservée entre les souches de *R. solanacearum*, se trouvant dans les régions promotrices des gènes *PopF*, les effecteurs de pathogénie de la bactérie.

Les candidats proposés seront testés expérimentalement et c'est seulement la validation biologique qui donnera une validation finale de notre approche. Néanmoins, des éléments de validation, renforçant le poids des preuves de nos candidats, sont arrivés au cours de la rédaction de la thèse. Premierement, la base de données Rfam, dans sa nouvelle mise a jour, contient la prédiction dans *R. solanacearum* d'un ARNnc putatif *suhB* figurant à la tête de liste de nos candidats. Deuxièment, une étude récente, basée sur la conservation des IGR dans l'ensemble des génomes procaryotes séquencés, présente une liste de candidats d'ARNnc dans *R. solanacearum*, entre autres. Sur 20 candidats proposés, trois sont présents dans la liste de nos meilleurs candidats tandis

que cinq autres sont présents dans les candidats RNAsim qui n'ont pas été proposés pour validation. Un des candidats communs, proposé pour validation, représente une zone conservée se trouvant en proximité des gènes impliqués dans la pathogénie de *R. solanacearum*.

Approche ab initio et approche comparative De façon intéressante, les candidats issus de l'étude mentionnée, présentent, pour la majorité, une composition en G+C plus basse que le reste du génome. Il en va de même pour les candidats que nous avons proposés pour validation et ceci coïncide avec l'observation que nous avons pu faire au cours de l'analyse, à savoir que la conservation partielle des zones intergéniques est souvent accompagnée d'un biais de composition dans ces zones. Or, l'utilisation des HMM sur le génome de *R. solanacearum* a démontré que l'utilisation du biais de composition dans ce génome n'est pas suffisante pour discriminer les ARNnc du reste du génome. En revanche, un couplage de deux méthodes donnant lieu à une recherche des zones conservées présentant un biais en composition pourrait constituer une piste intéressante pour augmenter la spécificité des deux approches qui a, dans les deux cas, été faible.

Index

Bibliographie

Aldon, D., Brito, B., Boucher, C. et Génin, S.A bacterial sensor of plant cell contact controls the transcriptional induction of Ralstonia solanacearum pathogenicity genes. t. *The EMBO Journal*, 19(10):2304–14. PMID : 10811621.

Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E. G., Margalit, H. et Altuvia, S.Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. t. *Current Biology : CB*, 11(12):941–50. PMID : 11448770.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. et Sherlock, G.Gene ontology : tool for the unification of biology. The Gene Ontology Consortium. t. *Nature Genetics*, 25(1):25–9. PMID : 10802651.

Atallah, M. J. (1998). *Complexity Classes*, pages 6.4–6.5. CRC Press.

Axmann, I. M., Kensche, P., Vogel, J., Kohl, S., Herzel, H. et Hess, W. R.Identification of cyanobacterial non-coding RNAs by comparative genome analysis. t. *Genome Biology*, 6(9):R73. PMID : 16168080.

Azad, R. K. et Borodovsky, M.Probabilistic methods of identifying genes in prokaryotic genomes : connections to the HMM theory. t. *Briefings in Bioinformatics*, 5(2):118–30. PMID : 15260893.

Babitzke, P. et Romeo, T.CsrB sRNA family : sequestration of RNA-binding regulatory proteins. t. *Current Opinion in Microbiology*, 10(2):156–63. PMID : 17383221.

Bateman, A., Birney, E., Durbin, R., Eddy, S., Finn, R. et Sonnhammer, E.Pfam 3.1 : 1313 multiple alignments and profile HMMs match the majority of proteins. t. *Nucl. Acids Res.*, 27(1):260–262.

Bindewald, E. et Shapiro, B. A.RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. t. *RNA*, 12(3):342–352. PMC1383574.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D. et Miller, W.Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. t. *Genome Research*, 14(4):708–715. PMC383317.

Boisset, S., Geissmann, T., Huntzinger, E., Fechter, P., Bendridi, N., Possedko, M., Chevalier, C., Helfer, A. C., Benito, Y., Jacquier, A., Gaspin, C., Vandenesch, F. et Romby, P.Staphylococcus aureus RNAIII coordinately represses the synthesis of virulence factors and the transcription regulator Rot by an antisense mechanism. t. *Genes & Development*, 21(11):1353–1366. PMC1877748.

Borer, P. N., Dengler, B., Tinoco, I. et Uhlenbeck, O. C.Stability of ribonucleic acid double-stranded helices. t. *Journal of Molecular Biology*, 86(4):843–53. PMID : 4427357.

Bouyer, J., HÉMON, D. et CORDIER, S. (1995). *Épidémiologie, Principes et méthodes quantitatives*. Inserm.

Brors, B.Microarray annotation and biological information on function. t. *Methods of Information in Medicine*, 44(3):468–72. PMID : 16113775.

Brown, D. G. (2008). *Bioinformatics Algorithms : Techniques and Applications*, chapitre A survey of seeding for sequence alignment, pages 126–152. Wiley-Interscience (I. Mandoiu, A. Zelikovsky).

Burge, C. et Karlin, S.Prediction of complete gene structures in human genomic DNA. t. *Journal of Molecular Biology*, 268(1):78–94. PMID : 9149143.

Burge, C. et Karlin, S.Prediction of complete gene structures in human genomic DNA. t. *J Mol Biol*, 268(1):78–94.

Caldwell, A., Kahng, A. et Markov, I.Improved Algorithms for Hypergraph Bipartitioning. t. *In ASPDAC '00*, pages 661–666. ACM/IEEE.

Caliebe, A. et Roesler, U.Convergence of the maximum a posteriori path estimator in hidden Markov model. t. *IEEE Transactions on information theory*, 45(7):1750–1759.

Carter, R. J., Dubchak, I. et Holbrook, S. R.A computational approach to identify genes for functional RNAs in genomic sequences. t. *Nucleic Acids Research*, 29(19): 3928–38. PMID : 11574674.

Cech, T. R. et Atkins, J. F. (2005). *The Rna World (Cold Spring Harbor Monograph Series) (Cold Spring Harbor Monograph Series)*. Cold Spring Harbor Laboratory Press.

Chargaff, E.Structure and function of nucleic acids as cell constituents. t. *Fed. Proc.*, 10:654–659.

Chen, S., Lesnik, E. A., Hall, T. A., Sampath, R., Griffey, R. H., Ecker, D. J. et Blyn, L. B.A bioinformatics based approach to discover small RNA genes in the Escherichia coli genome. t. *Bio Systems*, 65(2-3):157–77. PMID : 12069726.

Clote, P., Ferre, F., Kranakis, E. et Krizanc, D.Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. t. *RNA*, 11(5):578–591.

Cochrane, J. C. et Strobel, S. A.Riboswitch effectors as protein enzyme cofactors. t. *RNA*, 14(6):993–1002. PMC2390802.

Coenye, T., Drevinek, P., Mahenthiralingam, E., Shah, S. A., Gill, R. T., Vandamme, P. et Ussery, D. W.Identification of putative noncoding RNA genes in the Burkholderia cenocepacia J2315 genome. t. *FEMS Microbiology Letters*, 276(1): 83–92. PMID : 17937666.

Collett, D. (1991). *Modelling Binary Data*. Chapman & Hall/CRC.

Coppins, R. L., Hall, K. B. et Groisman, E. A.The intricate world of riboswitches. t. *Current opinion in microbiology*, 10(2):176–181. PMC1894890.

Corbino, K. A., Barrick, J. E., Lim, J., Welz, R., Tucker, B. J., Puskarz, I., Mandal, M., Rudnick, N. D. et Breaker, R. R.Evidence for a second class of S-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. t. *Genome Biology*, 6(8):R70. PMC1273637.

Corpet, F.Multiple sequence alignment with hierarchical clustering. t. *Nucleic Acids Research*, 16(22):10881–90. PMID : 2849754.

Coventry, A., Kleitman, D. J. et Berger, B.MSARi : Multiple sequence alignments for statistical detection of RNA secondary structure. t. *Proceedings of the National Academy of Sciences*, 101(33):12102–12107.

Cozzuto, L., Petrillo, M., Silvestro, G., Nocera, P. D. et Paolella, G.Systematic identification of stem-loop containing sequence families in bacterial genomes. t. *BMC Genomics*, 9(1):20.

Cros, M., Sallet, E., Moisan, A., Cierco-Ayrolles, C. et Gaspin, C.Visualizing and exploring genomic information for non-protein-coding RNA identification using ApolloRNA. t.

Cunnac, S. (2004). *Identification à l'échelle génomique des effecteurs dépendant du système de sécretion de type III de la bactérie phytopathogène* Ralstonia solanacearum. Thèse de doctorat, Université Paul Sabatier.

Cunnac, S., Boucher, C. et Genin, S.Characterization of the cis-Acting Regulatory Element Controlling HrpB-Mediated Activation of the Type III Secretion System and Effector Genes in Ralstonia solanacearum. t. *Journal of Bacteriology*, 186(8): 2309–2318. PMC412162.

Das, S., Paul, S., Bag, S. K. et Dutta, C.Analysis of Nanoarchaeum equitans genome and proteome composition : indications for hyperthermophilic and parasitic adaptation. t. *BMC Genomics*, 7:186. PMC1574309.

del Val, C., Rivas, E., Torres-Quesada, O., Toro, N. et Jiménez-Zurdo, J. I.Identification of differentially expressed small non-coding RNAs in the legume endosymbiont Sinorhizobium meliloti by comparative genomics. t. *Molecular Microbiology*, 66(5):1080–91. PMID : 17971083.

Delcher, A. L., Harmon, D., Kasif, S., White, O. et Salzberg, S. L.Improved microbial gene identification with GLIMMER. t. *Nucleic Acids Res*, 27(23):4636–4641.

Dempster, A. P., Laird, N. M. et Rubin, D. B.Maximum Likelihood from Incomplete Data via the EM Algorithm. t. *Journal of the Royal Statistical Society. Series B (Methodological)*.

di Bernardo, D., Down, T. et Hubbard, T.ddbRNA : detection of conserved secondary structures in multiple alignments. t. *Bioinformatics*, 19(13):1606–1611.

Diestel, R. (1997). *Graph Theory*. Springer-Verlag.

Doshi, K., Cannone, J., Cobaugh, C. et Gutell, R.Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. t. *BMC Bioinformatics*, 5(1):105.

Durbin, R., Eddy, S., Krogh, A. et Mitchison, G. (1998a). *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Durbin, R., Eddy, S. R., Krogh, A. et Mitchison, G. (1998b). *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Eddy, S. R.RNABOB : Fast Pattern searching for RNA secondary structures. t.

Ermolaeva, M. D., Khalak, H. G., White, O., Smith, H. O. et Salzberg, S. L.Prediction of transcription terminators in bacterial genomes. t. *Journal of Molecular Biology*, 301(1):27–33. PMID : 10926490.

Fegan, M. et Prior, P. (2005). *Bacterial wilt disease and the Ralstonia solanacearum species complex.*, chapitre How complex is the Ralstonia solanacearum species complex. APS Press, Madison, WI.

Flavier, A. B., Ganova-Raeva, L. M., Schell, M. A. et Denny, T. P.Hierarchical autoinduction in Ralstonia solanacearum : control of acyl-homoserine lactone production by a novel autoregulatory system responsive to 3-hydroxypalmitic acid methyl ester. t. *Journal of Bacteriology*, 179(22):7089–97. PMID : 9371457.

Forsdyke, D. R. et Mortimer, J. R.Chargaff's legacy. t. *Gene*, 261(1):127 – 137.

Frank, D. N. et Pace, N. R.Ribonuclease P : unity and diversity in a tRNA processing ribozyme. t. *Annual Review of Biochemistry*, 67:153–80. PMID : 9759486.

Galtier, N. et Lobry, J. R.Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. t. *Journal of Molecular Evolution*, 44(6):632–6. PMID : 9169555.

Gautheret, D., Major, F. et Cedergren, R.Modeling the three-dimensional structure of RNA using discrete nucleotide conformational sets. t. *Journal of Molecular Biology*, 229(4):1049–64. PMID : 7680379.

Genin, S. et Boucher, C.Lessons learned from the genome analysis of ralstonia solanacearum. t. *Annual Review of Phytopathology*, 42:107–34. PMID : 15283662.

Gerdes, K., Gultyaev, A. P., Franch, T., Pedersen, K. et Mikkelsen, N. D.Antisense RNA-regulated programmed cell death. t. *Annual Review of Genetics*, 31:1–31. PMID : 9442888.

Gish, W.WU-BLAST 2.0. t.

Goertzen, L. R., Cannone, J. J., Gutell, R. R. et Jansen, R. K.ITS secondary structure derived from comparative analysis : implications for sequence alignment and phylogeny of the Asteraceae. t. *Molecular Phylogenetics and Evolution*, 29(2):216–234.

Gottesman, S.The small RNA regulators of Escherichia coli : roles and mechanisms*. t. *Annual Review of Microbiology*, 58:303–28. PMID : 15487940.

Gottesman, S.Micros for microbes : non-coding regulatory RNAs in bacteria. t. *Trends in Genetics*, 21(7):399–404.

Grant, S. R., Fisher, E. J., Chang, J. H., Mole, B. M. et Dangl, J. L.Subterfuge and manipulation : type III effector proteins of phytopathogenic bacteria. t. *Annual Review of Microbiology*, 60:425–49. PMID : 16753033.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. et Eddy, S. R.Rfam : an RNA family database. t. *Nucleic Acids Research*, 31(1):439–441. PMC165453.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R. et Bateman, A.Rfam : annotating non-coding RNAs in complete genomes. t. *Nucleic Acids Research*, 33(Database issue):D121–4. PMID : 15608160.

Grissa, I., Vergnaud, G. et Pourcel, C.The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. t. *BMC Bioinformatics*.

Gruber, A. R., Neubock, R., Hofacker, I. L. et Washietl, S.The RNAz web server : prediction of thermodynamically stable and evolutionarily conserved RNA structures. t. *Nucl. Acids Res.*, page gkm222.

Guidot, A., Prior, P., Schoenfeld, J., Carrère, S., Genin, S. et Boucher, C.Genomic Structure and Phylogeny of the Plant Pathogen Ralstonia solanacearum Inferred from Gene Distribution Analysis. t. *Journal of Bacteriology*, 189(2):377–387. PMC1797399.

Hershberg, R., Altuvia, S. et Margalit, H.A survey of small RNA-encoding genes in Escherichia coli. t. *Nucl. Acids Res.*, 31(7):1813–1820.

Hofacker, I. L.Vienna RNA secondary structure server. t. *Nucleic Acids Research*, 31(13):3429–3431. PMC169005.

Huang, X. et Madan, A.CAP3 : A DNA sequence assembly program. t. *Genome Research*, 9(9):868–77. PMID : 10508846.

Huber, W., Carey, V. J., Long, L., Falcon, S. et Gentleman, R.Graphs in molecular biology. t. *BMC Bioinformatics*, 8(Suppl 6):S8. PMC1995545.

Hüttenhofer, A., Schattner, P. et Polacek, N.Non-coding RNAs : hope or hype ? t. *Trends in Genetics : TIG*, 21(5):289–97. PMID : 15851066.

Jacques, J.-F., ois, Jang, S., Pré, vost, K., Desnoyers, G., Desmarais, M., Imlay, J., Massé et EricRyhB small RNA modulates the free intracellular iron pool and is essential for normal growth during iron limitation in Escherichia coli. t. *Molecular Microbiology*, 62:1181–1190.

Johansen, J., Rasmussen, A. A., Overgaard, M. et Valentin-Hansen, P.Conserved small non-coding RNAs that belong to the sigmaE regulon : role in down-regulation of outer membrane proteins. t. *Journal of Molecular Biology*, 364(1):1–8. PMID : 17007876.

Karlin, S., Campbell, A. M. et Mrázek, J.Comparative DNA analysis across diverse genomes. t. *Annual Review of Genetics*, 32:185–225. PMID : 9928479.

Karlin, S., Ladunga, I. et Blaisdell, B. E.Heterogeneity of genomes : measures and values. t. *Proceedings of the National Academy of Sciences of the United States of America*, 91(26):12837–12841. PMC45535.

Kim, J. et Breaker, R.Purine sensing by riboswitches. t.

Klein, R. J. et Eddy, S. R.RSEARCH : Finding homologs of single structured RNA sequences. t. *BMC Bioinformatics*, 4:44.

Klein, R. J., Misulovin, Z. et Eddy, S. R.Noncoding RNA genes identified in AT-rich hyperthermophiles. t. *Proceedings of the National Academy of Sciences of the United States of America*, 99(11):7542–7. PMID : 12032319.

Koebnik, R., Krüger, A., Thieme, F., Urban, A. et Bonas, U.Specific Binding of the Xanthomonas campestris pv. vesicatoria AraC-Type Transcriptional Activator HrpX to Plant-Inducible Promoter Boxes. t. *Journal of Bacteriology*, 188(21):7652–7660. PMC1636286.

Kogan, J. A.Optimal segmentation of structural experimental curves by the DP method. t. *Automation and Remote Control*, 7(2):934–942.

Krogh, A., Larsson, B., von Heijne, G. et Sonnhammer, E. L.Predicting transmembrane protein topology with a hidden Markov model : application to complete genomes. t. *Journal of Molecular Biology*, 305(3):567–80. PMID : 11152613.

Krogh, A., Mian, I. S. et Haussler, D.A hidden Markov model that finds genes in E. coli DNA. t. *Nucleic Acids Research*, 22(22):4768–78. PMID : 7984429.

Kulkarni, P. R., Cui, X., Williams, J. W., Stevens, A. M. et Kulkarni, R. V.Prediction of CsrA-regulating small RNAs in bacteria and their experimental verification in Vibrio fischeri. t. *Nucleic Acids Research*, 34(11):3361–9. PMID : 16822857.

Lambert, A., Fontaine, J.-F., Legendre, M., Leclerc, F., Permal, E., Major, F., Putzer, H., Delfour, O., Michot, B. et Gautheret, D.The ERPIN server :

an interface to profile-based RNA motif identification. t. *Nucleic Acids Research*, 32(Web Server issue):W160–5. PMID : 15215371.

Le, S. Y., Nussinov, R. et Maizel, J. V.Tree graphs of RNA secondary structures and their comparisons. t. *Computers and Biomedical Research, an International Journal*, 22(5):461–73. PMID : 2776449.

Lease, R. A. et Belfort, M.Riboregulation by DsrA RNA : trans-actions for global economy. t. *Molecular Microbiology*, 38(4):667–72. PMID : 11115103.

Lenz, D. H., Mok, K. C., Lilley, B. N., Kulkarni, R. V., Wingreen, N. S. et Bassler, B. L.The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in Vibrio harveyi and Vibrio cholerae. t. *Cell*, 118(1):69–82. PMID : 15242645.

Lewis, S. E., Searle, S. M. J., Harris, N., Gibson, M., Lyer, V., Richter, J., Wiel, C., Bayraktaroglir, L., Birney, E., Crosby, M. A., Kaminker, J. S., Matthews, B. B., Prochnik, S. E., Smithy, C. D., Tupy, J. L., Rubin, G. M., Misra, S., Mungall, C. J. et Clamp, M. E.Apollo : a sequence annotation editor. t. *Genome Biology*, 3(12):RESEARCH0082. PMID : 12537571.

Li, W., Stolovitzky, G., Bernaola-galván, P. et Oliver, J. L.Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. t. *Genome Research*, 9:916–928.

Liu, Y., Cui, Y., Mukherjee, A. et Chatterjee, A. K.Characterization of a novel RNA regulator of Erwinia carotovora ssp. carotovora that controls production of extracellular enzymes and secondary metabolites. t. *Molecular Microbiology*, 29(1): 219–34. PMID : 9701816.

Livny, J., Brencic, A., Lory, S. et Waldor, M. K.Identification of 17 Pseudomonas aeruginosa sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2. t. *Nucl. Acids Res.*, 34(12):3484–3493.

Livny, J., Fogel, M. A., Davis, B. M. et Waldor, M. K.sRNAPredict : an integrative computational approach to identify sRNAs in bacterial genomes. t. *Nucl. Acids Res.*, 33(13):4096–4105.

Livny, J., Teonadi, H., Livny, M. et Waldor, M. K.High-Throughput, Kingdom-Wide Prediction and Annotation of Bacterial Non-Coding RNAs. t. *PLoS ONE*, 3(9):e3197. PMC2527527.

Livny, J. et Waldor, M. K.Identification of small RNAs in diverse bacterial species. t. *Current Opinion in Microbiology*, 10(2):96–101. PMID : 17383222.

Lobry, J. R.Asymmetric substitution patterns in the two DNA strands of bacteria. t. *Molecular Biology and Evolution*, 13(5):660–5. PMID : 8676740.

Lowe, T. M. et Eddy, S. R.tRNAscan-SE : a program for improved detection of transfer RNA genes in genomic sequence. t. *Nucleic Acids Research*, 25(5):955–64. PMID : 9023104.

Maccario, J. (1998). *Modélisation*, chapitre Modèles logistiques. SVS.

Macke, T. J., Ecker, D. J., Gutell, R. R., Gautheret, D., Case, D. A. et Sampath, R.RNAMotif, an RNA secondary structure definition and search algorithm. t. *Nucleic Acids Research*, 29(22):4724–35. PMID : 11713323.

Massé, E., Escorcia, F. E. et Gottesman, S.Coupled degradation of a small regulatory RNA and its mRNA targets in Escherichia coli. t. *Genes & Development*, 17(19):2374–83. PMID : 12975324.

Massé, E. et Gottesman, S.A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. t. *Proceedings of the National Academy of Sciences of the United States of America*, 99(7):4620–5. PMID : 11917098.

Masse, E., Vanderpool, C. K. et Gottesman, S.Effect of RyhB Small RNA on Global Iron Use in Escherichia coli. t. *J. Bacteriol.*, 187(20):6962–6971.

Massé, E., Salvail, H., Desnoyers, G. et Arguin, M.Small RNAs controlling iron metabolism. t. *Current Opinion in Microbiology*, 10(2):140–5. PMID : 17383226.

McCullagh, P. et Nelder, J. A. (1989a). *Generalized Linear Models, Second Edition*, chapitre 4.5.2. Chapman & Hall.

McCullagh, P. et Nelder, J. A. (1989b). *Generalized Linear Models, Second Edition*, chapitre 12.7.1. Chapman & Hall.

McCullagh, P. et Nelder, J. A. (1989c). *Generalized Linear Models, Second Edition*, chapitre Appendix A. Chapman & Hall.

McCutcheon, J. P. et Eddy, S. R.Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. t. *Nucleic Acids Res*, 31(14): 4119–4128.

McInerney, J. O.Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. t. *Proceedings of the National Academy of Sciences of the United States of America*, 95(18):10698–703. PMID : 9724767.

Meister, G. et Tuschl, T.Mechanisms of gene silencing by double-stranded RNA. t. *Nature*, 431(7006):343–9. PMID : 15372041.

Merhav, N. et Ephraim, Y.Maximum likelihood hidden markov modeling using a dominant sequence of states. t. *IEEE Transactions on signal processing*, 39(9).

Meyer, D., Cunnac, S., Guéneron, M., Declercq, C., Gijsegem, F. V., Lauber, E., Boucher, C. et Arlat, M.PopF1 and PopF2, two proteins secreted by the type III protein secretion system of Ralstonia solanacearum, are translocators belonging to the HrpF/NopX family. t. *Journal of Bacteriology*, 188(13):4903–17. PMID : 16788199.

Mogk, A., Homuth, G., Scholz, C., Kim, L., Schmid, F. X. et Schumann, W.The GroE chaperonin machine is a major modulator of the CIRCE heat shock regulon of Bacillus subtilis. t. *The EMBO Journal*, 16(15):4579–4590. PMC1170084.

Mole, B. M., Baltrus, D. A., Dangl, J. L. et Grant, S. R.Global virulence regulation networks in phytopathogenic bacteria. t. *Trends in Microbiology*, 15(8):363–71. PMID : 17627825.

Moulton, V.Tracking down noncoding RNAs. t. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2269–2270.

Narberhaus, F.Negative regulation of bacterial heat shock genes. t. *Molecular Microbiology*, 31(1):1–8.

Newman, M. E. J.The structure and function of complex networks. t. *SIAM Review*, 45:167.

Nicolas, P. (2003). *Mise au point et utilisation de modèles de chaînes de Markov cachées pour l'étude des séquences d'ADN*. Thèse de doctorat.

Nicolas, P., Tocquet, A.-S. et be, F. M.-M.SHOW User Manual. t.

Noirot, C.Rapport de stage : RNAsim, une approche comparative pour la prédiction des ARNnc. t.

Occhialini, A., Cunnac, S., Reymond, N., Genin, S. et Boucher, C.Genome-wide analysis of gene expression in Ralstonia solanacearum reveals that the hrpB gene acts as a regulatory switch controlling multiple virulence pathways. t. *Molecular Plant-Microbe Interactions : MPMI*, 18(9):938–49. PMID : 16167764.

Ohler, U., Niemann, H., Gc, L. et Rubin, G. M.Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. t. *Bioinformatics (Oxford, England)*, 17 Suppl 1:S199–206. PMID : 11473010.

Ostberg, Y., Bunikis, I., Bergstrom, S. et Johansson, J.The Etiological Agent of Lyme Disease, Borrelia burgdorferi, Appears To Contain Only a Few Small RNA Molecules. t. *J. Bacteriol.*, 186(24):8472–8477.

Pánek, J., Bobek, J., Mikulík, K., Basler, M. et Vohradský, J.Biocomputational prediction of small non-coding RNAs in Streptomyces. t. *BMC Genomics*, 9:217. PMC2422843.

Pichon, C. et Felden, B.From the Cover : Small RNA genes expressed from Staphylococcus aureus genomic and pathogenicity islands with specific expression among pathogenic strains. t. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14249–14254. PMC1242290.

Poueymiro, M. et Genin, S.Secreted proteins from Ralstonia solanacearum : a hundred tricks to kill a plant. t. *Current Opinion in Microbiology*, 12(1):44–52. PMID : 19144559.

Pré, vost, K., Salvail, H., Desnoyers, G., Jacques, J.-F., ois, Phaneuf, E., milie, Massé et EricThe small RNA RyhB activates the translation of shiA mRNA encoding a permease of shikimate, a compound involved in siderophore synthesis. t. *Molecular Microbiology*, 64:1260–1273.

Pregibon, D.Logistic Regression Diagnostics. t. *The Annals of Statistics*.

R Development Core Team (2008). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rabiner, L.A tutorial on hidden Markov models and selected applications in speech recognition. t. *Proceedings of the IEEE*, 77(2):257–286.

Ray, A. K., Ghosh, A. K. et Majumdar, A. K.Patscan—a microprocessor-based pattern scanner system. t. *J. Microcomput. Appl.*, 10(1):71–82.

Reeder, J., Reeder, J. et Giegerich, R.Locomotif : from graphical motif description to RNA motif search. t. *Bioinformatics*, 23(13):i392–400.

Reka, A. et Albert-Laszlo, B.Statistical mechanics of complex networks. t. *Reviews of Modern Physics*, 74:47.

Rivas, E. et Eddy, S. R.Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. t. *Bioinformatics (Oxford, England)*, 16(7):583–605. PMID : 11038329.

Rivas, E. et Eddy, S. R.Noncoding RNA gene detection using comparative sequence analysis. t. *BMC Bioinformatics*, 2:8. PMID : 11801179.

Rivas, E., Klein, R. J., Jones, T. A. et Eddy, S. R.Computational identification of noncoding RNAs in E. coli by comparative genomics. t. *Current Biology : CB*, 11(17):1369–73. PMID : 11553332.

Roberts, R. C., Toochinda, C., Avedissian, M., Baldini, R. L., Gomes, S. L. et Shapiro, L.Identification of a Caulobacter crescentus operon encoding hrcA, involved in negatively regulating heat-inducible transcription, and the chaperone gene grpE. t. *Journal of Bacteriology*, 178(7):1829–1841. PMC177876.

Robin, S., Rodolphe, F. et Schbath, S. (2003). *ADN, mots et modèles*, chapitre 5, page 76. Belin.

Rocha, E. P. C., Danchin, A. et Viari, A.Universal replication biases in bacteria. t. *Molecular Microbiology*, 32(1):11–16.

Romby, P., Vandenesch, F. et Wagner, E. G. H.The role of RNAs in the regulation of virulence-gene expression. t. *Current Opinion in Microbiology*, 9(2):229–36. PMID : 16529986.

Rudner, R., Karkas, J. D. et Chargaff, E.Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. t. *Proceedings of the National Academy of Sciences of the United States of America*, 60(3):921–922.

Salanoubat, M., Genin, S., Artiguenave, F., Gouzy, J., Mangenot, S., Arlat, M., Billault, A., Brottier, P., Camus, J. C., Cattolico, L., Chandler, M., Choisne, N., Claudel-Renard, C., Cunnac, S., Demange, N., Gaspin, C., Lavie, M., Moisan, A., Robert, C., Saurin, W., Schiex, T., Siguier, P., Thebault, P., Whalen, M., Wincker, P., Levy, M., Weissenbach, J. et Boucher, C. A.Genome sequence of the plant pathogen Ralstonia solanacearum. t. *Nature*, 415(6871):497–502.

Salzberg, S. L., Delcher, A. L., Kasif, S. et White, O.Microbial gene identification using interpolated Markov models. t. *Nucleic Acids Res*, 26(2):544–548.

Schattner, P.Searching for RNA genes using base-composition statistics. t. *Nucleic Acids Research*, 30(9):2076–82. PMID : 11972348.

Schell, M. A.CONTROL OF VIRULENCE AND PATHOGENICITY GENES OF RALSTONIA SOLANACEARUM BY AN ELABORATE SENSORY NETWORK. t. *Annual Review of Phytopathology*, 38:263–292. PMID : 11701844.

Shimoni, Y., Friedlander, G., Hetzroni, G., Niv, G., Altuvia, S., Biham, O. et Margalit, H.Regulation of gene expression by small non-coding RNAs : a quantitative view. t. *Molecular Systems Biology*, 3:138. PMID : 17893699.

Silvaggi, J. M., Perkins, J. B. et Losick, R.Genes for small, noncoding RNAs under sporulation control in Bacillus subtilis. t. *Journal of Bacteriology*, 188(2):532–41. PMID : 16385044.

Silvaggi, J. M., Perkins, J. B. et Losick, R.Genes for Small, Noncoding RNAs under Sporulation Control in Bacillus subtilis. t. *J. Bacteriol.*, 188(2):532–541.

Storz, G., Altuvia, S. et Wassarman, K. M.An abundance of RNA regulators. t. *Annual Review of Biochemistry*, 74:199–217. PMID : 15952886.

Thebault, P., de Givry, S., Schiex, T. et Gaspin, C.Searching RNA motifs and their intermolecular contacts with constraint networks. t. *Bioinformatics*, 22(17):2074–2080.

Thebault, P., Servant, F., Schiex, T., Kahn, D. et Gouzy, J.iANT (integrated ANnotation Tool). t. *In JOBIM Conference Proceedings*, (ENSA & LIRM, Montpellier, France.

Tjaden, B.Prediction of small, noncoding RNAs in bacteria using heterogeneous data. t. *J Math Biol.*

Tjaden, B., Saxena, R. M., Stolyar, S., Haynor, D. R., Kolker, E. et Rosenow, C.Transcriptome analysis of Escherichia coli using high-density oligonucleotide probe arrays. t. *Nucleic Acids Research*, 30(17):3732–8. PMID : 12202758.

Touzet, H. et Perriquet, O.CARNAC : folding families of related RNAs. t. *Nucl. Acids Res.*, 32(suppl_2):W142–145.

Tsuge, S., Terashima, S., Furutani, A., Ochiai, H., Oku, T., Tsuno, K., Kaku, H. et Kubo, Y.Effects on Promoter Activity of Base Substitutions in the cis-Acting Regulatory Element of HrpXo Regulons in Xanthomonas oryzae pv. oryzae. t. *J. Bacteriol.*, 187(7):2308–2314.

Ueki, T. et Lovley, D. R.Heat-shock sigma factor RpoH from Geobacter sulfurreducens. t. *Microbiology*, 153(3):838–846.

Ulvé, V. M., Sevin, E. W., Chéron, A. et Barloy-Hubler, F.Identification of chromosomal alpha-proteobacterial small RNAs by comparative genome analysis and

detection in Sinorhizobium meliloti strain 1021. t. *BMC Genomics*, 8:467. PMID : 18093320.

Upadhyay, R., Bawankar, P., Malhotra, D. et Patankar, S.A screen for conserved sequences with biased base composition identifies noncoding RNAs in the A-T rich genome of Plasmodium falciparum. t. *Molecular and Biochemical Parasitology*, 144(2):149–58. PMID : 16183147.

Valentin-Hansen, P., Eriksen, M. et Udesen, C.The bacterial Sm-like protein Hfq : a key player in RNA transactions. t. *Molecular Microbiology*, 51(6):1525–33. PMID : 15009882.

Valls, M., Genin, S. et Boucher, C.Integrated regulation of the type III secretion system and other virulence determinants in Ralstonia solanacearum. t. *PLoS Pathogens*, 2(8):e82. PMID : 16933989.

Valverde, C., Livny, J., Schlüter, J.-P., Reinkensmeier, J., Becker, A. et Parisi, G.Prediction of Sinorhizobium meliloti sRNA genes and experimental detection in strain 2011. t. *BMC Genomics*, 9:416. PMID : 18793445.

Vinogradov, A. E.Measurement by flow cytometry of genomic AT/GC ratio and genome size. t. *Cytometry*, 16(1):34–40. PMID : 7518377.

Vogel, J.A rough guide to the non-coding RNA world of Salmonella. t. *Molecular Microbiology*. PMID : 19007416.

Vogel, J., Bartels, V., Tang, T. H., Churakov, G., Slagter-Jäger, J. G., Hüttenhofer, A. et Wagner, E. G. H.RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria. t. *Nucleic Acids Research*, 31(22):6435–6443. PMC275561.

Vogel, J. et Sharma, C. M.How to find small non-coding RNAs in bacteria. t. *Biological Chemistry*, 386(12):1219–38. PMID : 16336117.

Voinnet, O.Origin, biogenesis, and activity of plant microRNAs. t. *Cell*, 136(4):669–87. PMID : 19239888.

Washietl, S. (2005). *Prediction of Structural Non-Coding RNAs by Comparative Sequence Analysis*. Thèse de doctorat.

Washietl, S. et Hofacker, I. L.Consensus Folding of Aligned Sequences as a New Measure for the Detection of Functional RNAs by Comparative Genomics. t. *Journal of Molecular Biology*, 342(1):19–30.

Washietl, S., Hofacker, I. L. et Stadler, P. F.Fast and reliable prediction of noncoding RNAs. t. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2454–9. PMID : 15665081.

Wassarman, K. M.6S RNA : a small RNA regulator of transcription. t. *Current Opinion in Microbiology*, 10(2):164–8. PMID : 17383220.

Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G. et Gottesman, S.Identification of novel small RNAs using comparative genomics and microarrays. t. *Genes & Development*, 15(13):1637–1651. PMC312727.

Wassarman, K. M., Zhang, A. et Storz, G.Small RNAs in Escherichia coli. t. *Trends in Microbiology*, 7(1):37–45.

Weilbacher, T., Suzuki, K., Dubey, A. K., Wang, X., Gudapaty, S., Morozov, I., Baker, C. S., Georgellis, D., Babitzke, P. et Romeo, T.A novel sRNA component of the carbon storage regulatory system of Escherichia coli. t. *Molecular Microbiology*, 48(3):657–70. PMID : 12694612.

Wetzstein, M., Völker, U., Dedio, J., Löbau, S., Zuber, U., Schiesswohl, M., Herget, C., Hecker, M. et Schumann, W.Cloning, sequencing, and molecular analysis of the dnaK locus from Bacillus subtilis. t. *Journal of Bacteriology*, 174(10): 3300–10. PMID : 1339421.

Wicker, E., Grassart, L., Coranson-Beaudu, R., Mian, D., Guilbaud, C., Fegan, M. et Prior, P.Ralstonia solanacearum strains from Martinique (French West Indies) exhibiting a new pathogenic potential. t. *Applied and Environmental Microbiology*, 73(21):6790–801. PMID : 17720825.

Wilderman, P. J., Sowa, N. A., FitzGerald, D. J., FitzGerald, P. C., Gottesman, S., Ochsner, U. A. et Vasil, M. L.Identification of tandem duplicate

regulatory small RNAs in Pseudomonas aeruginosa involved in iron homeostasis. t. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26):9792–7. PMID : 15210934.

Willkomm, D. K., Minnerup, J., Huttenhofer, A. et Hartmann, R. K.Experimental RNomics in Aquifex aeolicus : identification of small non-coding RNAs and the putative 6S RNA homolog. t. *Nucl. Acids Res.*, 33(6):1949–1960.

Workman, C. et Krogh, A.No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. t. *Nucleic Acids Research*, 27(24):4816–22. PMID : 10572183.

Zhang, A., Wassarman, K. M., Rosenow, C., Tjaden, B. C., Storz, G. et Gottesman, S.Global analysis of small RNA and mRNA targets of Hfq. t. *Molecular Microbiology*, 50(4):1111–24. PMID : 14622403.

Zhang, S., Haas, B., Eskin, E. et Bafna, V.Searching Genomes for Noncoding RNA Using FastR. t. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 2(4):366–379.

Zuber, U. et Schumann, W.CIRCE, a novel heat shock element involved in regulation of heat shock operon dnaK of Bacillus subtilis. t. *Journal of Bacteriology*, 176(5): 1359–1363. PMC205200.

Zuker, M.Mfold web server for nucleic acid folding and hybridization prediction. t. *Nucleic Acids Research*, 31(13):3406–15. PMID : 12824337.

Zuker, M. et Stiegler, P.Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. t. *Nucleic Acids Research*, 9(1):133–48. PMID : 6163133.

Zytnicki, M. (2007). *Localisation d'ARN non-codants par réseaux de contraintes pondérées*. Thèse de doctorat, Université Paul Sabatier.

Zytnicki, M., Gaspin, C. et Schiex, T.DARN ! A Weighted Constraint Solver for RNA Motif Localization. t. *Constraints*, 13(1-2):91–109.

Annexe A

Liste des ARN impliqués dans la virulence

ARNnc	Bactérie	Gènes cibles connus	Fonction	Mécanisme
MicA	*E. coli*	*ompA*	Porines/protéines de la membrane extérieure	Inhibition de la traduction et dégradation d'ARNm
MicC		*ompC*		
MicF		*ompF*		
OmrA-OmrB		*ompT*, *cirA*, *fecA*, *fepA*	Protéines de la membrane extérieure	Inhibition de la traduction et dégradation d'ARNm
RyhB		*sodB*, *acnA*, *sdhD*, *fumA*, *bfr*, *ftn*	Métabolisme du Fer	Inhibition de la traduction et dégradation d'ARNm
IstR		*tisAB*	Système antitoxin/toxin	Inhibition de la traduction et le clivage de la cible
SgrS		*ptsG*	Transport de glucose	Inhibition de la traduction et dégradation d'ARNm
GadY		*gadX*	Réponse à la condition d'acidité	Stabilisation d'ARNm
LhtA	*Chlamydia trachomatis*	*hc1*	Structure du nucléoide	Traduction ?
RatA	*Bacillus subtilis*	*txpA*	Système Antitoxintoxin	Dégradation d'ARNm

TAB. A.1 – Les ARNnc connus impliqués dans la virulence. Repris de Romby *et al.* (2006).

ARNnc	Bactérie	Gènes cibles connus	Fonction	Mécanisme
PrrF1, PrrF2	*P. aeruginosa*	*sodB*	Métabolisme du Fer	Inhibition de la traduction et dégradation d'ARNm
		sdhD		
		bfr		
Qrr1-Qrr4	*V. cholerae*	*hapR*	Virulence	Inhibition de la traduction et dégradation d'ARNm
	Vibrio harveyi	*luxR*	Bioluminescence	
RNAIII	*S. aureus*	*hla*	Synthèse de hémolysine	Activation de la traduction
		spa	Interaction hôte-pathogène	Inhibition de la traduction et dégradation d'ARNm
		sa1000	Virulence	Inhibition of translation and mRNA degradation
SprA-SprG		?	Virulence ?	?
FasX	*S. pyogenes*	?	Virulence (les facteurs secrétés)	?
	Pel	?	Virulence	?

TAB. A.2 – Les ARNnc connus impliqués dans la virulence. Repris de Romby *et al.* (2006).

ARNnc	Bactérie	Gènes cibles connus	Fonction	Mécanisme
VR ARN	*Clostridium perfringens*	?	Virulence (les toxines secrétées)	?
VirX		?	Virulence (les toxines secrétées)	?
CsrB/CsrC	*E. coli*	CsrA	Biosynthèse de glycogène, formation de biofilm , interaction hôte-bactérie	Séquestration de protéines
CsrB	*Salmonella typhymurium*	CsrA	Virulence	Séquestration de protéines
RsmZ/RsmY /RsmX	*Pseudomonas fluorescens*	RsmA	Exoenzymes, les metabolites secondaires	Séquestration de protéines
RsmZ/RsmB	*P. aeruginosa*	RsmA	Elastase, pyocyanin, métabolits secondaires	Séquestration de protéines
CsrB/CsrC /CsrD	*V. cholerae*	CsrA	Virulence	Protein sequestration

Tab. A.3 – Les ARNnc connus impliqués dans la virulence. Repris de Romby *et al.* (2006).

Annexe B

Test d'égalité de la proportion de G et C dans une séquence d'ADN

On considère une séquence d'ADN, et on est intéressé à tester les hypothèses suivantes :

$$\begin{aligned} H0 &: P_A = P_T \text{ et } P_C = P_G \\ H1 &: \text{Il existe au moins un } P_u \text{ différent des autres} \end{aligned}$$

où $P_u, u \in \{A, C, G, T\}$ est la probabilité d'observer le nucléotide u.

Les nucléotides sont supposés indépendants et identiquement distribués.

Il s'agit d'un test de comparaison de deux modèles emboités, le modèle M_0, associé à l'hypothèse nulle, étant un cas particulier du modèle M_1 associés à l'hypothèse alternative. On utilisera la théorie des tests de modèles emboités qui donne la statistique ainsi que sa loi asymptotique suivantes.

$$T = -2(\log L_0 - \log L_1) \sim \chi^2_2$$

où L_0 est la vraisemblance maximale sous l'hypothèse nulle, L_1 est la vraisemblance maximale sous l'hypothèse alternative. Le nombre de degré de liberté du modèle M_1 vaut 3 (car la somme des P_i est nulle), et le nombre de degré de liberté du modèle M_0 vaut 1. Sous H_0, T suit asymptotiquement une loi de χ^2 dont le nombre de degré de liberté vaut 3-1=2.

B.1 Modèle M_1

Le problème est de dériver les maxima de vraisemblance L_0 et L_1. En ce qui concerne, L_1, le résultat est bien connu, et on se contente de le redonner ici. La vraisemblance sous H_1 vaut

$$\mathcal{L}(X_1, ..., X_n, P_A, P_C, P_G, P_T) = \prod_{u \in \{A,C,G,T\}} P_u^{N_u}$$

où N_u est le nombre d'occurrence du nucléotide u dans la séquence $X_1, ..., X_n$.

La statistique du maximum de vraisemblance des paramètres du modèle M_1 est

$$\left\{ \widehat{P_u} = \frac{N_u}{N}, u \in \{A, C, G, T\} \right\}$$

où $N = \sum_{u \in \{A,C,G,T\}} N_u$. d'où la valeur maximal de la log vraisemblance suivante :

$$\log L_1 = N_A \log\left(\frac{N_A}{N}\right) + N_C \log\left(\frac{N_C}{N}\right) + N_G \log\left(\frac{N_G}{N}\right) + N_T \log\left(\frac{N_T}{N}\right)$$

B.2 Modèle M_0

En ce qui concerne le modèle M_0, les résultats demandent un peu plus de précautions. En effet, l'écriture du modèle général M_1 laisse supposer que le modèle possède 4 paramètres, ce qui n'est pas vrai, car il n'en possède que 3 linéairement indépendant, le quatrième paramètres pouvant se déduire des autres paramètres : $P_T = 1 - P_A - P_C - P_G$. En utilisant ce fait, la vraisemblance s'ecrit :

$$\mathcal{L}(X_1, ..., X_n, P_A, P_C, P_G) = (1 - P_A - P_C - P_T)^{N_T} \prod_{u \in \{A,C,G\}} P_u^{N_u}$$

L'hypothèse nulle devient :

$$\begin{aligned} H0 &: P_A = 1 - P_A - P_C - P_G \text{ et } P_C = P_G \\ H0 &: P_A = 1 - P_A - 2 \times P_C \text{ et } P_C = P_G \\ H0 &: P_A = \frac{1 - P_C}{2} \text{ et } P_C = P_G \end{aligned}$$

On se propose maintenant de dériver les estimateurs du maximum de vraisemblance des paramètres sous les contraintes spécifiées par H_0. Comme habituellement, on maximise de manière équivalente, la log-vraisemblance :

$$\begin{aligned}\log \mathcal{L}(X_1, ..., X_n, P_A, P_C, P_G) &= N_T \log(1 - P_A - P_C - P_T) \prod_{u \in \{A,C,G\}} N_u \log P_u \\ &= (N_T + N_A) \log \frac{1 - P_C}{2} + (N_C + N_G) \log P_C\end{aligned}$$

On dérive et on cherche le point qui annule la dérivée afin de trouver la valeur de P_C qui maximise la vraisemblance.

$$\frac{d \log \mathcal{L}(X_1, ..., X_n, P_A, P_C, P_G)}{dP_C} = \frac{-2(N_T + N_A)}{1 - 2P_C} + \frac{N_C + N_G}{P_C}$$

$$\begin{aligned}(1) \qquad & \frac{d \log \mathcal{L}(X_1, ..., X_n, P_A, P_C, P_G)}{dP_C} = 0 \\ (1) \iff & -2(N_A + N_T)P_C + (N_C + N_G)(1 - 2P_C) = 0 \\ (1) \iff & P_C = \frac{N_C + N_G}{2(N_C + N_G + N_A + N_T)}\end{aligned}$$

Finalement les estimateurs du maximum de vraisemblance sont donnés par

$$\begin{aligned}\widehat{P_C} &= \frac{N_C + N_G}{2N} \\ \widehat{P_G} &= \widehat{P_C} \\ \widehat{P_A} &= \frac{1 - \widehat{P_C}}{2} \\ \widehat{P_T} &= \frac{1 - \widehat{P_C}}{2}\end{aligned}$$

où $N = N_C + N_G + N_A + N_T$, et la log-vraisemblance maximal vaut :

$$\log L_0 = (N_C + N_G) \log \left(\frac{N_C + N_G}{2N} \right) + (N_A + N_T) \log \left(\frac{N_A + N_T}{2N} \right)$$