Université Paris-Sud 11

Ecole doctorale Gènes, Génomes, Cellules

Doctorat de l'université Paris-Sud XI

Mention : BIOLOGIE

MARCHAIS Antonin

ETUDE DE LA CO-EVOLUTION DES ELEMENTS CODANTS ET NON-CODANTS DES GENOMES BACTERIENS

Thèse dirigée par : GAUTHERET Daniel

Membres du Jury :

Mme Cécile Fairhead, Présidente du Jury

M. Daniel Gautheret, Directeur de thèse

M. François Major, Rapporteur

M. Mathias Springer, Rapporteur

Mme Hélène Touzet, Examinatrice

RESUME

Les ARN non-codants (ARNnc) forment une classe hétérogène d'éléments transcrits mais non-traduits qui participent, par l'intermédiaire de leur structure et/ou leur séquence, à une grande diversité de processus cellulaire selon de multiple modalités. Ces dernières années, un grand nombre d'ARNnc ont été détectés expérimentalement dans les génomes et aujourd'hui les ARNnc sont considérés comme des acteurs majeurs de la biologie des cellules eucaryotes, bactériennes et archaeénnes.

Malgré ces récentes avancées, l'étape de détection expérimentale des ARNnc dans les génomes nécessite toujours un investissement important en temps et reste encore rarement suivie d'une caractérisation fonctionnelle des éléments ainsi identifiés. Partant de ces constatations, nous avons développé une méthode de détection *in-silico* des ARNnc dans les génomes bactériens que nous avons appelée NAPP pour « Nucleic Acid Phylogenetic Profiling ». NAPP est une adaptation du profilage phylogénétique, une méthode utilisée pour prédire la fonction de protéine de fonction inconnue. NAPP analyse la co-occurrence d'éléments codants et non-codants d'un génome de référence dans toutes les espèces bactériennes disponibles. Il construit ainsi des groupes d'éléments d'histoire phylogénétique similaire et donc probablement dépendants les uns des autres, contraints par la nécessité de conserver l'intégrité d'un processus cellulaire. Parmi ces groupes, certains présentent un enrichissement en ARNnc connus, ce qui nous a permis logiquement de faire de NAPP un outil de détection des ARNnc dans les génomes bactériens. En comparaison à d'autres outils bioinformatiques, les performances de notre programme ont été quantifiées et se sont avérées très favorables dans plusieurs espèces modèles. Mais notre validation a été également expérimentale, permettant la découverte de 7 nouveaux ARNnc chez *S. aureus*.

L'un de ces ARNnc, RsaOG, a particulièrement retenu notre attention. En effet, fortement exprimé et conservé uniquement dans le genre Staphylococcus, RsaOG pourrait présenter une structure en pseudo-nœud encore jamais observée à notre connaissance dans un ARNnc agissant en *trans*. Après une étape de prédiction de ses cibles putatives, nous recherchons actuellement à les valider expérimentalement et à intégrer RsaOG dans la physiologie des Staphylocoques.

Mais la caractéristique la plus intéressante de NAPP est probablement sa capacité intrinsèque à fournir une indication fonctionnelle sur les éléments qu'il classifie. En effet, l'analyse de l'enrichissement en certaines fonctions des groupes d'éléments codants et non-codants peut, dans le meilleur des cas, constituer un indice sur la fonction des éléments de fonction inconnue contenus dans ces groupes. Ce type d'analyse nous a permis, chez *B. subtilis*, d'observer qu'un nouvel ARNnc, que nous avons appelé CsfG, se regroupait phylogénétiquement avec quasiment la moitié des gènes impliqués dans la sporulation. Nous avons donc inféré que CsfG était potentiellement lui-même impliqué dans la sporulation et des études expérimentales et *in-silico* nous ont permis de confirmer l'implication de CsfG dans la sporulation en validant sa régulation par les facteurs Sigma F et G spécifiques de la formation de la préspore.

ABSTRACT

Non-coding RNAs (ncRNA) form a heterogeneous group of transcribed, non-translated elements that play, through their structure, sequence and mechanistic diversity, an important role in many cellular processes. In recent years, thousands of ncRNA have been experimentally detected in genomes and ncRNAs are now accepted as key components of the cellular biology of Eukaryota, Prokaryota and Archaea.

Despite these recent advances, the experimental detection of ncRNAs remains a time-consuming task and is rarely followed by a functional analysis of the identified transcripts. To address this issue, we developed an *in-silico* method for ncRNA detection in bacterial genomes, named NAPP - Nucleic Acid Phylogenetic Profiling. This method derives from phylogenetic profiling, a method used to predict the function of unknown proteins. Based on a reference genome sequence, NAPP computes the co-inheritance of coding and non-coding elements in all available bacterial genomes and clusters elements that share a similar phylogenetic history. Several of these clusters are enriched in known ncRNAs, which enables using NAPP as an ncrNA classifier. Performance benchmarks indicate that NAPP predictive accuracy is equivalent to that of methods designed specifically for ncRNA detection. We further validate our predictions by the description of seven new ncRNAs in *S. aureus*.

We further studied RsaOG, one of the new *S. aureus* small RNAs identified by NAPP. RsaOG is highly expressed and specifically conserved in the Staphylococcus genus. This RNA may involve a pseudoknotted structure, a new observation for a *trans*-acting ncRNA. After a round of computational target predictions, we are now trying to validate RsaOG targets to integrate this small RNA in Staphylococcus physiology.

The most attractive feature of NAPP, however, is its intrinsic capacity to provide functional information on the classified elements. Indeed, functional enrichment of coding and non-coding clusters can, in favorable cases, provide clues about the function of unannotated elements in these clusters. Using this type of functional analysis in *B. Subtilis*, we focused on CsfG, an ncRNA detected by NAPP. Half of the genes in the CsfG cluster are involved in sporulation and we inferred that CsfG could be a sporulation-related RNA. Experimental and *in-silico* studies confirmed this prediction, demonstrating that CsfG is directly regulated by two Sigma factors sigG and sigF, that are specific to the prespore formation.

REMERCIEMENTS

Je tiens à remercier particulièrement Daniel Gautheret, mon directeur de thèse. Grace à la formation qu'il m'a octroyée, à son écoute et à son esprit d'analyse, j'ai pu m'épanouir scientifiquement. Je remercie mes co-équipiers Magali Naville et Chongjian Chen qui m'ont accompagné tout au long de ma thèse et de la leur. Je remercie Alban, Jean et Claire pour le soutien, les discussions et la vie de l'équipe. Je ne peux oublier de remercier Philippe Bouloc et Chantal Bohn car en plus de collaborateurs précieux ceux sont des gourmets de mets écossais comme moi. Je remercie Patrick Stragier pour notre collaboration fructueuse. Je remercie tous mes collègues de l'IGM avec lesquels j'ai passé de si bons moments. Je remercie les amis que j'ai pu découvrir sur les campus d'Orsay et d'ailleurs, Stéphane, Julien, Thierry, Marion. Je remercie les membres de mon jury de thèse pour le temps et l'intelligence qu'ils ont consacrés à la lecture de mon mémoire. Je remercie mon père, ma mère et mon frère Victor pour leur amour. Et je remercie ma compagne, Catherine, pour son soutien quotidien et sans laquelle, cette aventure n'aurait pas pu être aussi agréable.

TABLE DES MATIERES

Introduction 1

De l'acide ribonucléique aux ARN 3

Une période charnière 11

Ni des ARNt, ni des ARNr, ni des ARNm 13

Les ARN bactériens 15

La transcription chez les procaryotes 16

La traduction chez les procaryotes 17

Les ARNnc chez les procaryotes 18

la détection *in silico* des ARNnc bactériens 27

Les méthodes de détection basées sur la conservation de structure 28

Les autres méthodes de detection d'ARNnc 31

Articles 34

Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles. 34

Résumé 34

Article 37

RsaOG, a new staphylococcal family of highly transcribed non-coding RNA 46

Résumé 46

Article 47

A sporulation-specific, small noncoding RNA highly conserved in endospore formers 51

Résumé 51

Article 52

Articles en co-auteur 71

Experimental discovery of small RNAs in Staphylococcus aureus reveals a riboregulator of central metabolism 71

ARNold: A web tool for the prediction of Rho-independent transcription terminators 72

Discussion 73

Références 81

Annexes 88

Experimental discovery of small RNAs in Staphylococcus aureus reveals a riboregulator of central metabolism 89

ARNold: a web tool for the prediction of Rho-independent transcription terminators 106

INTRODUCTION

" Comment dire alors que l'esprit n'attend que des idées nouvelles pour s'en emparer et les exploiter ? Comment voir le développement des sciences comme guidé par la seule finalité de la logique ? Celle-ci ne peut manœuvrer que dans l'espace assigné par l'attitude du moment, autour des objets alors offerts à l'analyse. »

A propos de la redécouverte des travaux de G. Mendel, F. Jacob, « La logique du vivant », 1970

Mes recherches de doctorat se sont principalement focalisées sur l'étude de la coévolution des éléments codants et non-codants des génomes bactériens. L'objectif final de ce travail étant de développer une méthode *in silico*, qui, simultanément, détecte les ARN non-codants (ARNnc) dans les séquences de ces génomes, et assigne une fonction putative à ces ARN.

Depuis leur redécouverte au début des années 80, ces ARN, dits non-codants car ils ne sont pas traduits en protéine, résistent à l'analyse, en particulier à celle des bioinformaticiens. Néanmoins, grâce à la persévérance de chercheurs dans les 30 dernières années, nous savons aujourd'hui que les ARNnc forment une classe regroupant des éléments hétérogènes, qui peuvent aussi bien catalyser des réactions que réguler la transcription ou la traduction d'un nombre parfois important de gènes. Aujourd'hui, leurs rôles majeurs et leur présence nécessaire au bon fonctionnement de tous les êtres vivants ne sont plus discutés. En revanche ces résultats laissent supposer que nous sous-évaluons encore certainement la diversité de fonction et le nombre des ARNnc. Aujourd'hui il n'existe pas de méthode *in silico* réellement satisfaisante pour détecter leur signature dans des génomes entiers et pour leur attribuer une fonction putative, comme cela peut être effectué pour les gènes codants des protéines. Ainsi, à l'ère du séquençage massif des génomes, il semblait, et semble toujours, indispensable de s'attarder sur l'annotation positionnelle et fonctionnelle des ARNnc par des méthodes bioinformatiques automatisées.

Nous parlions plus haut d'une « redécouverte » des ARNnc car, en réalité, les premières hypothèses concernant leur existence, reprises par Crick dans son « dogme central », remontent aux années 50 et se réfèrent à une machinerie traductionnelle ribonucléique, base moléculaire du transfert d'information de l'ADN vers les protéines. Dans les années qui suivent, grâce à des démonstrations expérimentales élégantes, ses hypothèses sont confirmées et enrichies par la description dans tous les êtres vivants des pierres angulaires de la traduction que sont les ARN de transfert et les ARN ribosomaux. N'étant pas traduits, ces ARN constituent donc les premiers ARNnc décrits. Ces ARNnc présentant certaines spécificités – je les expliciterai plus tard --, je ne me suis pas concentré sur leur étude au cours de mon doctorat. Il me paraît cependant important de m'y attarder maintenant pour cette introduction. En effet, à l'époque déjà, la détermination du rôle des ARN dans les cellules vivantes aura nécessité toute la persévérance et la perspicacité, voire le génie, des plus grands biologistes du XXème siècle. Aussi, à la lecture des articles de cette période, on s'aperçoit que dès l'origine ont émergé des réflexions et des observations qui font toujours écho à nos questionnements contemporains. C'est pourquoi, pour bien comprendre les enjeux et les difficultés soulevés par l'étude des ARNnc et, en parallèle, introduire leurs caractéristiques communes et spécifiques, il me parait judicieux de débuter cette introduction par une description chronologique des recherches menées dans ce domaine. Cette chronologie ne sera pas nécessairement exhaustive car je souhaite qu'elle serve surtout à présenter les grands concepts qui ont émergé à propos des ARNnc. Je la découperai en trois chapitres. Le premier chapitre s'étendra de la caractérisation moléculaire de l'ARN dans une cellule vivante en 1939 à la caractérisation des principaux ARN impliqués dans le transfert de l'information en 1964. Le second couvrira la période charnière qui a suivi et qui, d'une certaine manière, a permis de préparer les esprits et les méthodes d'analyse nécessaires à la caractérisation de nouveaux types d'ARNnc, qui débuta dans les années 80. Enfin je terminerai cette chronologie par une brève présentation des travaux de recherche de ces trente dernières années qui ont révélé la complexité du monde moderne des ARNnc. Afin d'inscrire dans un continuum nos recherches et les méthodologies que nous avons utilisées, je poursuivrai cette réflexion en détaillant les différentes catégories d'ARNnc bactériens connues à ce jour ainsi que les recherches plus spécifiquement menées en bioinformatique depuis 10 ans dans le domaine de la détection des ARN non-codants. Finalement je conclurai cette introduction par une synthèse des questions que nous nous sommes posées au cours de mon

doctorat et j'expliciterai les raisons qui nous ont orienté, pour y répondre, vers l'étude de la coévolution des éléments codants et non-codants des génomes bactériens.

DE L'ACIDE RIBONUCLEIQUE AUX ARN

C'est en 1939, en plein développement de la biochimie, que Caspersson et Schultz décrivaient pour la première fois la présence d'acide ribonucléique (ARN) dans le cytoplasme des cellules de levure[1]. L'année suivante ils confirmaient aussi la présence de cette molécule dans le nucléole[2,3]. Simultanément, des travaux menés sur les constituants du protoplasme par Claude Albert, et qu'il synthétisera en 1943 dans Science[4], permettaient de discrétiser dans le cytoplasme cette substance ribonucléique comme le constituant d'une particule qu'il nomme alors microsome, et qui se révélera par la suite être le ribosome. Tous ces travaux seront réalisés soit en parallèle, soit dans la foulée, dans des cellules d'organismes couvrant les différents règnes du vivant avec des résultats identiques en ce qui concerne la présence et la localisation cellulaire de l'ARN. Ils permettront de définir l'ARN comme une particule centrale du vivant. Dès lors l'ARN ne cessera d'être l'un des objets de recherche majeur de la biologie moderne et son rôle dans la cellule sera activement étudié.

L'ARN, qui se différencie de l'ADN par son sucre -- un ribose au lieu d'un désoxyribose --, et par la présence d'uracile en place de la thymine, possède donc aussi une localisation cellulaire différente de l'ADN puisque ce dernier ne s'observe qu'au niveau du noyau et des mitochondries ou chloroplastes dans les cellules eucaryotes. Cependant plusieurs chercheurs des années 30 et 40 remarquent que l'ARN partage avec l'ADN, mais aussi les protéines, une caractéristique importante : toutes ces molécules semblent se répliquer puisqu'on les retrouve en quantité égale dans la cellule mère et dans chacune des cellules filles obtenues après la division mitotique[4]. Bien sûr ces observations sont encore difficilement interprétables et l'hypothèse dominante concernant l'hérédité ne leur permettra pas de faire coïncider un modèle cohérent avec les observations. En effet, les meilleurs candidats comme support de l'information permettant de répliquer les enzymes, qui sont à cette époque considérés comme les principaux éléments fonctionnelles de la cellule, semblent encore être les enzymes eux-mêmes. Cette interprétation est induite par

l'observation que leurs séquences, constituées à partir de 20 acides aminés, sont trop complexe pour être codées par les 4 bases de l'ADN ou de l'ARN. Ce qui peut se comprendre puisque l'hypothèse énoncée par Levene et Bass en 1931[5], selon laquelle l'ADN est une suite régulière et répétée des 4 nucléotides, est très répandue et ne correspond pas à ce que l'on attend d'une séquence informative. Ainsi, c'est l'autoréplication des enzymes qui reste l'explication la plus raisonnable de l'hérédité pendant encore quelques années, malgré des travaux explorant des directions différentes et l'accumulation d'indices qui désignent l'ADN comme étant lié à l'information génétique.

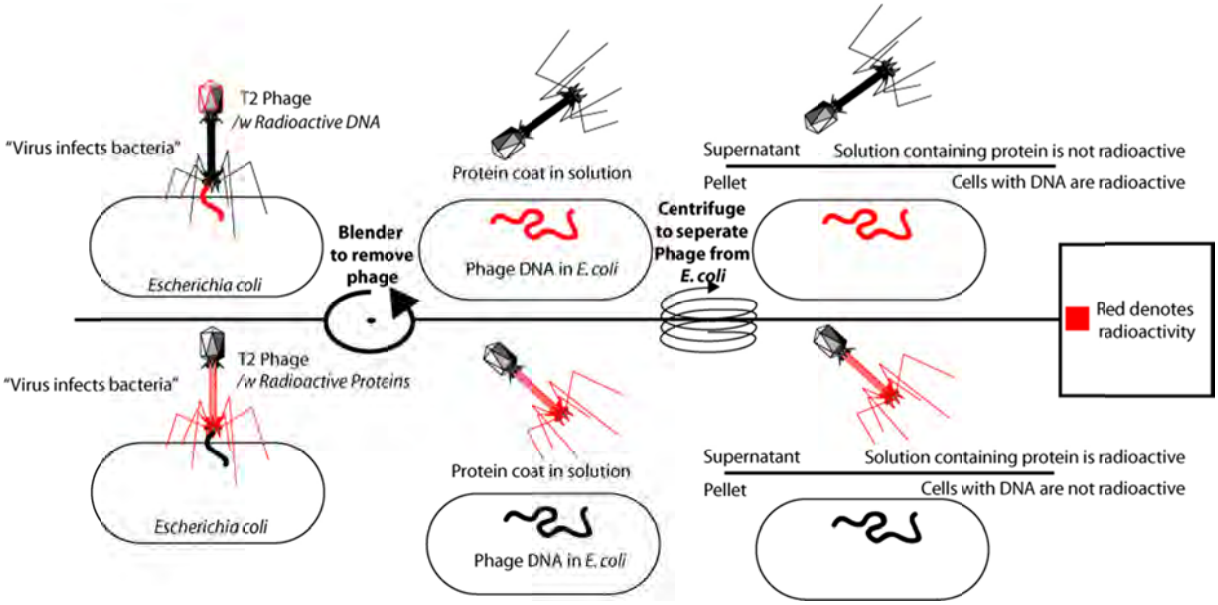
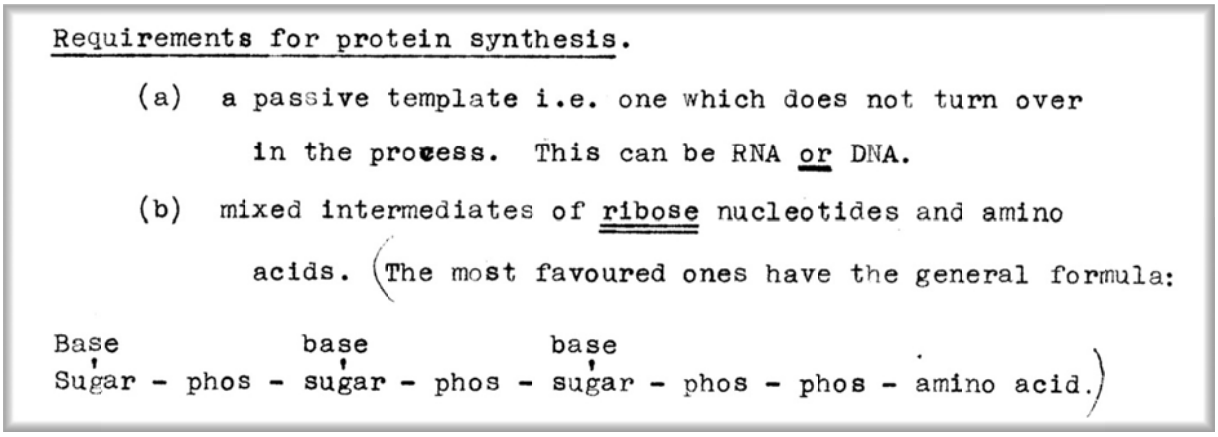


Figure 1. Hershey et Chase ont utilisé le phage à ADN T2 pour infecter *E. coli*. Ils débutent par le marquage de l'ADN phagique au P32 puis infectent *E. coli* avec ces bactériophages marqués. Ils observent après centrifugation que le culot, qui contient les cellules, est radio-marqué alors que le surnageant, qui contient des protéines de la capside, est non-marqué. Dans une deuxième étape ils vont marquer les protéines de la capside au S32 et après infection suivi d'une centrifugation, observer que cette fois-ci c'est le surnageant qui est marqué et non le culot. Ainsi ils prouvent que seul l'ADN phagique rentre dans la bactérie, et en infèrent que seul l'ADN, comme support de l'information, est nécessaire à la production de nouveaux phages T2.

Pour prouver définitivement que l'ADN est le support de l'hérédité, il faudra attendre la confirmation des résultats obtenus en 1944 par Avery[6], qui étudie la transformation de pneumocoques par de l'ADN, confirmation qui se fera en 1952 avec la démonstration par Hershey et Chase que l'ADN viral est le seul élément qui entre dans la bactérie hôte[7] (Figure 1). En ancrant moléculairement la génétique, cette découverte permettra de débloquer et d'accélérer

au cours des années 50 le développement d'un modèle valable des rôles de l'ADN et de l'ARN dans les cellules vivantes.

Dès 1952, Dounce[8] énonce pour la première fois avec clarté un modèle plus convaincant pour un concept qui a émergé en 1947[9] sur le transfert de l'information dans une cellule et qui peut se résumer par «DNA makes RNA, and RNA makes proteins ». En effet, les observations faites au cours des années 40 sur la localisation des différents constituant de la cellule et les preuves que l'ADN est le support de l'information, lui permettent de spéculer que l'ADN nucléaire spécifie la structure de l'ARN qui passe dans le cytoplasme où il spécifie à son tour la structure des protéines. Mais pour passer de ce concept à la validation expérimentale le chemin reste long. L'année suivante, Watson et Crick[10] démontrent que l'ADN est structuré en hélice double brin. Chaque base d'un brin est liée à une base du brin opposé selon la règle A-T/G-C, bien connue aujourd'hui. De plus, toutes les séquences de paires de base sont possibles. Cette dernière caractéristique est essentielle pour expliquer le rôle de l'ADN comme support de l'information. En effet, pour qu'une séquence constituée à partir de 4 bases puisse coder la séquence d'une protéine constituée de 20 acides aminés il est nécessaire que les 4 bases puissent s'arranger en au moins 20 combinaisons différentes.



Requirements for protein synthesis.

(a) a passive template i.e. one which does not turn over in the process. This can be RNA or DNA.

(b) mixed intermediates of ribose nucleotides and amino acids. (The most favoured ones have the general formula:

Base base base
Sugar - phos - sugar - phos - sugar - phos - phos - amino acid.)

Figure 2. Extrait des notes de Crick de 1956 sur la synthèse des protéines où pour la première fois il énonce sa théorie du dogme central et de l'adaptateur ARN qui, grâce à sa fixation à un acide aminé et au décodage des codons par un triplet complémentaire, permet la synthèse d'une protéine à partir d'un brin d'ARN.

Ainsi la description physico-chimique du support de l'information réactiva logiquement la question du codage de cette information. Rapidement il fut admis dans la communauté scientifique travaillant sur ce problème de codage que la taille du mot codant pour un acide

aminé, et appelé codon, devait être de trois nucléotides. Après quelques hypothèses infructueuses comme le « diamond code » où l'ADN interagit directement avec les acides aminés, Crick propose en 1955 de manière informelle, au sein du « RNA Tie club », l'hypothèse d'un adaptateur ARN (Figure 2). Partant de l'observation qu'il est peu probable que l'ARN interagisse directement avec les acides aminés à cause de problèmes de spécificité, il spécule que les acides aminés s'attachent à des trinucléotides spécifiques par à un mécanisme alors inconnu et que ces adaptateurs doivent être complémentaires aux codons d'une séquence d'ARN. Ainsi ils peuvent s'apparier comme le font les deux brins opposés d'ADN et permettre de traduire la séquence d'ARN en une séquence d'acide aminé. Cet adaptateur a en réalité été découvert la même année grâce à des études de fractionnement, et a été appelé sRNA pour ARN soluble[11]. Cependant, ce travail étant complètement indépendant du travail de Crick et la taille des ARN étant plus grande que celle attendue selon son modèle, il faudra attendre 1957 avec la poursuite des travaux de Hoagland et Zamecnik[12] pour que des expériences montrent que ces ARN se retrouvent dans la même fraction que les enzymes activant les acides aminés, et qu'ils sont chargés d'un acide aminé. Les années suivantes, plusieurs études prouvent que ces enzymes activateurs, connus aujourd'hui sous le nom d'ARNt-synthétases, catalysent aussi la transacétylation, étape de chargement spécifique des différents ARN solubles avec différents acides aminés[13,14]. Enfin, en 1960, de Allen et al[15] attribuent à cette fraction ARN son nom définitif d'ARN de transfert (ARNt).

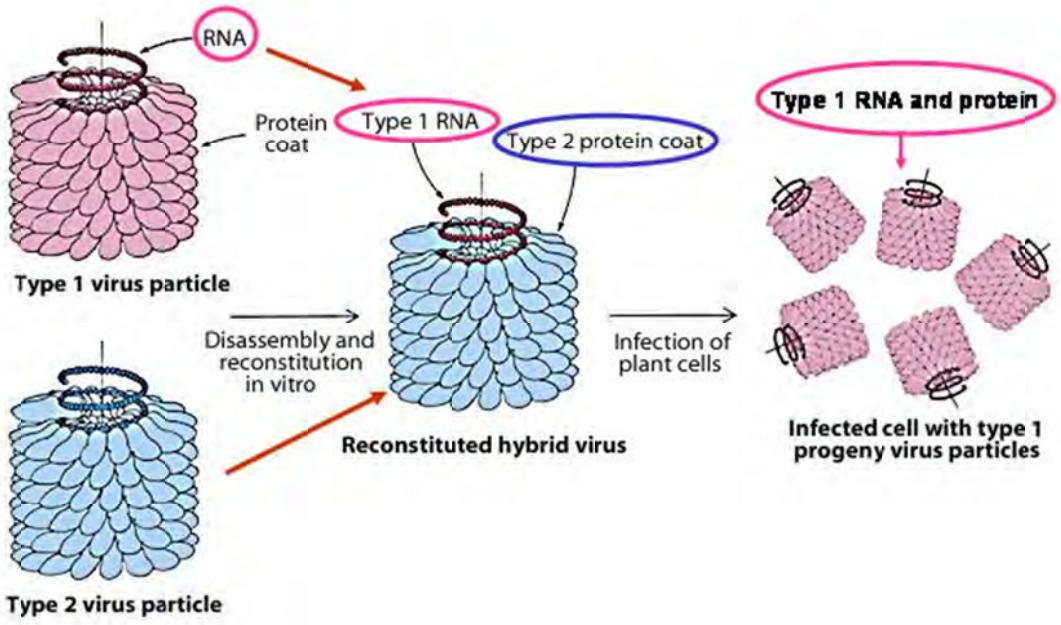


Figure 3. En insérant l'ARN du virus du tabac de type 1 dans une capside du virus de type 2, Fraenkel-Conrat et al[16-18] observent que les particules virales produites après infection sont constituées d'ARN et de capside de type 1. Ainsi l'ARN viral semble être l'élément qui supporte l'information nécessaire à la production des protéines de capsides de type 1.

Parallèlement à cette question du décodage, une autre problématique émergea sur le support ARN qui pouvait spécifier cette information et permettre la synthèse des protéines. En effet depuis les expériences sur le virus à ARN de la mosaïque du tabac (Figure 3), des arguments sont réunis pour pointer l'ARN comme le support de l'information nécessaire à la synthèse des protéines[16-18]. Cependant des études de composition de l'ARN total semblaient contredire ces résultats puisque cette composition se révélait beaucoup moins variable d'une espèce à l'autre en regard de la grande variation observée pour l'ADN. Ce sont Spirin et ses collaborateurs[19] qui, en 1957, par des expériences réalisées dans plusieurs bactéries suggèrent les premiers l'existence de deux sous-fractions d'ARN différentes. La première, représentant la majorité de l'ARN cellulaire, varie peu en composition, quand la seconde, beaucoup plus petite, suit la même composition que l'ADN des différentes espèces. Mais pour confirmer ces intuitions, il faut attendre les résultats des expériences de «pulse-chase » sur la composition de l'ARN nouvellement synthétisé, menés par Volkin et al de 1956 à 1960[20-23] et poursuivies dans la levure par Ycas et Vincent[24], qui confirment qu'il existe bien une sous-fraction respectant la composition de l'ADN. Ils appelleront cette sous-fraction DNA-like RNA (dRNA).

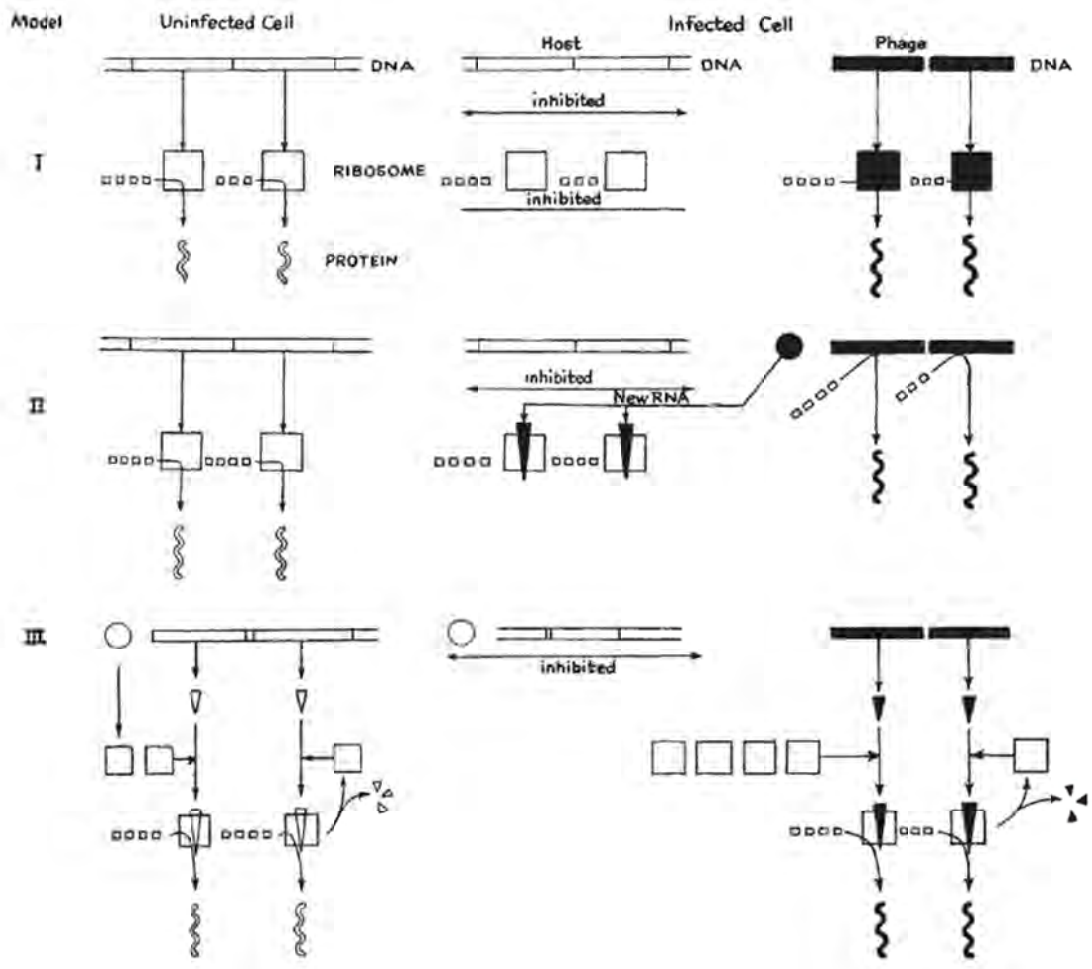


Figure 4. Brenner et al ont conduit une série d'expérience utilisant le bactériophage T2 à ARN connu pour inhiber la production de protéines de la bactérie qu'il infecte. Les cellules infectées produisent ensuite rapidement des protéines phagiques. Ces connaissances acquises, Brenner et ses collègues vont, par des expériences de « pulse-chase », tester trois modèles. Le premier soutient l'hypothèse selon laquelle le bactériophage inhibe la traduction chez la bactérie infectée et utilise sa propre machinerie traductionnelle. Le second ajoute la possibilité que les ARNm du bactériophage soient aussi traduits par la machinerie bactérienne. Enfin le troisième suggère que les ribosomes soient non-spécialisés et que les protéines phagiques soient toutes produites par la machinerie bactérienne détournée à cette fin. Ce sera le troisième modèle qui sera confirmé par les expériences où seuls les ribosomes de la bactérie sont dans la même fraction que les protéines phagiques neo-synthétisées. Ainsi le ribosome est non-spécialisé et peut traduire des supports ARN bactérien ou phagique.

L'année suivante, 1961, est une année charnière pour la biologie. En effet, c'est en 1961, grâce à la méthode d'hybridation de l'ADN avec l'ARN développée par Hall et Spiegelman[25], que l'on pourra déterminer que le dRNA possède la même séquence en bases que l'ADN[25,26] et que l'ARNr est écarté comme support de cette information puisqu'il s'hybride sur un nombre trop restreint de loci pour contenir l'information nécessaire au codage des protéines[27,28]. C'est aussi cette année-là que Jacob et Monod, dans leur célèbre étude sur les enzymes inductibles chez *Escherichia coli*[29], postulent que cette fraction dRNA serait de l'ARN messager transportant l'information de l'ADN vers la protéine. En effet, cette portion de dRNA est reportée par de nombreuses études comme disposant d'une très courte demi-vie, contrastant avec la grande

stabilité des ARNr. Pour Jacob et Monod, ceci est la condition nécessaire pour expliquer l'immédiateté de la réponse, par la synthèse d'enzyme, à l'ajout ou au retrait d'inducteurs. Toujours la même année, ce concept d'ARN messager (ARNm) sera appuyé par les élégantes expériences de « pulse-chase » de Brenner et al chez *E. coli*[30] (Figure 4) dans lesquelles, en plus de consolider l'idée selon laquelle les ARNm existent, qu'ils possèdent une courte demi-vie et qu'ils sont différents de l'ARNr, ils remarquent que « les ARNr sont des structures non spécialisées qui, à un temps donnée, synthétisent les protéines selon l'ARN messager qu'ils contiennent ». A partir de ce moment, le schéma général de la traduction des ARN messagers en protéines tel que nous le connaissons commence réellement à émerger. Des preuves fournies par Takanami[31,32] en 1962 démontrant que les ARNt chargés sont transférés dans le ribosome viendront compléter ce schéma général de la traduction.

Enfin la question du transfert de l'information de l'ADN vers l'ARN et donc les bases physico-chimiques du concept d'ARN messager seront en partie résolues par plusieurs groupes[33-35], qui parviendront à isoler une ARN polymérase dépendante de l'ADN et possédant la capacité de polymériser des ribonucléotides triphosphates pour produire une séquence d'ARN complémentaire d'un brin d'ADN. En 1963, ces évidences réunies auront convaincu la majorité des chercheurs que l'ADN est exprimé par transcription en ARN et que cette transcription dans la plupart des cas produit un ARNm qui détermine la séquence en acides aminés d'une protéine, produite au sein du ribosome grâce aux adaptateurs que sont les ARNt.

Outre les bases moléculaires du transfert de l'information, d'autres caractéristiques essentielles des ARN sont soulignées par ces fantastiques découvertes. La première est qu'il existe différentes fractions d'ARN. Une fraction codante, constituée des ARNm qui codent pour des protéines, et une fraction non-codante, dont font partie les ARNt et les ARNr. La confirmation qu'ils sont effectivement non-codants se fera quelques années plus tard avec le séquençage en 1964 de l'ARNt de l'alanine chez la levure et le déchiffrage du code qui gouverne la traduction des codons en acides aminés. Une autre caractéristique, qui pouvait passer inaperçu, était l'implication des ARNm dans la régulation de la synthèse des protéines. En effet, Jacob et Monod démontrent qu'il existe un programme cellulaire nécessitant l'implication des ARNm. Ils vont même plus loin dans la discussion qu'ils ouvrent à la fin de leur fameux article[29] de 1961, en proposant que l'inducteur soit aussi de l'ARN. Il se révélera que cet inducteur est d'une autre

nature, mais on peut noter l'incroyable esprit d'analyse de ces deux scientifiques qui ont touché du doigt ce qui sera découvert 30 ans plus tard, à savoir la régulation par les ARN. Evidemment d'autres feront de même. Ainsi des mises en garde sur le problème de la caractérisation des ARNm apparaissent dès 1963 dans un article de Roberts et al[36] et ne cesseront de se justifier par la suite. En effet, ces auteurs observent que définir les ARNm par leur court turn-over et une ressemblance en composition et en séquence avec l'ADN n'est pas suffisant. Plusieurs contre-exemples sont déjà observés en 1963. La seule validation définitive serait de montrer *in-vivo* qu'une séquence d'acide aminé correspondante existe. Or au début des années 60 il est très difficile d'obtenir ce type de résultat. D'une certaine manière, ces mises en garde soulèvent le fait qu'il pourrait exister dans la cellule vivante d'autres catégories d'ARN n'étant ni des ARNt ni des ARNr ni des ARNm. Cependant les chercheurs de cette époque ne disposaient pas des outils d'analyse moléculaire et biochimique adéquats pour enrichir leurs observations et par induction développer un modèle où d'autres types d'ARN seraient impliqués. C'est en cela que les années qui suivent joueront un rôle central. En effet, avant que de nouveaux ARN soient découverts à la fin des années 70, les biologistes vont approfondir leur connaissance sur la séquence, la structure, la fonction des ARN, mais aussi découvrir l'évolution moléculaire et la complexité des génomes et des systèmes de régulation.

UNE PERIODE CHARNIERE

Dès 1965, différents ARNt seront progressivement séquencés[37-43] et immédiatement des observations feront état de suites de bases complémentaires antiparallèles d'autres parties de la même molécule. Ainsi des régions d'un ARNt pourraient former des structures secondaires en tige-boucle en se repliant par l'intermédiaire d'appariements intramoléculaires. La structure en feuille de trèfle que nous connaissons sera dès l'origine pressentie comme étant la plus favorable des structures secondaires proposées par Holley[38] pour l'ARNt Alanine (Figure 5). L'article de Holley sera le point de départ d'une émulation internationale pour déterminer la structure tertiaire de ces ARNt qui s'achèvera par la publication simultanée de plusieurs articles entre 1974 et 1975 établissant la structure 3D de l'ARNt phe [44,45](Figure 5). Ces structures aideront rapidement à mieux comprendre l'empilement des bases, leur appariement, et à décrire un grand nombre des interactions tertiaires connues aujourd'hui dans les structures d'ARN[46](Figure 5).

Au-delà d'une période d'exploration et de développement technique fantastique les années 70 ont vu l'émergence de concepts nouveaux comme le « junk DNA ». Si aujourd'hui ce concept est dépassé, il représenta, quand il fut formulé en 1971 par Susumu Ohno, une véritable révolution et s'accompagna du développement de plusieurs notions. Partant du constat qu'un locus humain possède une probabilité de 1/100000 d'acquérir une mutation délétère par génération, Ohno infére qu'il ne peut pas exister plus de 30000 gènes dans le génome humain pour que le maintien de l'espèce soit assuré. Ainsi selon ces estimations et la quantité d'ADN contenue par un noyau de cellule humaine seul 6% du génome humain serait utilisé comme gène et le reste serait soit des « régions non-fonctionnelles utiles » (« the importance of doing nothing ») soit des régions contenant des gènes qui dégénèrent (pseudogènes) libérés de la pression de la sélection naturelle et pouvant servir à la création de nouveaux gènes (protogènes). Bien sûr, nous savons aujourd'hui que la réalité est différente et que ces grandes zones non traduites ou non transcrites ne sont pas nécessairement non-fonctionnelles. Néanmoins ce concept, qui a pu ralentir dans un premier temps la recherche d'éléments dans ces grandes régions non-codantes, aura, je pense, permis l'émergence de notions, qui quoique confuses, ont entraîné des questionnements précis qui seront en fin de compte utiles à la découverte des régions non-codantes fonctionnelles.

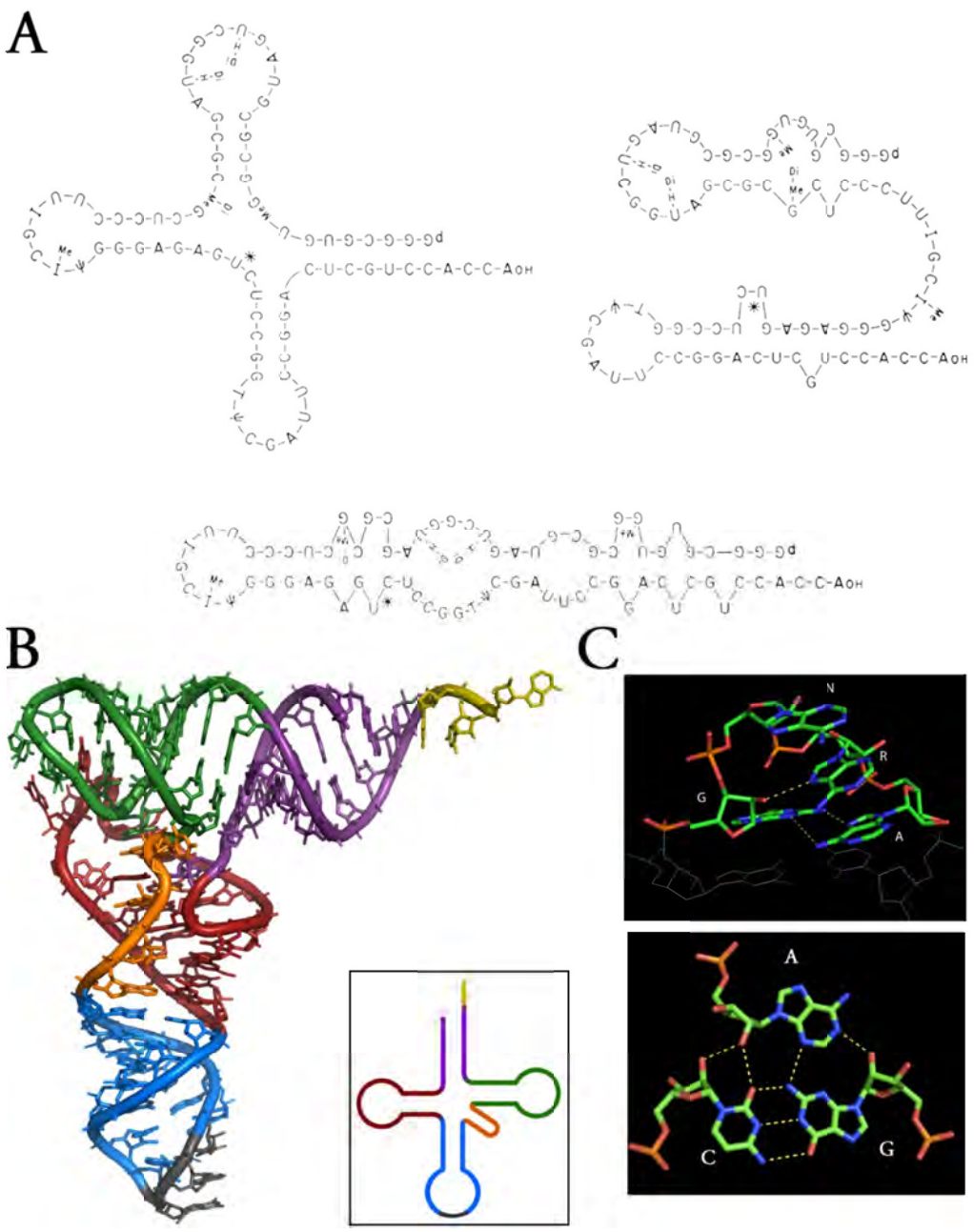


Figure 5. Le panneau A représente les structures secondaires proposées par Holley en 1965 à partir de la séquence de l'ARNt Alanine. La structure 3D obtenue en 1974 et 1975 pour l'ARNt Phénylalanine (panneau B) confirmera définitivement la structure secondaire en feuille de trèfle et servira de base à la découverte des motifs structuraux dans les autres ARNnc. Le panneau C reprend deux motifs structuraux aujourd'hui très bien documentés et retrouvés fréquemment dans les ARNnc, qui sont respectivement la tétra-boucle GNRA et le motif A-minor. Comme d'autres motifs découverts ces trente dernières années, on les retrouve pour la plupart sous de nombreuses variantes et combinaisons. Ainsi Hogley et Rich décrivent en 1976[46] le U-turn de l'ARNt phe, qui par la suite, sera observé à l'intérieur de certains motifs GNRA[47].

Une étape supplémentaire qui focalise, à nouveau, la communauté scientifique sur le problème des éléments non-codants, et plus spécifiquement de l'ARN non-codant, est la

découverte des introns dans les gènes eucaryotes en 1977, indépendamment par Roberts et Sharp[48,49]. En effet, ces éléments intragéniques sont transcrits et présents dans l'ARNm précoce mais ils sont absents de l'ARNm mature et par conséquent non-traduits en protéines. Ainsi ces éléments paraissent non-fonctionnels. Cependant dès l'année suivante, Walter Gilbert publie un article dans Nature[50] qui, en plus d'introduire le nom d'intron, pose la question essentielle « Why genes in pieces ? ». A cette époque les mêmes notions de dégénérescence de partie de gènes sont avancées comme possibles raisons de la présence de régions non-codantes au sein des gènes eucaryotes.

D'autres découvertes viendront alimenter des réflexions plus générales à propos des ARN. On peut citer, entre autre, le cas des virus à ARN double brin mais aussi le travail de Carl Woese sur l'évolution des séquences d'ARNr, lequel s'accélérera en 1977 grâce à la méthode de séquençage de Sanger[51].

Avec le recul, il semble que cette période d'accumulation d'observations toujours plus fines de l'ADN et de l'ARN ait permis de soulever des questions essentielles concernant la notion de fonctionnalité des éléments non-codants et ressemble à bien des égard aux débuts des années 50 que Crick décrivait comme « vagues » mais « nécessaires » au développement d'un optimisme scientifique[52].

NI DES ARNT, NI DES ARNR, NI DES ARNM

Ainsi, beaucoup d'ADN ne semblaient pas transcrits et beaucoup d'ARN ne semblaient pas traduits. Mais cela signifiait-t-il nécessairement que ces éléments étaient non-fonctionnels ? Dès 1978, deux études indépendantes, qui vaudront un prix Nobel de chimie en 1989 à Cech et Altman, apportent une première réponse. En effet, ces auteurs rapportent l'existence de deux types de transcrits non-traduits qui possèdent une activité catalytique, l'intron de groupe I[53] et la sous-unité de la ribonucléase P[54]. Ces ribozymes comme ils seront appelés, sont depuis les ARNt et les ARNr, les premiers ARNnc fonctionnels décrits. L'intron de groupe I catalyse sa propre excision de l'ARNm précoce dans lequel il est présent, et la sous unité de la ribonucléase P, ribonucléase que l'on retrouve dans toutes les bactéries, participe à la maturation des ARNt. Ainsi

des éléments non-traduits pouvaient posséder des activités comparables aux protéines et cette découverte, sachant qu'une grande partie des génomes eucaryotes et une partie des génomes procaryotes n'étaient pas traduites, permis d'envisager qu'il en existe bien d'autres.

Les 30 années suivantes seront riches en découvertes dans le domaine des ARNnc fonctionnels. Chez les eucaryotes sont découverts les petits ARN nucléaires. Ces ARNnc, en s'associant avec des protéines, sont impliqués dans des processus essentiels comme l'épissage, la régulation de facteurs de transcription[55] et le maintien des télomères. Une sous-catégorie, celle des petits ARN nucléolaires (snoRNA), guide même des modifications chimiques d'autres petits ARN nucléaires, d'ARN ribosomaux et de transfert, en s'appariant spécifiquement sur des régions de ces ARN. L'autre grande catégorie d'ARNnc chez les eucaryotes, les micro ARN (miRNA), seront découverts en 1993 par Lee et al[56]. Ces ARNnc sont impliqués dans la régulation post-transcriptionnelle. Avant d'exercer leur fonction, ils subissent une étape de maturation au cours de laquelle un transcrit primaire (pri-miRNA) formant une structure en tige boucle est coupé en une ou plusieurs tiges boucles plus petites appelées pré-miRNA. Ces pré-miRNA sont alors exportés dans le cytosol où ils subissent une nouvelle étape de maturation réalisée par la protéine DICER qui produit un micro ARN mature simple brin. Associé au complexe RISC, ce micro ARN mature s'apparie spécifiquement à un ARNm et ainsi guide le complexe RISC qui entraîne soit la dégradation soit le blocage de la traduction de cet ARNm cible. Ces mécanismes ont été initialement découvert chez *Arabidopsis thaliana*[57] avec l'étude des petits ARN interférents (siRNA), une autres classe d'ARN interférents comme nous appelons aujourd'hui tous ces petits ARN agissant en trans. Les siRNA guident, par ce mécanisme, l'inhibition de l'expression d'ARNm produits par des virus, ou le réarrangement de la chromatine. Les rôles de tous ces ARN interférents sont loin d'être complètement élucidés et de nouveaux types ou de nouvelles sous-catégories sont décrits périodiquement.

De la même manière, chez les procaryotes, de nombreux ARNnc fonctionnels seront décrits et impliqués dans des mécanismes aussi divers que la signalisation cellulaire, la régulation transcriptionnelle ou traductionnelle, la réponse antivirale et le marquage des protéines incomplètes. Je reviendrai plus en détail sur ces ARNnc bactériens dans le chapitre suivant. Plus généralement on peut remarquer que c'est aussi au cours de ces trente dernières années que la structure 3D du ribosome sera résolue[58,59] permettant ainsi une meilleure compréhension de son

fonctionnement et de la diversité des structures 3D que forme préférentiellement l'ARN. Ainsi, comme pressentie dans les années 80, l'activité enzymatique de la partie ribonucléique du ribosome sera confirmée[60].

En 1986 l'hypothèse du monde ARN, déjà énoncé par Carl Woese dans son livre The Genetic Code[61], était reprise par Walter Gilbert dans Nature[62]. Cette hypothèse propose que le précurseur à l'origine de la vie soit l'ARN, puisqu'il a la capacité de stocker l'information et de catalyser des réactions. Cette hypothèse permet de surcroît de contourner le paradoxe de l'œuf ou la poule induit par l'interdépendance qui relie l'ADN aux protéines, et la conservation de certains ARN dans tous les domaines du vivant ne fait qu'argumenter en sa faveur. Le rôle de l'ARN dans l'origine de la vie viendrait donc compléter le rôle central de l'ARN moderne établi pendant ces années frénétiques de découvertes et de révolutions méthodologiques.

Aujourd'hui, de nouvelles technologies de séquençage massif permettent d'obtenir des millions de séquences en quelques heures et l'on sait que les génomes sont dans leur quasi-totalité transcrits. Par conséquent, ce monde moderne des ARN s'enrichit à une vitesse qui est en train de dépasser nos capacités d'analyse, et de nouvelles problématiques théoriques et pratiques ont émergé. Aussi, les objectifs principaux de ma thèse, la détection automatisée de ces éléments dans les génomes séquencés et la prédiction de leur fonction, s'inscrivent dans ces récentes problématiques. Pour les introduire, je vais débuter par un rappel des connaissances actuelles sur le monde ARN chez les Bactéries.

LES ARN BACTERIENS

Avant de décrire les différentes catégories d'ARNnc bactérien, il me semble nécessaire de rappeler brièvement les mécanismes de transcription et de traduction au sein des bactéries. J'en profiterai pour préciser la structure classique d'un ARNm bactérien car cet élément joue souvent un rôle central dans le mode d'action des ARNnc chez ces organismes.

LA TRANSCRIPTION CHEZ LES PROCARYOTES

Chez les Bactéries la transcription d'un ARN est catalysée par un seul enzyme, l'ARN polymérase ADN-dépendante, constituée de cinq sous-unités : 2α, β, β'et σ. La transcription bactérienne se divise en trois étapes : l'initiation, l'élongation et la terminaison. L'initiation s'effectue au niveau du promoteur en amont de la séquence codante. Ce promoteur d'une taille d'environ 50 nucléotides est composé des régions -35 et -10 qui sont reconnues par la sous-unité σ de la polymérase, ce qui entraîne le recrutement des 4 sous-unités restantes (Figure 6). Le promoteur joue un rôle essentiel dans la régulation de la transcription des gènes. En effet les régions -10 et surtout -35 peuvent varier en séquence d'un gène à l'autre et recruter différentes sous-unités σ spécifiques. Par exemple, et nous le détaillerons plus tard, RpoS est un facteur σ spécifique de la réponse au stress et vient initier la transcription des gènes spécifiques de cette réponse en reconnaissant une séquence spécifique comme promoteur. D'autres protéines, appelées facteurs de transcription, peuvent aussi venir se fixer dans cette région promotrice pour activer ou réprimer la transcription d'un gène.

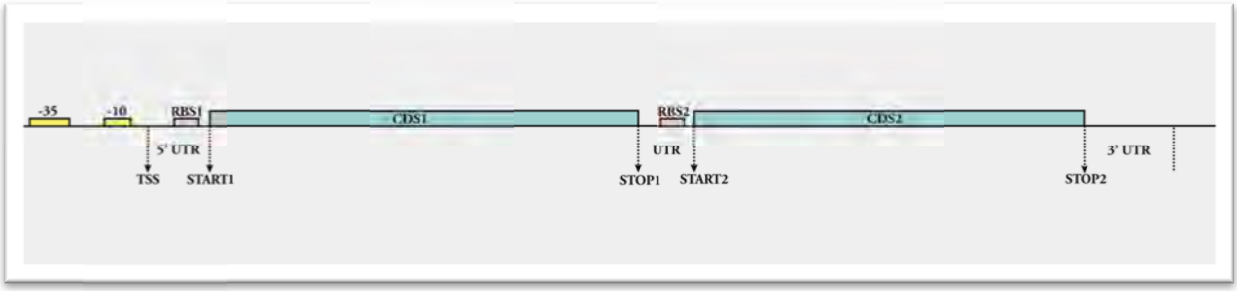


Figure 6. Schéma d'un segment d'ADN où sont illustrées les régions engagées dans la transcription et la traduction bactérienne d'un opéron. En jaune, le promoteur constitué des régions -10 et -35 en amont du site de début de la transcription (TSS). En rouge, les sites de fixation du ribosome (RBS) qui permettent d'initier la traduction de chacune des séquences codantes, en bleu, y étant associées. Ces deux séquences codantes commencent par un codon Start et finissent par un codon Stop. Entre le site de démarrage de la transcription et le premier codon Start se trouve la région 5' non traduites souvent impliquée dans les régulations par des ARNnc. Entre le dernier codon Stop et le site de terminaison de la transcription se localise la région 3' non traduite parfois impliquée dans des régulations par des ARNnc qui viennent stabiliser le transcrit.

L'élongation, qui commence après l'assemblage de l'ARN polymérase au niveau du promoteur, est l'étape de synthèse d'un ARN complémentaire du brin d'ADN transcrit dans le

sens 5' vers 3'. Sur le brin d'ADN plusieurs gènes consécutifs et souvent impliqués dans un même processus cellulaire peuvent être arrangés en opéron, c'est-à-dire placés sous le contrôle d'un seul promoteur. L'ARNm synthétisé à partir de cette séquence d'ADN contient donc plusieurs séquences codantes et est qualifié de polycistronique. L'étape finale de terminaison de la transcription peut s'effectuer selon deux modalités. La première est la terminaison Rho-dépendante. Elle fait intervenir la protéine Rho qui reconnaît une séquence riche en cytosines en amont du site de terminaison. Cette protéine rejoint l'ARN polymérase qui a ralenti au niveau de ce site de terminaison riche en bases CG. Puis son activité hélicase spécifique au duplex ARN-ADN libère l'ARN polymérase de ce duplex. Le second mécanisme de terminaison est Rho-indépendant et fait intervenir principalement la structure secondaire de l'ARNm en cours de synthèse pour bloquer l'ARN polymérase. En effet l'ARNm forme une structure en tige-boucle très stable d'une trentaine de nucléotides riche en base CG, appelée terminateur, qui ralentit la progression de l'ARN polymérase. La protéine NusA, en se fixant à cette tige boucle, entraîne alors une pause de l'ARN polymérase. Cette pause coïncide avec la transcription d'une région riche en adénine qui forme un duplex relativement instable avec l'uracile entraînant la dissociation du complexe et la fin de la transcription.

LA TRADUCTION CHEZ LES PROCARYOTES

Chez les bactéries la traduction est couplée à la transcription et débute dès que le site de fixation du ribosome a été transcrit. Ce site, appelé Shine-Dalgarno, est constitué par un motif de 5 bases se trouvant environ à 10 nucléotides en amont du codon Start dans la région 5' non traduite de l'ARNm (Figure 6). La petite sous-unité du ribosome vient se fixer par appariement à cette séquence. Puis la grande sous-unité est recrutée et l'élongation du polypeptide codé par le messager peut commencer à partir du codon Start dans le sens 5' vers 3'. Enfin, la synthèse du polypeptide se termine par la lecture du codon stop par le ribosome. La traduction nécessite l'intervention de multiples facteurs protéiques qui catalysent les réactions d'initiation, d'élongation et de terminaison. Comme chez les eucaryotes, plusieurs ribosomes peuvent traduire simultanément un ARN messager et former ce que l'on appelle un polysome. En revanche, chez les bactéries, l'existence d'ARNm polycistroniques fait qu'un ARNm peut contenir plusieurs

séquences codantes avec en amont de la plupart une séquence Shine-Dalgarno. Ainsi ces polycistrons sont traduits en plusieurs polypeptides de manière simultanée et indépendamment les uns des autres.

LES ARNNC CHEZ LES PROCARYOTES

Comme nous l'avons vu précédemment, un grand nombre des transcrits d'une cellule ne sont pas traduits et peuvent néanmoins posséder des fonctions essentielles. La détection de ces éléments dans les génomes bactériens a été le sujet central de ma thèse. Aussi, avant d'aborder les méthodes existantes de détection bioinformatique, il est nécessaire de bien comprendre la diversité que recouvre cette classe d'éléments non-codants chez les procaryotes. En effet grâce aux études menées depuis les années 80 dans ce domaine on a pu classifier ces ARNnc selon leur mode d'action. Malgré une petite taille, généralement comprise entre 100 et 200 nucléotides, ces ARNnc sont impliqués dans une multitude de fonctions comme la modulation de la transcription, de la traduction, le contrôle de la stabilité des ARNm ou bien la séquestration de protéines.

Les ARNnc constitutifs

Cette catégorie regroupe un nombre limité d'éléments. Ils sont souvent ubiquitaires aux différents domaines du vivant et essentiels à la survie des cellules. Ces ARN possèdent comme autre caractéristique commune des séquences très conservées à travers les espèces, et leur détection dans les génomes est maintenant presque triviale. Les ARNt et les ARNr qui composent le ribosome en font partie. Cependant, plus récemment cette catégorie s'est étoffée. Ainsi on peut nommer l'ARN 4,5S, un composant du système de « signal recognition particle », la sous-unité ARN de la RNase P[63], impliquée dans la maturation de l'ARNt et d'autres ARN, et le tmRNA qui marque les protéines traduites de manière incomplète pour les rediriger vers le dégradosome[64].

Les ARNnc agissant en *cis*

Souvent présentés comme les ARN régulateurs les plus simples, ces éléments sont des séquences localisées dans la région 5' non-traduite des ARNm et qui peuvent adopter des structures alternatives en réponse à un signal environnemental cytoplasmique comme un ribosome bloqué, un ARNt non chargé, une élévation de température ou la présence en concentration suffisante d'une petite molécule. Si le signal est un métabolite on parlera de riboswitch. Ces éléments sont généralement organisés en deux parties : la région « aptamère » qui reconnait le ligand et la plateforme d'expression qui régule l'expression du gène selon la structure alternative qu'elle adopte. Dans le cas des riboswitchs, la fixation d'un ligand entraîne un changement de conformation qui habituellement se traduit par la formation d'une tige boucle alternative qui stabilisera ou déstabilisera un terminateur ou un antiterminateur transcriptionnel ou qui séquestrera ou libérera un site de liaison au ribosome[65] (Figure 7). Cependant la plupart des riboswitchs répriment la transcription ou la traduction et l'activation ne concerne qu'un petit nombre d'entre eux. En général, les riboswitchs régulent des gènes qui sont impliqués dans l'utilisation des métabolites auxquels ils sont sensibles. Ainsi les riboswitchs qui fixent un mononucléotide flavine régulent l'expression d'enzyme de biosynthèse de la flavine, alors que ceux qui fixent la lysine régulent certains gènes impliqués dans la synthèse de la lysine.

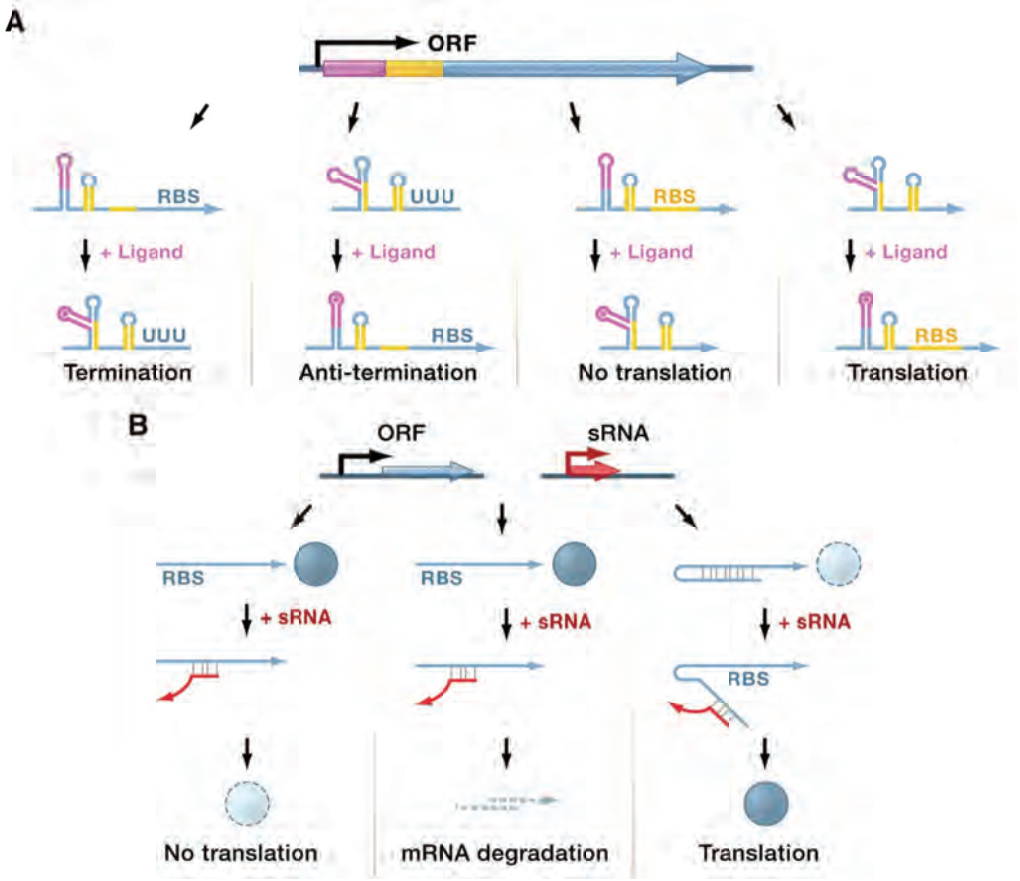


Figure 7. Représentation schématique des mécanismes de régulation impliquant les ARNnc. A : les ARNnc agissant en *cis* et. B ; les ARNnc agissant en trans. Sur le panneau A ne sont représentés que les ARN agissant en *cis* localisés dans le 5' UTR des transcrits. Ceux-ci, en réponse à un ligand, adoptent une nouvelle structure secondaire qui peut stabiliser ou déstabiliser un terminateur de la transcription ou bien agir sur la traduction du transcrit par séquestration ou libération du site de fixation au ribosome. Sur la panneau B, seul les ARNnc agissant en *trans* sur la région 5' UTR d'un transcrit sont représentés. Ceux-ci agissent principalement en s'appariant sur la région du site de fixation du ribosome pour le séquestrer et inhiber la traduction. Cette inhibition peut être accompagnée d'une dégradation de l'ARN par des ribonucléases. Cependant ces ARNnc peuvent aussi libérer un site de fixation au ribosome en entrant en compétition avec une tige-boucle déjà formée par la région 5' UTR. D'autres mécanismes qui ne sont pas illustrés ici entrainent une modulation de l'expression des transcrits ciblés. Ainsi quelques ARNnc ciblent la région 3'UTR des ARNm et entraînent une augmentation de la stabilité de ces transcrits[66]. D'autres encore ciblent des 5' UTR classiques tout en étant eux même titrés par des ARNm polycistroniques[67]. Cette figure est adaptée de Waters et al[68].

Ces riboswitchs ont une nature modulaire intrinsèque et extrinsèque. En effet un même aptamère peut induire différents types de régulation selon la plateforme d'expression à laquelle il est associé, et on peut retrouver un riboswitch identique en amont de plusieurs gènes dans un génome. Certains se retrouvent même en tandem avec un autre riboswitch répondant à un signal différent[69]. D'autres, comme GlmS, sont des ribozymes qui créent une boucle de rétro-contrôle négatif en inactivant l'ARNm qui code pour leur ligand en s'auto-clivant[70]. On peut noter que la distribution des riboswitchs ne semble pas homogène selon les espèces. Ainsi dans les bactéries Gram-positives les riboswitchs sont majoritairement des atténuateurs transcriptionnels alors que

les riboswitchs de bactéries Gram-négative sont plus fréquemment des inhibiteurs de la traduction[71]. Ces organismes Gram-positif semblent aussi contenir un plus grand nombre d'ARN agissant en *cis* alors que les bactéries Gram-négatives semblent préférer les ARN régulateurs agissant en *trans*. Toutefois ces différences apparentes s'estompent depuis que ces éléments sont recherchés plus activement dans toutes les bactéries[72].

Les ARNnc interagissant en *cis*-antisens

Au moins aussi répandus que les riboswitchs, ces ARN régulent l'expression des gènes de manière très différente en s'appariant sur les ARNm. On peut séparer ces ARNnc en deux sous-catégories. La première, que je vais décrire dans ce chapitre, est transcrite sur le brin d'ADN opposé d'un ARNm et partage donc avec lui une grande région strictement complémentaire[73]. La seconde, qui sera l'objet du chapitre suivant, partage une complémentarité souvent incomplète avec un ARNm ciblé, localisé à une autre position du génome.

Nombre d'ARN interagissant en *cis*-antisens sont exprimés par des bactériophages, des plasmides ou des transposons et leur permettent de maintenir un nombre de copie suffisant d'éléments mobiles par l'intermédiaire de mécanismes très variés. Un autre groupe très représenté concerne les systèmes toxine-antitoxine où l'ARNnc permet de réprimer suffisamment le niveau de traduction de protéines toxiques pour que la cellule survive[74]. Ils sont probablement aussi impliqués dans la défense contre des plasmides qui contiendraient des protéines toxiques homologues. Mais ces ARN peuvent aussi être exprimés en anti-sens d'un opéron. C'est le cas de GadY qui, exprimé en antisens de l'ARNm GadXW, provoque son clivage en deux unités de transcription GadX et GadW[75].

Ces ARNnc *cis*-antisens sont impliqués dans bien d'autres mécanismes et nous sommes certainement encore loin de disposer d'une liste complète de ces ARN et de leurs fonctions. En effet, plusieurs études récentes de séquençage massif des petits ARN[76,77] ont rapporté l'existence de milliers d'ARN antisens d'ARNm.

Les ARNnc interagissant en *trans*

Cette autre catégorie d'ARNnc fonctionnant par appariement de base est assurément la plus étudiée ces dernières années chez les bactéries. Ces ARNnc régulent la traduction et/ou la stabilité des ARNm auxquels ils s'apparient. Cette régulation est essentiellement négative et induit donc une inhibition de la traduction et/ou une augmentation du taux de dégradation de l'ARNm ciblé[78]. Cette inhibition traductionnelle s'effectue le plus souvent par la séquestration du site de liaison au ribosome (RBS) de l'ARNm, par l'appariement avec l'ARNnc, mais d'autres stratégies sont recensées (Figure 7). Par exemple, on peut citer les ARNnc GcvB et RyhB qui inhibent la traduction par l'intermédiaire d'un appariement très en amont du RBS[79], ou encore MicC qui s'apparie en aval de l'ATG de l'ARNm et doit probablement bloquer la course du ribosome[80]. La formation de tels duplexes conduit souvent à leur dégradation par la RNase E. Néanmoins certains ARNnc agissant en *trans* peuvent aussi activer l'expression d'un ARNm. Ainsi DsrA et RprA, deux ARNnc qui régulent positivement l'expression de l'ARNm *rpoS*, rentrent en compétition avec une tige boucle qui séquestre naturellement le RBS et ainsi le libèrent[81].

Dans leur grande majorité, les ARNnc agissant en *trans* peuvent s'apparier avec de multiples ARNm[82,83]. Cette multiplicité des cibles explique probablement en partie la complémentarité limitée que l'on observe entre les ARNnc et leurs ARNm cibles. En effet les sites d'appariement de ces ARNnc sur les ARNm sont souvent aussi le lieu d'interactions avec d'autres éléments qui peuvent varier d'une cible à l'autre. De plus si l'ARNnc cible la région du RBS, l'appariement doit nécessairement comporter des irrégularités afin d'éviter de s'apparier à tous les RBS. Enfin ce type d'appariement procure certainement un avantage évolutif que je discuterai plus tard.

En plus d'être limités par des mésappariements, ces appariements sont souvent courts. Ils impliquent typiquement 10 à 25 nucléotides et tous les cas qui ont été étudiés en détail prouvent que seul le cœur de l'appariement est essentiel. Ainsi les interactions des ARNnc avec leur cible paraissent peu stable et pourrait même être transitoires. Cependant, il faut nuancer ces observations, car la plupart des régulations par ARNnc agissant en *trans* nécessitent l'intervention de la protéine chaperonne à ARN Hfq qui facilite ces interactions[84]. Cette protéine homohéxamèrique en anneau, qui possède un domaine sm-like, est en effet une protéine centrale

de la régulation en *trans* par les ARNnc. En plus de faciliter les duplexes ARN/ARN elle est impliquée dans la stabilisation de certains ARNm et ARNnc agissant en *trans* en s'y fixant et pourrait être nécessaire au recrutement de protéines tierces comme la RNase E. Des études récentes chez *S.thypimurium*[85], où tous les ARN co-immunoprécipitant avec Hfq ont été séquencés, ont révélé qu'au moins un cinquième des ARNm interagirait avec Hfq. Dès lors on pourrait même proposer que Hfq « scanne » tous les ARNm d'une cellule. Cependant certains ARNnc agissant en *trans* ne nécessitent pas l'intervention d'Hfq et on peut remarquer qu'il s'agit en général des ARN qui forment de plus longs appariements. De plus certaines bactéries, essentiellement des firmicutes, ne disposent pas d'une protéine Hfq active vis-à-vis de l'action des ARNnc. D'autres protéines et mécanismes sont sûrement requis dans ces bactéries puisque leurs duplexes ARN/ARN partagent les mêmes propriétés que ceux observés dans les espèces disposant d'une protéine Hfq active.

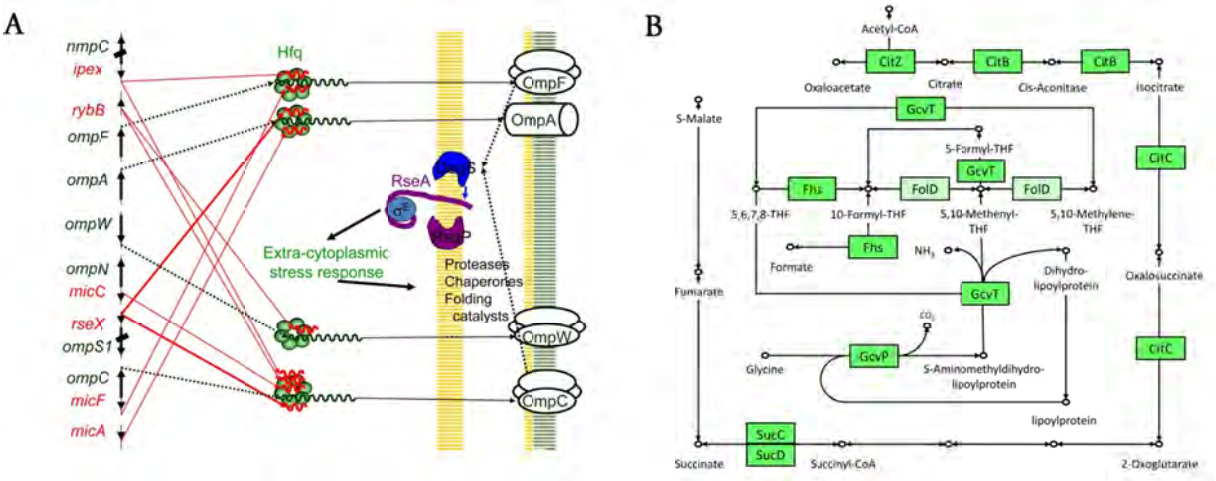


Figure 8. Le panneau A illustre un réseau de régulation impliquant les ARNnc ipeX, rybB, micC, rseX, micF et micA dans la régulation de l'expression des protéines de la membrane externe en réponse à des stress divers comme des modifications de l'osmolarité ou des problèmes de repliement des protéines de la membrane externe[86]. Le panneau B représente les réseaux métaboliques du folate et du succinate chez *S. aureus*. La traduction des gènes en vert est régulée post-transcriptionellement par l'ARNnc RsaE[87]. Ces deux figures illustrent la diversité de l'implication des ARNnc dans des réseaux.

Tous ces ARNnc agissant en *trans* sont majoritairement exprimés dans des conditions spécifiques de croissance (Figure 8). Par exemple chez *E. coli* ils sont induits par des stress oxydants, des stress de la membrane, des concentrations faibles en fer ou des changements dans la quantité de glucose disponible[88-91]. On les retrouve aussi impliqués dans le quorum-sensing chez *V. cholerae*[92] et dans la régulation de facteurs de virulence chez d'autres bactéries[93]. En vertu de

leur capacité à réguler de multiples cibles ces ARN sont des facteurs de régulations centraux de la cellule, qui peuvent moduler globalement des réponses à des conditions physiologiques particulières.

Les ARNnc interagissant avec les protéines

Si l'on exclue les ARN classés comme constitutifs, cette classe d'ARNnc n'est composée actuellement que de quatre ARNnc : CsrC, CsrB, GlmY et l'ARN 6S. Ces ARN agissent tous comme des compétiteurs négatifs qui viennent titrer des protéines en mimant les séquences nucléotidiques auxquelles elles sont associées habituellement. CsrC et csrB chez *E. coli* modulent ainsi l'activité de CsrA[94]. Cette protéine homodimérique se lie à des motifs GGA dans les 5' non-traduits de ses cibles ARNm avec pour conséquence d'affecter leur stabilité et/ou leur traduction. CsrC et csrB contiennent aussi respectivement 13 et 22 motifs GGA. Ainsi plus leur niveau d'expression augmente et plus elles titrent d'homodimères CsrA. De la même façon l'ARN 6S mime un promoteur qui lie et séquestre le facteur σ 70 au cours la phase stationnaire de croissance modifiant ainsi la compétition entre ce facteur et le facteur σS spécifique de cette phase de croissance. Enfin, plus récemment, GlmY a été reporté comme étant éventuellement impliqué dans ce même type de mécanisme[95].

Les CRISPR

Le système CRISPR, découvert récemment, est constitué d'une catégorie particulière d'éléments non-codants qui se distinguent des autres ARNnc par leur organisation, leur mécanisme d'action et la manière dont ils évoluent. Ces ARN CRISPR et les protéines CAS qui y sont associées sont impliqués dans la résistance aux bactériophages et préviennent la conjugaison de plasmides[96]. On les retrouve dans 40% des bactéries séquencées.

La séquence ADN des ARN CRISPR est organisée comme suit. Une séquence leader d'environ 500 paires de bases est suivie d'une série d'éléments alternant entre une répétition directe et un espaceur. Au sein d'un CRISPR donné, toutes les séquences répétées sont quasiment

identiques et peuvent être répétées de 2 à des centaines de fois. Leur taille varie de 24 à 47 paires de bases selon les CRISPR. Ces répétitions alternent avec des espaceurs qui peuvent varier en taille et en séquence. Les espaceurs, à la différence des répétitions, ne sont pas conservés au sein d'un même locus CRISPR et entre différentes espèces. Les espaceurs sont souvent homologues avec des régions de phages et de plasmides. Il a été observé après des études d'infection phagique l'acquisition par la séquence CRISPR de nouveaux espaceurs homologues à des séquences de phages. De plus ces nouveaux espaceurs confèrent alors une résistance de la bactérie à ces phages[97].

La séquence ADN des ARN CRISPR est transcrite en un long ARN. Les protéines CAS maturent ces longs ARN pour produire des petites unités, appelées crRNA, constituées d'une répétition et d'un espaceur. Puis de manière similaire au fonctionnement des siRNA eucaryotes, des protéines CAS sont conduites, par l'appariement de ces petites unités sur l'ADN phagique ou plasmidique, à inhiber le fonctionnement de ces éléments[97]. Les mécanismes impliqués dans cette chaîne de réactions sont généralement inconnus. Leur étrangeté, et certainement aussi le fait qu'ils évoquent un système immunitaire rudimentaire, font des CRISPR un sujet d'étude particulièrement actif. On peut remarquer que des éléments transcrits appelés STARs[98] (*Staphylococcus aureus* repeat), de fonction inconnue et que l'on retrouve chez *S. aureus*, ressemblent aux CRISPR dans leur organisation génomique.

De nombreuses questions en suspens

Les 10 dernières années ont été particulièrement riches en nouvelles connaissances au sujet des ARNnc bactériens. Cependant ce domaine de recherche est encore récent et de nombreuses questions n'ont pas encore trouvé de réponse.

Ainsi, tous ces ARN semblent se replier pour former des structures secondaires et tertiaires assez stables. Pour certaines classes d'ARNnc, comme les ribozymes ou les riboswitchs, ces structures sont directement impliquées dans leurs fonctions clairement identifiées. Cependant pour la classe d'ARNnc agissant en *trans,* le rôle de leur structure dans leur mécanisme/mode d'action reste plus problématique. En effet, au-delà d'un rôle de stabilisation de l'ARN ces

structures n'ont pas a priori d'utilité dans l'appariement nécessaire à la régulation. La plupart des appariements impliquent d'ailleurs les parties simple brin de l'ARNnc. Peut-être est-ce justement l'antagonisme intrinsèque de ces structures à la formation d'un appariement qui permet de moduler différemment plusieurs cibles en étendant l'appariement par l'ouverture de ces structures selon leur complémentarité avec la cible? Le rôle de protéines comme Hfq, la RNase E et d'autres hélicases pose également des questions. Ces protéines ont-elles besoin de ces structures pour agir ou sont-elles simplement séquence spécifique ? Dans ce dernier cas, quels sont les motifs nucléotidiques reconnus, étant donné que les séquences des ARNnc et des régions ciblés sont très variables ? Cette variation induite par la complémentarité entre les ARNnc et leurs cibles est-elle compatible avec des règles d'appariement générales? Il est probable que ces règles varient selon la localisation de l'appariement, au niveau du RBS, dans les séquences codantes ou autres. Plus généralement, il existe probablement d'autres mécanismes d'actions des ARNnc. Ainsi on peut imaginer des riboswitchs transcrits indépendamment d'un ARNm qui agiraientt en *trans* en régulant différents ARNm selon leur position « fermée » ou « ouverte » induite par un ligand. On peut également proposer des ARN qui associeraient des activités catalytiques à des activités de régulation du type ARNnc agissant en *trans*.

La question du rôle physiologique des ARN régulateurs est tout aussi importante. La grande majorité des ARNnc bactériens décrits n'ont pas de fonctions associées car les étapes expérimentales qui suivent la détection de leur signal de transcription se révèlent encore longues et difficiles. Ainsi, on peut imaginer que dans les prochaines années de nouvelles fonctions seront découvertes puisque, à l'image des ARNnc eucaryotes, on peut parfaitement concevoir leur implication dans des mécanismes de modification des ARN ou de réparation et de modification de l'ADN. De plus, un champ de recherche encore peu exploré est la phylogénie de ces ARN. Leur faible conservation restreint en effet fortement les travaux qui peuvent être menés pour retracer leur histoire évolutive. L'avalanche de génomes séquencés permettra probablement, dans les années à venir, de reconstruire des phylogénies plus résolutives et de déterminer certaines règles sur ces mécanismes évolutifs. Ces règles sont encore très peu étudiées. Pourtant la plupart des méthodes *in silico*, que je décrirai dans le chapitre suivant, s'appuient sur elles pour détecter des ARNnc.

LA DETECTION *IN SILICO* DES ARNNC BACTERIENS

La problématique de détection massive des ARNnc a émergé au début des années 2000 de la constatation que de plus en plus d'ARNnc étaient découverts chez les eucaryotes et les procaryotes, mais toujours de manière ponctuelle dans le cadre d'études particulières, et que leur nombre, certainement plus important, nécessitait une accélération du processus de détection. Or les ARNnc ne disposent pas de caractéristiques simples à identifier comme cela peut être le cas avec les gènes codants qui sont tous constitués d'une organisation régulière en codons. Toutefois, dès 2001, deux méthodes de détections des ARNnc sont publiées. La première combine la recherche par bioinformatique de régions intergéniques conservées dans des espèces proches et des données de microarray[99]. La seconde, appelée QRNA et développée par Rivas et Eddy[100], est uniquement *in silico* et sera le précurseur de la majorité des méthodes bioinformatiques qui verront le jour dans les années suivantes. Ces deux méthodes s'appuient sur la généralisation du séquençage des génomes qui permet des études comparatives entre génomes et la détection de régions conservées potentiellement fonctionnelles. Cependant, QRNA s'appuie sur une autre caractéristique qui semble commune à la majorité des ARNnc, la conservation des structures. En effet nous avons vu précédemment que la majorité des ARNnc se structurait et que ces structures devaient être conservées entre des espèces proches puisqu'elles paraissent indispensables à la fonction de ces ARNnc. La plupart des méthodes *in silico* développées depuis 2001 se sont inspirées de ces observations. Cependant toutes ces méthodes présentent des limites importantes en termes de temps de calcul et sont fondées sur un *a priori* fort qui est la conservation des structures. D'autres méthodes, utilisant des caractéristiques différentes ont été développées en parallèle pour tenter d'échapper à ces restrictions.

Dans les chapitres suivants, je me propose de décrire les méthodes fondées sur la conservation de structure puis celles, alternatives, qui utilisent chacune des caractéristiques propres. Je ferai l'impasse pour le moment sur les méthodes de détection moléculaire puissantes développées au cours des dernières années mais j'y reviendrai dans la discussion.

LES METHODES DE DETECTION BASEES SUR LA CONSERVATION DE STRUCTURE

A l'instar de QRNA, toutes ces méthodes s'appliquent à rechercher des signaux de structures conservées dans les séquences génomiques, qu'elles interprètent ensuite comme des preuves de la présence d'ARNnc. Dans ces outils, la conservation dans plusieurs génomes est essentielle car la structure seule ne constitue pas un signal suffisant depuis qu'il a été montré que des séquences d'ARNnc permutées aléatoirement présentent une énergie de repliement moyenne semblable à celle d'ARNnc natifs[101]. Cette conservation des structures suggère qu'elles sont soumises à une pression de sélection, et sont donc probablement fonctionnelles.

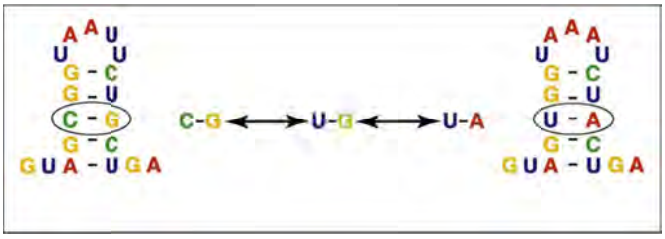


Figure 9. Les mutations compensatoires permettent de conserver la stabilité de la structure secondaire mais augmentent la variabilité de la séquence primaire. Ainsi la paire C-G, après la mutation du C vers un U va former la paire wobble U-G, dite intermédiaire, légèrement moins stable. Cette paire pourra à son tour être modifiée par un second évènement de mutation pour rétablir une structure plus stable en formant la paire initiale C-G ou la nouvelle paire U-A. Cette figure est adaptée de Gorodkin et al[102].

Il est couramment admis que la conservation de la séquence primaire seule n'est pas suffisante pour caractériser un ARNnc car ces molécules auraient une séquence qui évolue plus rapidement que leur structure. Les structures conservées ont, elles, tendance à accumuler entre des espèces éloignées des mutations compensatoires qui leur permettent de conserver leur topologie et leur stabilité (Figure 9). Ces doubles événements de mutation entraînant logiquement une évolution rapide de la séquence primaire sont aussi les marqueurs évolutifs d'une structure conservée d'ARNnc. En résumé, ces méthodes exigent que les éléments recherchés soient structurés et que ces structures soient conservées, mais aussi que les séquences primaires soient un minimum conservées pour affirmer qu'elles sont homologues et détecter des mutations compensatoires. Plusieurs séquences alignées sont donc nécessaires au fonctionnement de ces méthodes. Selon les cas, les séquences homologues seront alignées soit avant le repliement en structure secondaire soit après ou encore conjointement.

La première catégorie, qui nécessite un alignement préalable des séquences, regroupe les méthodes les plus utilisées à l'heure actuelle. QRNA se base sur des alignements deux à deux des séquences obtenu avec blastn[103]. Ces alignements sont analysés avec trois modèles probabilistes. L'un correspond aux structures d'ARN modélisées par une grammaire stochastique hors contexte qui décrit les mutations compensatoires et donc la contrainte évolutive sur les structures. Le deuxième correspond aux gènes codants modélisés par un HMM pour paire qui décrit les substitutions synonymes dans les codons. Enfin le troisième modélise les séquences restantes par un HMM pour paire qui considère l'hypothèse nulle et donc que les mutations sont indépendantes de la position. Ainsi, chaque alignement deux à deux est classé en RNA, CODING ou OTHER. Il faut remarquer que QRNA ne prend pas en compte la stabilité du repliement, lequel, couplé à la conservation, peut être un critère puissant de discrimination, comme le prouve la méthode RNAz[104]. Ce logiciel, à l'origine développé pour la détection d'ARNnc dans les génomes eucaryotes, mais maintenant utilisé dans tous les phylums[105-107], se concentre sur le calcul de l'énergie de repliement de la structure consensus obtenue à partir d'un alignement multiple de séquences primaires. Cette énergie est comparée à l'énergie de repliement obtenue après la permutation aléatoire des bases dans les colonnes de l'alignement multiple. RNAz mesure également la conservation structurale en introduisant la notion de mutation compensatoire. Finalement, grâce à un classificateur SVM, RNAz prédit si un alignement multiple contient un segment d'ARNnc. Cet outil de prédiction des ARNnc est aujourd'hui certainement le plus utilisé. Les outils de prédiction les plus récents qui se basent aussi sur des alignements multiples de séquences primaires sont Pfold[108] et Evofold[109]. A la différence de QRNA et RNAz ils ajoutent à leur modèle probabiliste l'information phylogénétique qui relie les séquences individuelles dans l'alignement selon des critères comme la structure de l'arbre ou la longueur des branches. Ainsi, ces programmes comparent deux modèles, un correspondant à des séquences non-structurées et un correspondant à des ARN structurés, pour décider celui qui correspond le mieux à l'alignement multiple tout en respectant la phylogénie.

QRNA, RNAz, Pfold et Evofold utilisent tous des fenêtres pour parcourir les alignements multiples, ce qui peut induire des problèmes de prédiction aux bords des fenêtres et entraîner de mauvaises interprétations pour des ARNnc qui seraient d'une taille plus grande que la fenêtre. De plus ces méthodes s'appuient toutes sur des alignements. Si la conservation des séquences est trop

grande il y sera difficile de détecter des mutations compensatoires. A l'inverse si la conservation est trop faible il sera difficile de faire coïncider les structures de chacune des séquences et par conséquent d'y détecter des mutations compensatoires. Mais le plus gros défaut de ces méthodes reste qu'elles partent d'alignements multiples produits par des logiciels comme ClustalW[110] ou Muscle[111], qui se révèlent fréquemment incorrects, d'autant plus que les éléments sont souvent peu conservés,. Toutefois, en exploitant des alignements multiples, ces méthodes permettent de gagner un temps de calcul important par rapport à des méthodes cherchant à produire la structure simultanément à l'alignement.

En effet certaines méthodes permettent d'éviter les désavantages liés aux alignements multiples de séquences classiques par le traitement conjoint de la structure et de la séquence, mais nécessitent en comparaison des temps de calcul beaucoup plus important qui en limitent aujourd'hui l'utilisation. Les méthodes comme FOLDALIGN[112], Dynalign[113] et LocARNA[114] qui sont basées sur l'énergie de repliement s'appuient toutes sur l'algorithme de Sankoff[115] que je ne détaillerai pas ici. La communauté de chercheurs, qui implémente ces méthodes est très active, et des solutions pour les accélérer sont prévisibles. En effet leur capacité à équilibrer l'apport de la structure et de la séquence dans un alignement devrait aboutir à de meilleurs résultats.

La dernière catégorie de méthodes qui s'attaque au problème de la détection d'ARNnc d'un point de vue structural, est uniquement représenté par l'outil CMfinder[116] qui est aussi souvent utilisé pour trouver des ARNnc homologues à des ARNnc déjà connus. CMfinder a permis, par exemple, l'analyse des séquences humaines conservées produites par le projet ENCODE[117] pour y détecter des motifs structuraux d'ARNnc[118]. Dans ce travail, les auteurs débutent par l'analyse de toutes les séquences alignées avec MULTIZ[119], qui n'aligne pas les séquences primaires mais recherche le meilleur ordonnancement des orthologues, pour y détecter des motifs d'ARN structurés. A partir des résultats de cette première étape, ils construisent les alignements locaux de ces motifs et ils en infèrent des modèles de covariance qu'ils raffinent par étapes successives en même temps que les alignements locaux. Cette méthode est donc totalement libérée des contraintes liées aux fenêtres de « scan » et à un alignement multiple préalable. Cependant, afin de limiter le temps de calcul très important il est nécessaire de paramétrer CMfinder à l'avance en indiquant la complexité des motifs structuraux recherchés (nombre de tiges-boucles).

Toutes ces méthodes partent du principe qu'un ARNnc est structuré et que cette structure est conservée car fonctionnelle. Cet *a priori* s'appuie essentiellement sur des observations empiriques, même si les résultats obtenus avec ce type d'outil, et validés expérimentalement, confirment dans de nombreux cas cette hypothèse. Néanmoins il existe des ARNnc espèce-spécifiques dont la séquence n'est pas conservée, comme il existe des ARNnc dont la structure n'est pas conservée d'une espèce à l'autre. De plus, la plupart de ces méthodes utilisent des algorithmes de repliement d'ARN en structure secondaire exigeants en temps de calcul, qui rendent impossible leur application à l'ensemble des génomes séquencés. Enfin on peut remarquer que le recouvrement des résultats entre méthodes semble faible, et ce plus particulièrement entre RNAz et Evofold qui ont été comparé par l'équipe de Peter Stadler[106]. Sachant qu'elles possèdent pourtant chacune une bonne spécificité, on peut se demander si les critères utilisés sont tous pertinents.

Tous ces outils de prédiction ne prennent pas en compte d'autres caractéristiques que l'on peut qualifier de communes aux ARNnc au même titre que la conservation de structure, telles que la présence de terminateurs Rho-indépendants et de sites de fixation à des facteurs de transcription dans les régions intergéniques. D'autres méthodes qui s'écartent complètement des modes d'analyses précités ont été développées afin de prendre en compte ces autres caractéristiques des ARNnc tout en permettant une application à très large échelle en un temps raisonnable.

LES AUTRES METHODES DE DETECTION D'ARNNC

Ces méthodes, peu nombreuses, se sont développées en grande partie grâce à l'annotation de plus en plus détaillée des génomes séquencés dans les années 2000. En effet elles reposent toute sur un principe simple soustractif. Elles considèrent que l'annotation des gènes codants de ces génomes est quasi exhaustif et que par conséquent dans les régions intergéniques, localisation de la majorité des ARNnc agissant en *trans*, des signaux partagés par les gènes codants et non-codants comme les terminateurs Rho-indépendants ou les sites de liaison aux facteurs de transcription seront très probablement associés à la présence d'un gène non-codant. On peut citer

la méthode développée dès 2005 par Livny, appelée initialement sRNApredict[120] puis améliorée récemment et renommée SIPHT[121]. Cette méthode se fonde sur de nombreux critères appliqués aux régions intergéniques extraites des génomes : présence de terminateurs Rho-indépendants, présence de sites de liaison aux facteurs de transcription, conservation de la séquence, présence d'une structure conservée validée par QRNA, conservation de la synténie avec des ARNnc déjà connus et, enfin, similitude des séquences avec des ARNnc connus. Ainsi par la calibration de la méthode sur un jeu d'ARNnc et une exécution séquentielle ingénieuse avec des conditions d'arrêt à certaines étapes, cette suite de critères permet d'obtenir un grand nombre de prédictions avec une bonne sensibilité en un temps raisonnable. La procédure est totalement automatisée et Livny et al ont pu publier en 2008 un article répertoriant 47263 prédictions de segments d'ARNnc pour 932 génomes bactériens[121].

Une autre méthode, plus marginale car pour le moment uniquement appliquée dans le cas précis de la recherche d'ARNnc chez *Staphylococcus aureus*, se fonde sur des principes similaires, quoique utilisant des critères de discrimination en partie différents et moins nombreux. Cette méthode développée par Pichon et al[122] recherche des îlots riches en CG dans les régions intergéniques de *S. aureus* riches en AT présentant au moins une faible conservation dans d'autres firmicutes. Elle vérifie ensuite la présence de promoteurs et de terminateurs Rho-indépendants au voisinage de ces régions candidates. Faute de test sur des ARNnc de référence, il est difficile d'évaluer cette méthode, mais elle aura néanmoins permis d'identifier 12 nouveaux ARNnc chez *S.aureus*.

Les méthodes de génomique comparative comme celles reposant sur la recherche de structures conservées ne peuvent pas détecter d'ARNnc espèce-spécifiques. Cependant elles s'en approchent plus que les précédentes car leur critère de conservation minimum est souvent très faible, en utilisant blastn avec des seuils d'e-value compris entre 0,01 et 0,1. Un autre point à souligner est que l'utilisation de filtres multicritères qui fondent ces méthodes est à la fois leur force et leur faiblesse. En effet, en évitant de limiter la définition d'un ARNnc à une structure conservée, la multiplicité des critères rend le calibrage compliqué et exige, pour conserver une bonne sensibilité, une baisse importante de la spécificité.

Toutes les méthodes que je viens de décrire partagent des limites communes. En effet elles ne peuvent pas détecter les ARNnc antisens d'un ARNm puisque, ne disposant pas de critère discriminant pour ce type d'élément, elles n'analysent que les régions intergéniques. De plus ces méthodes *in-silico* comme la plupart des méthodes de détection moléculaire massive ne fournissent aucun indice sur la fonction de ces ARNnc. La constatation des limites communes et spécifiques de ces méthodes nous a conduits dans le cadre de ma thèse à tenter de développer une méthode qui pourrait contourner au moins partiellement ces limites et proposer de nouveaux signaux spécifiques des ARNnc. Cette méthode, appelée « profilage phylogénétique » est déjà utilisée pour la prédiction de fonction des protéines à partir de leur séquence en acides aminés. Notre adaptation de cette méthode à l'analyse de séquences nucléiques intergéniques et la détection d'ARNnc bactériens a donc logiquement l'avantage de pouvoir dans le meilleur des cas nous fournir aussi des indices sur leur fonction. Se fondant sur l'étude phylogénétique des séquences, cette méthode se distingue des autres par son absence d'*a priori* concernant la stabilité et la structuration des ARNnc, le seul présupposé étant la localisation intergénique et un minimum de conservation de la séquence primaire. De plus, les phylogénies des ADN non-codants produits par notre méthode devraient pouvoir nous permettre d'analyser la manière dont évoluent plus spécifiquement ces ARNnc.

ARTICLES

SINGLE-PASS CLASSIFICATION OF ALL NONCODING SEQUENCES IN A BACTERIAL GENOME USING PHYLOGENETIC PROFILES.

Antonin Marchais, Magali Naville, Chantal Bohn, Philippe Bouloc, and Daniel Gautheret.
Published in Genome Research, 2009.[123]

RESUME

La détection *in silico* des ARN non-codants s'est développée ces dernières années avec l'implémentation de nombreuses méthodes s'appuyant pour la majorité sur des critères de conservation de séquence, de structure et sur la présence de co-variations de nucléotides distants entre espèces proches. En plus des biais induits par les critères utilisés, ces méthodes nécessitent une expertise importante de la part de l'utilisateur et leur besoin en temps de calcul ne leur permet pas d'analyser la grande quantité de données produite actuellement par le séquençage massif de génomes bactériens. Par ailleurs toutes ces méthodes éludent l'annotation fonctionnelle de leur prédiction. Le « profilage phylogénétique d'acide nucléique » (NAPP), que nous avons développé, pallie à la plupart de ces limitations. En effet, totalement automatisé, il ne nécessite pas l'intervention d'un spécialiste ni d'importants moyens de calcul, et peut dans le meilleur des cas informer sur une fonction putative des ARNnc prédits. Cette méthode, à l'origine développée pour prédire la fonction de protéines de fonction inconnue et donc appliquée à des séquences d'acides aminés, se fonde sur l'hypothèse que des gènes impliqués dans un même processus cellulaire sont interdépendants et que par conséquent la pression de sélection aura tendance à co-conserver ces gènes dans un ensemble d'organismes afin de maintenir l'intégrité du processus.

Cette méthode nécessite donc la création d'un index, appelé aussi profil phylogénétique, qui répertorie la présence et l'absence de chaque gène dans un jeu d'organismes donné. L'adaptation de cette méthode aux acides nucléiques a exigé le développement d'une première étape d'extraction des séquences conservées (même très faiblement) dans les régions intergéniques (CNE, pour « Conserved Non-coding Elements ») des 422 génomes séquencés référencés à la date du début du projet. L'étape suivant est la construction des profils phylogénétiques des CNE d'une

espèce. Elle consiste à chercher pour chaque CNE de cette espèce de référence, grâce à l'outil blastn, des séquences similaires dans les 421 espèces restantes. Ainsi, pour un CNE donné, nous répertorions le meilleur « bit score » fournit par blast pour chaque espèce, que l'on normalise par le « bit score » maximum obtenu en blastant le CNE contre lui-même. Nous obtenons au final pour chaque CNE un profil phylogénétique constitué de valeur continue illustrant sa conservation dans chaque espèce.

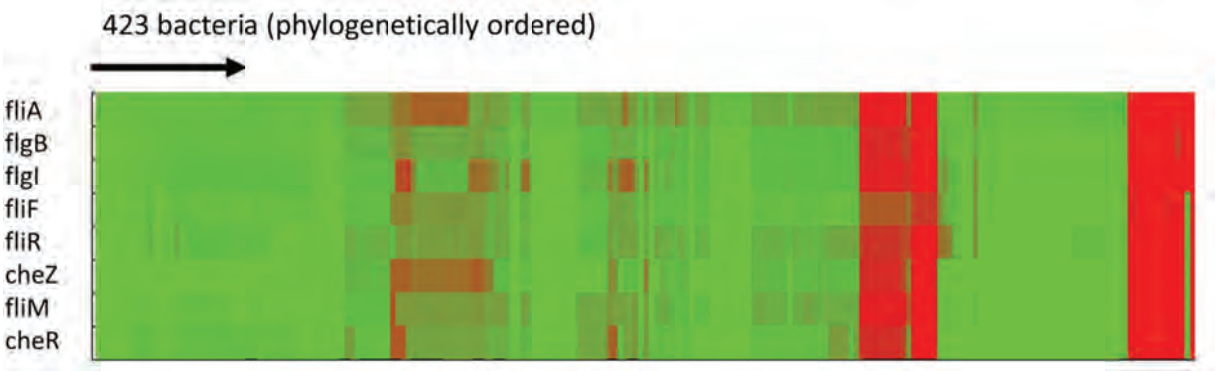


Figure 10. Profils phylogénétiques de 8 gènes impliqués dans la motilité chez *E. coli*. En abscisse, sont indiqués les 8 gènes correspondant au profil. En ordonnée, chaque barre représente une des 423 bactéries dans lesquels nous avons cherché à déterminé la conservation de chacun des 8 gènes. Le gradient de couleur illustre la conservation qui augmente du vert jusqu'au rouge. Les espèces sont ordonnées phylogénétiquement. Les profils de ces 8 gènes impliqués dans la mise en place de la motilité sont très similaires et nous pouvons en conclure qu'ils co-évoluent. Si le profil d'un gène de fonction inconnue se regroupe avec les profils de ces gènes de fonction connue nous pouvons inférer qu'il co-évolue avec et que par conséquent il est lui-même impliqué dans la motilité.

Nous avons ensuite réalisé les mêmes profils pour les gènes codant pour des protéines et nous avons classifié l'ensemble des profile en fonction de leur similitude. Cette classification nous permet en principe de détecter la co-occurrence des gènes et des CNE dans des espèces communes. Quelque-soit l'espèce de référence choisie, ce regroupement par la méthode des K-means produit toujours un ou deux groupes particulièrement intéressants qui réunissent une majorité de CNE correspondant à des ARNnc avec des gènes le plus souvent de ménage et, pour un sous-ensemble, impliqués dans l'interaction, la régulation ou la modification des ARN. Cette observation nous a permis d'utiliser ces groupes pour prédire 48 ARNnc et en valider expérimentalement sept chez *S. aureus*. Sur les trois espèces modèles *E. coli*, *B. subtilis* et *V. cholerae*, la capacité de détection de NAPP est comparable à celle des méthodes les plus utilisées chez les bactéries que sont RNAz[104] ou SIPHT[121]. Cependant NAPP fournit une information

supplémentaire sur la phylogénie de ces éléments et sur leur co-occurence avec des gènes de fonction connue. Ainsi nous pouvons espérer associer à un CNE un groupe de protéines et en inférer son niveau d'implication dans la cellule ou dans le meilleur des cas sa fonction dans un processus spécifique.

Downloaded from genome.cshlp.org on September 8, 2010 - Published by Cold Spring Harbor Laboratory Press

Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles

Antonin Marchais, Magali Naville, Chantal Bohn, Philippe Bouloc, and Daniel Gautheret[1]

Université Paris-Sud 11, CNRS, UMR8621, Institut de Génétique et Microbiologie, F-91405 Orsay Cedex, France

Identification and characterization of functional elements in the noncoding regions of genomes is an elusive and time-consuming activity whose output does not keep up with the pace of genome sequencing. Hundreds of bacterial genomes lay unexploited in terms of noncoding sequence analysis, although they may conceal a wide diversity of novel RNA genes, riboswitches, or other regulatory elements. We describe a strategy that exploits the entirety of available bacterial genomes to classify all noncoding elements of a selected reference species in a single pass. This method clusters noncoding elements based on their profile of presence among species. Most noncoding RNAs (ncRNAs) display specific signatures that enable their grouping in distinct clusters, away from sequence conservation noise and other elements such as promoters. We submitted 24 ncRNA candidates from *Staphylococcus aureus* to experimental validation and confirmed the presence of seven novel small RNAs or riboswitches. Besides offering a powerful method for de novo ncRNA identification, the analysis of phylogenetic profiles opens a new path toward the identification of functional relationships between co-evolving coding and noncoding elements.

[Supplemental material is available online at www.genome.org.]

In all living organisms, the genome regions located between protein-coding sequences are home to a wide diversity of functional elements that include noncoding RNA (ncRNA) genes, DNA regulatory elements, untranslated regions (UTRs) of genes, transposable and self-replicating elements, and a variety of other transcribed or nontranscribed functional sequences. As these elements are often key players in gene regulation and thus in the global cell interaction network, their systematic identification and characterization has become a major challenge in biology.

Computational protocols developed to collect and characterize noncoding elements in genomic sequences rely, to a large extent, on comparative genomics. The most common strategies involve, first, collecting sequences under selective pressure and, second, analyzing the aligned sequences using various classifiers that exploit criteria such as nucleotide composition, folding potential, fold conservation, or covariation between distant positions (Rivas and Eddy 2001; Washietl et al. 2005; Pedersen et al. 2006; Torarinsson et al. 2006). In general, such classifiers are designed to detect structured RNAs among noncoding elements with no further distinction between regulatory elements, repeats, or artifacts produced by sequence comparison algorithms.

Comparative genomics entails a significant amount of expert intervention, especially in obtaining the right genome set to optimize the specificity and sensitivity of RNA detection. Although a number of studies have successfully identified ncRNAs in several animal (Missal et al. 2005; Washietl et al. 2007) and microbial genomes (Altuvia 2007), the pace at which such studies are performed and published lags far behind the rate of genome sequence output. Most of the complete genomes sequenced to date have escaped the scrutiny of RNA experts, and their potential for novel RNA functions lies unexplored. Recently, Livny et al. (2008) introduced an automated procedure, SIPHT, which combines conserved sequence detection and the presence of adjacent Rho-independent terminators. The procedure is sufficiently automated to be applied to all available bacterial genomes; however, the requirement for a terminator motif introduces a bias against the significant fraction of RNA elements that are not followed by a detectable terminator.

Nucleic acid phylogenetic profiling (NAPP) addresses both the issue of high-throughput noncoding sequence identification and that of their functional characterization. Phylogenetic profiling (Ragan and Gaasterland 1998; Pellegrini et al. 1999) posits that genes belonging to the same metabolic or regulatory pathway tend to occur concomitantly in a given set of organisms. By clustering proteins based on their occurrence profile in many species, one therefore obtains clusters of functionally related proteins, which enables functional assignments to uncharacterized sequences. While this principle has been successfully applied to protein annotation (Pellegrini et al. 1999; Enault et al. 2003; Srinivasan et al. 2005), to date, no application to ncRNA annotation has emerged. An obstacle to building nucleic acid profiles over a large evolutionary time frame is certainly the lower sensitivity of DNA sequence similarity searches compared to searches at the amino acid sequence level. However, in spite of this limitation, Janky and van Helden (2008) were able to obtain clear phylogenetic profiles for promoter DNA elements, by focusing their analysis on upstream sequences of orthologous genes. In this study, we apply phylogenetic profiling for the first time to the complete noncoding DNA in bacterial genomes, with the objective of classifying all conserved noncoding elements in a single computing process.

Results

Collecting noncoding conserved elements

To collect loosely conserved noncoding elements in a selected reference genome, we performed a BLAST search of all noncoding regions against a set of 422 complete bacterial genomes and

[1]Corresponding author.
E-mail daniel.gautheret@u-psud.fr; fax 33-4-1-69-15-46-29.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.089714.108.

19:1084–1092 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

Downloaded from genome.cshlp.org on September 8, 2010 - Published by Cold Spring Harbor Laboratory Press

obtained a conservation index for each intergenic position. Note that "intergenic" here also comprises gene UTRs. The conservation index represents the number of genomes in which a BLAST hit was recorded at this position, weighted by the phylogenetic proximity of each genome with the reference genome (see Methods). For each noncoding region, we produced a conservation histogram (Fig. 1), and extracted all conserved noncoding elements (CNEs) using empirical conservation index cutoffs, fragment sizes, and minimal distances between conserved fragments. An overview of the resulting CNEs in *Escherichia coli* and *Bacillus subtilis* is presented in Table 1. At the selected cutoffs, CNEs detected 76%–87% of all known ncRNAs in these two species. Examples of CNEs (Fig. 1) map to known noncoding elements, including small RNAs (Fig. 1A,B,D,E,G), a riboswitch (Fig. 1H), RNase P RNA genes (Fig. 1C,F) and the CRISPR locus (Fig. 1I). While CNEs do not strictly map the boundaries of RNA genes or elements, they generally overlap functional, and hence the most conserved regions, such as target- or substrate-binding regions in small RNAs (sRNAs) and riboswitches (black areas in Fig. 1A–E). In the CRISPR loci, each repeat is accurately represented by an independent CNE (Fig. 1I). The vast majority of CNEs, however, do not map any known ncRNA (3427 in *E. coli* and 2653 in *B. subtilis*). It is the classification of these heterogeneous elements that we will now address using phylogenetic profiling.

Phylogenetic profiling efficiently clusters ncRNAs

In order to obtain the phylogenetic profiles of CNEs in the reference species, we locally realigned each CNE against 422 complete bacterial genomes and retained the best hit in each genome in the form of a size-normalized BLAST score (Enault et al. 2003). As we were interested in comparing the profiles of noncoding elements to those of coding elements, we also generated profiles for each coding DNA sequence (CDS) in the reference genome using the same procedure. Each CNE/CDS was thus attributed a vector of 422 values between 0 and 1 (0: no hit; 1: perfect match). We computed a distance matrix by measuring Pearson's distances between all pairs of vectors and used this matrix for K-means clustering. To assess clustering quality, we computed

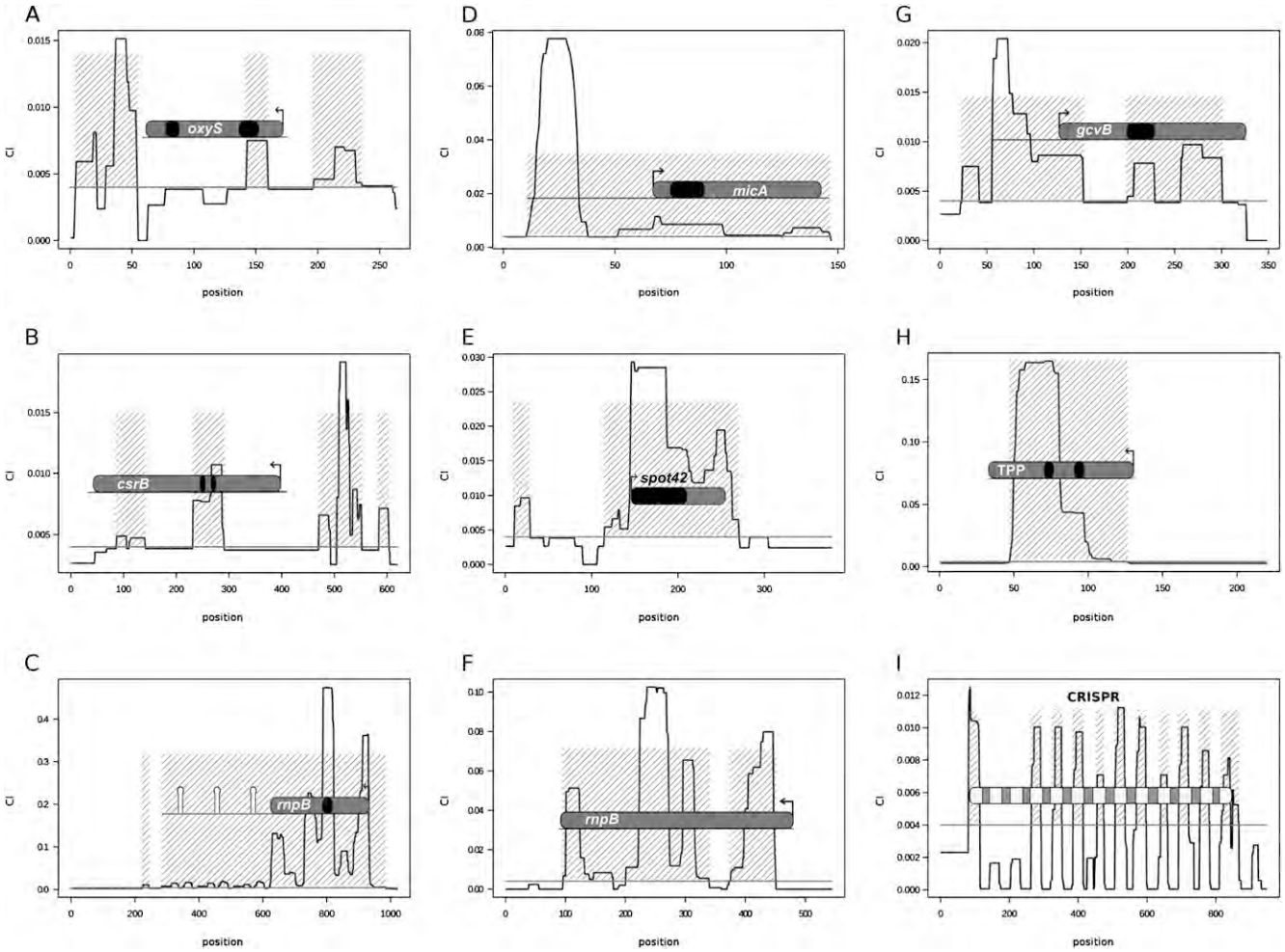


Figure 1. Conservation profiles in selected bacterial noncoding regions. The *vertical* axis represents the conservation index (CI) and the *horizontal* axis represents genome coordinates in relative numbering. The CI threshold used for defining conserved noncoding elements (CNEs) is represented by a horizontal line at CI = 0.004. CNEs are denoted by shaded areas. Color code for RNA elements: RNA genes are gray with black areas representing known target-binding regions, CRISPR spacers are gray and CRISPR repeats are light gray. Transcription initiation sites are shown with broken arrows and terminators with hairpins. (*A,B,D,E,G*) *E. coli* regulatory sRNA genes; (*H*) *E. coli* TPP riboswitch; (*C*) *E. coli* RNase P RNA gene; (*F*) *B. subtilis* RNase P RNA gene; and (*I*) *E. coli* CRISPR locus.

Downloaded from genome.cshlp.org on September 8, 2010 - Published by Cold Spring Harbor Laboratory Press

Table 1. CNE statistics

	E. coli	*B. subtilis*
Number of CNEs	3483	2714
Min/max size (nt)	15/701	15/284
Mean size (nt)	39	34
Fraction of known ncRNA captured (excl. tRNA/rRNA)	56/74	61/70

the enrichment of clusters for known ncRNAs (excluding tRNA and rRNA), promoters, terminators, and transcription factor binding sites. Figure 2 presents a two-dimensional representation of K-means clusters of all nucleic acids elements in *E. coli* and *B. subtilis*, showing cluster enrichment in known ncRNAs. Enrichments in other functional elements are shown in Supplemental Figure S1 for *E. coli*. While some clusters are mildly enriched in promoters and terminators ($P \sim 10^{-3}$; Supplemental Fig. S1), enrichment in ncRNA is extremely significant: between 54% and 79% of all known ncRNAs are grouped into only two clusters in both *E. coli* and *B. subtilis* ($P = 10^{-4}$ to 10^{-15}; Fig 2).

Inclusion of protein-coding genes in profiles improves cluster quality with respect to ncRNA enrichment. When CNEs are clustered alone, the best ncRNA enrichment obtained in *E. coli* is not as significant as when protein-coding genes are present ($P = 2 \times 10^{-7}$; Supplemental Fig. S2). Moreover, the efficient ncRNA clustering is not caused by the higher conservation or longer size of ncRNAs compared to other elements. When performing K-means clustering of CNEs based on their size (Supplemental Fig. S3) or conservation level (Supplemental Fig. S4), the highest ncRNA enrichment observed are $P = 4 \times 10^{-3}$ and $P = 10^{-2}$, respectively. This shows that information contained in a nucleic acid phylogenetic profile is significantly richer than size or conservation alone.

Performance of NAPP as an ncRNA classifier

NAPP is as efficient in bacterial ncRNA identification as specialized programs. We compared NAPP to RNAz (Washietl et al. 2005) and SIPHT (Livny et al. 2008). RNAz classifies multiple sequence alignments into ncRNA/non-ncRNA based on the detection of significant conserved secondary structure and base pair (bp) covariation. It has been applied to several organism families, including bacteria, where it led to experimental confirmation of novel ncRNAs (del Val et al. 2007; Panek et al. 2008; Sonnleitner et al. 2008). We compared NAPP clustering and RNAz in their abilities to distinguish true ncRNAs from other conserved elements. RNAz recommends input alignments containing at least four sequences with a minimal size of 50 nucleotides (nt) and less than 25% gaps. After realigning all *E. coli* CNEs against their best hits in other genomes, we obtained 512 alignments meeting these criteria, of which 23 correspond to known ncRNAs. Figure 3A compares predictions of NAPP (pink) and RNAz (blue) in four bacterial species. Both methods identify nearly the same fraction of known ncRNAs (yellow). NAPP appears more specific, as it produces fewer predictions in general. However, as complete sets of ncRNAs are unknown in either species,

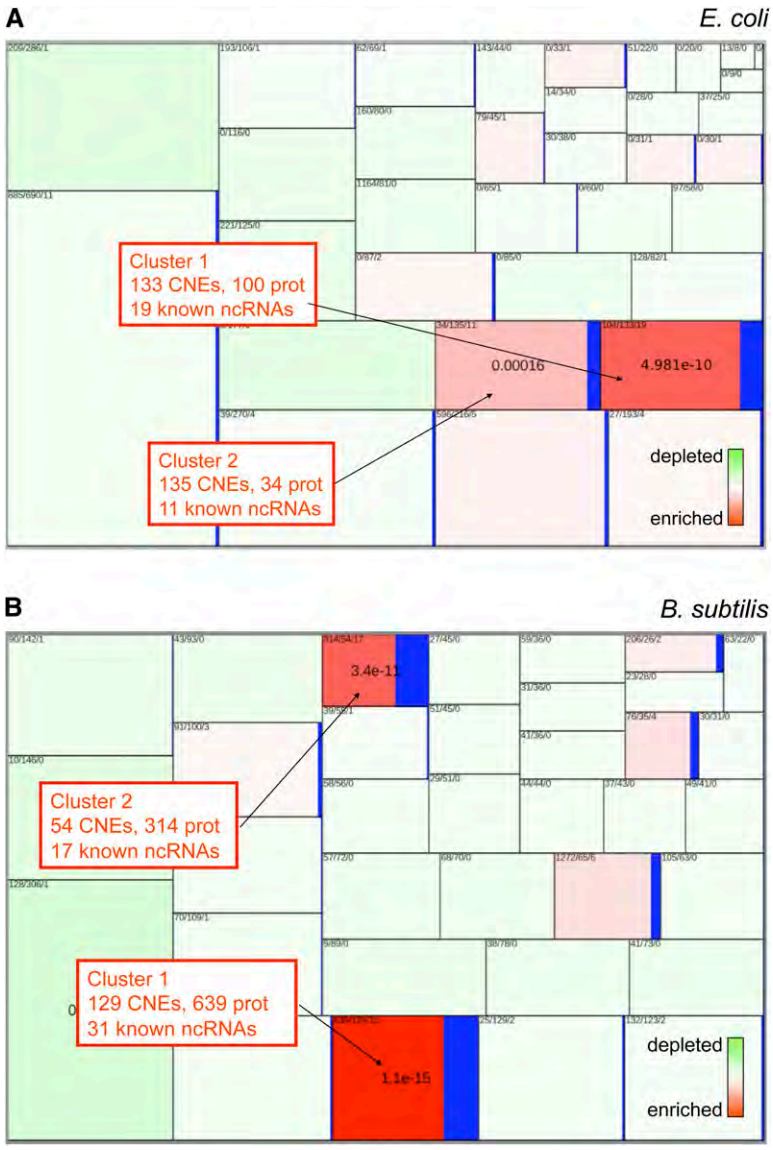


Figure 2. K-means clustering of CNEs and CDSs in (*A*) *E. coli* and (*B*) *B. subtilis*. A rectangle represents a cluster of CNEs/CDSs, and its size is proportional to the number of CNEs in the cluster. Numbers in each rectangle represent numbers of protein coding genes, CNEs and known ncRNA in this cluster, respectively. The color scale represents the enrichment/depletion of known ncRNA in each cluster. For ncRNA-rich clusters, the enrichment *P*-value is shown in the *center* of the rectangle. ncRNA enrichment measures do not include tRNA or rRNA.

Downloaded from genome.cshlp.org on September 8, 2010 - Published by Cold Spring Harbor Laboratory Press

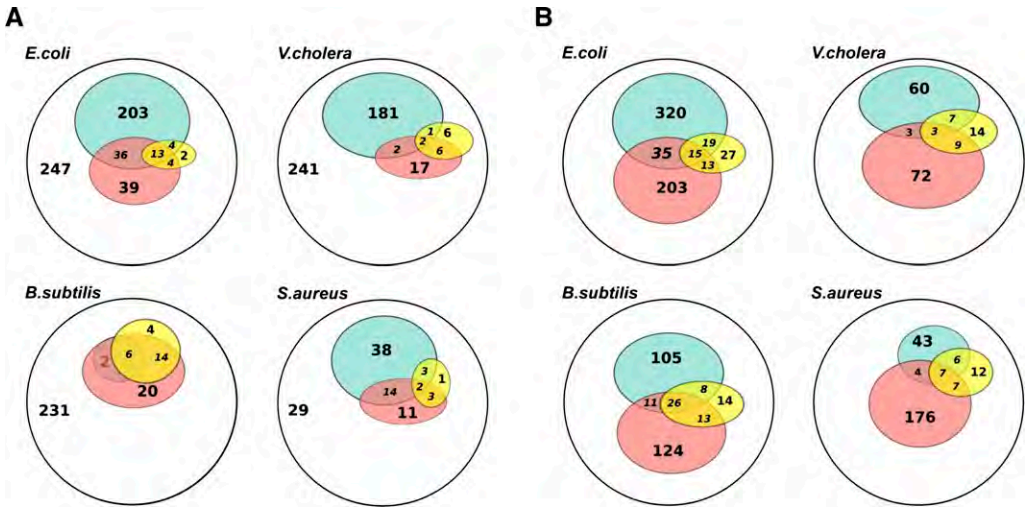


Figure 3. Comparison of NAPP ncRNA predictions with RNAz and SIPHT predictions in four bacterial genomes. Numbers in Venn diagram areas represent the area exclusive of included subareas. Blue: ncRNA predictions from RNAz or SIPHT; pink: ncRNA predictions from the NAPP program; and yellow: known ncRNAs present in the tested set (excluding tRNA, rRNA). (*A*) Predictions from the RNAz program. White areas represent total numbers of candidate sequences that can be tested with RNAz. (*B*) Prediction from the SIPHT program. Both SIPHT and NAPP screen all conserved intergenic elements.

the larger RNAz prediction sets do not necessarily correspond to false positives. Figure 3B compares NAPP to SIPHT, an ncRNA detection procedure based on intergenic sequence conservation and the presence of transcription terminators (Livny et al. 2008). When tested on the four organisms, the two programs show similar performances in terms of sensitivity and apparent specificity. These results show that NAPP clustering of ncRNAs compares favorably with two recent, specialized ncRNA prediction systems. Known ncRNAs predicted by the three methods present little overlap (Supplemental Table S1). Indeed, each program uses completely different criteria to classify conserved noncoding elements. As NAPP predictions are independent of conserved secondary structures (used in RNAz) and presence of a terminators (used in SIPHT), they are likely to contain a large fraction of ncRNA genes or elements that escaped previous scrutiny.

ncRNAs display distinctive phylogenetic signatures

Sequence comparisons are less sensitive when performed on nucleic acids than when performed on proteins and, consequently, whereas protein phylogenetic profiles often span hundreds of bacterial species, nucleic acids profiles span a much shorter evolutionary distance. In spite of this limitation, however, a large fraction of the nucleic acid profiles have very distinctive features when examined on shorter timescales. Figure 4 and Supplemental Figure S6 present a sample of K-means clusters represented as heat-maps. A histogram representing the average phylogenetic profile is superimposed onto each cluster (histograms for the full set of 35 K-means clusters are shown in Supplemental Fig. S7).

Most NAPP clusters that are not functionally enriched have a narrow signature and a significant background noise visible in the form of red "dust" (e.g., see Fig. 4D). This dust occurs across the taxonomic spectrum, and likely corresponds to nonsignificant similarities. In this respect, profile clustering is an efficient means to condense the bulk of nonsignificant similarities usually produced by comparative genomics protocols into "garbage clusters" that can be singled out. Garbage clusters comprise the majority of CNEs and few CDSs. Most CDSs cluster in distinct groups that contain few CNEs and display relatively narrow signatures. The narrow signature of CDSs may be surprising at first sight because proteins are often conserved across a wide phylogenetic spectrum; however, as we consider here size-normalized DNA-level conservation, we did not expect protein-coding genes to be more conserved than any other functional DNA sequence. Indeed, conservation analysis in *E. coli* previously showed that sRNA genes and protein-encoding genes display similar degrees of conservation at the nucleotide level (Zhang et al. 2004). On the other end of the conservation spectrum, sequences with the broadest phylogenetic signature are tRNAs and rRNAs (Fig. 4E,F). In *E. coli*, most tRNAs are grouped into a single cluster (Fig. 4E) and most 5S, 16S, and 23S rRNAs are grouped into a second cluster together with ribosomal spacers, a few remaining tRNAs and some ribosomal protein genes (Fig. 4F). A similar pattern of tRNA and rRNA clustering is observed in other species (not shown).

The ncRNA-rich clusters (excluding tRNA and rRNA) often have intermediate signatures with a species distribution between that of rRNA and that of most coding sequences. One of the *E. coli* ncRNA-rich clusters contains elements present primarily in the *Salmonella* and *Escherichia* families (Fig. 4A), while the other cluster has a slightly broader signature extending to *Yersinia* (Fig. 4B). Most sequences in these two clusters have low-scoring putative homologs in more distant species such as *Pseudomonas* and *Haemophilus*, and have no visible homolog outside γ-proteobacteria. Not all ncRNA-rich clusters, however, have such signatures. In *E. coli*, one subcluster contains nine ncRNAs among a total of 21 CNEs, but nevertheless has a very narrow species distribution (Supplemental Fig. S6C). Similar distributions were seen for two ncRNA-rich clusters in *B. subtilis* and *Vibrio cholerae* (Supplemental Fig. S6A,E).

Functional analysis of clustered RNAs and CDSs

Noncoding RNA clusters contain both universal or kingdom-wide RNAs and the bulk of phylum-specific ncRNAs. For instance,

Downloaded from genome.cshlp.org on September 8, 2010 - Published by Cold Spring Harbor Laboratory Press

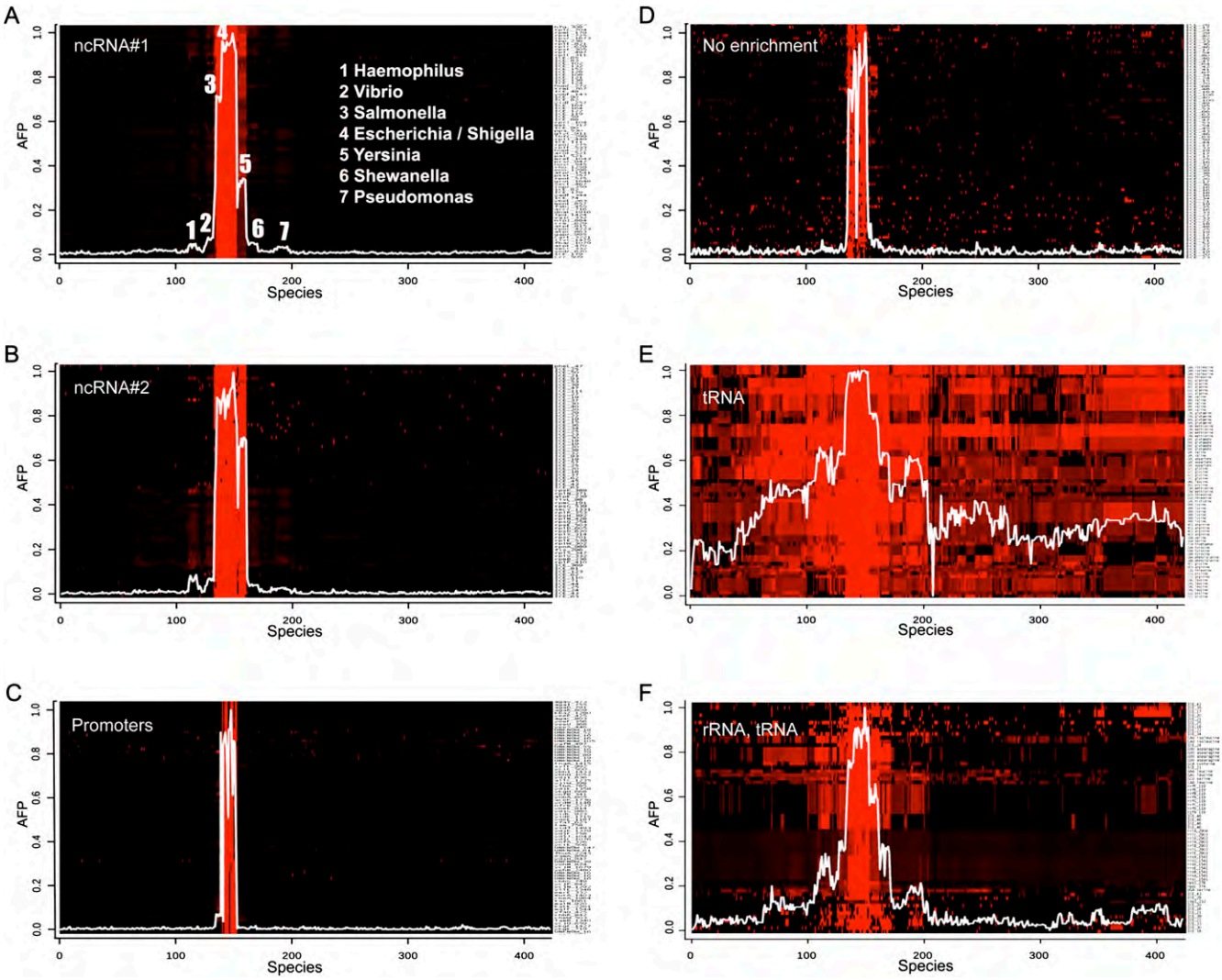


Figure 4. Heat-map representations of NAPP clusters in *E. coli* K12. Each panel shows a representative part of a K-means cluster from Figure 2A. Each K-means cluster was reclassified by hierarchical clustering and displayed as a heat-map using Eisen software (Eisen et al. 1998). The *horizontal* axis represents the 423 bacterial species ordered according to a 16S rRNA tree. The *right* vertical axis represents CNEs or CDSs. A red dot indicates the presence of a given CNE/CDS in this species with color intensity proportional to normalized BLAST scores. The histogram superimposed on each cluster represents the average phylogenetic profile for this cluster (AFP, *left* vertical axis) obtained by averaging dot intensities in columns. (*A*) ncRNA-rich cluster #1 from Figure 2A; (*B*) ncRNA-rich cluster #2 from Figure 2A; (*C*) a promoter-rich cluster; (*D*) a cluster with no apparent functional enrichment; (*E*) the main tRNA cluster; and (*F*) a cluster containing mostly tRNAs and ribosomal RNAs.

ncRNA clusters #1 and #2 in *E. coli* (Fig. 4A,B; Supplemental Table S2) contain enterobacterial RNAs, such as *sroB*, *csrB*, *csrC*, and *sraB*, together with universal RNAs such as RNase P, SRP RNA, and tmRNA, as well as TPP and cobalamin riboswitches that are present in all bacteria (Barrick and Breaker 2007). We observe the same pattern in other species (Supplemental Table S2). Therefore, the width of phylogenetic signatures is not a direct function of the age of an RNA family. Ancient RNAs that were present at the root of bacterial evolution may exhibit a variety of phylogenetic signatures ranging from relatively narrow (RNase P, SRP RNA, tmRNA) to medium (rRNA) or extremely wide (tRNA). This diversity of profiles reflects large differences in the evolutionary rates of different RNA families.

As expected from the distribution of known RNAs in bacteria, we observed mostly *cis*-regulatory elements, such as riboswitches and leader sequences in Gram positive bacteria, and mostly small RNAs in Gram negative bacteria. In *E. coli*, where annotated ncRNA include both sRNA genes and *cis*-regulatory RNA elements, we observed instances of both classes in all RNA clusters. Numbers of known ncRNAs are currently too low to evaluate clustering by more specific functional criteria, such as the involvement of a ncRNA in a type of metabolism. However, the mixture of universal ncRNAs and species-specific regulatory ncRNAs in the same clusters suggests that the primary driving force in ncRNA clustering is neither biological nor molecular function, but instead relates to the evolutionary rate of each of the RNAs in different species.

Supplemental Table S3 lists known *E. coli* RNAs that were missed by the NAPP procedure. Eighteen known RNAs fell below our conservation threshold and therefore were not fed into the phylogenetic profile screens. Another 16 RNAs are distributed among various NAPP clusters. This list includes a single ancient

Downloaded from genome.cshlp.org on September 8, 2010 - Published by Cold Spring Harbor Laboratory Press

RNA (6S) and a larger fraction of RNAs with unknown function than ncRNA-rich clusters (50% vs. 32%), suggesting that missed ncRNAs may be generally less reliable than predicted ncRNAs. Supporting this, four of the 16 missed RNAs (t44, c0343, c0465, is102) were detected in a single DNA microarray study (Tjaden et al. 2002).

We analyzed functional biases in the CDSs present in ncRNA clusters using KEGG-term enrichment (Kanehisa et al. 2002). In all species studied, at least one ncRNA cluster was enriched in major housekeeping functions such as "ribosome," "RNA polymerase," or "phosphorylative oxidation" (Supplemental Table S4). When comparing ncRNA clusters from distant species (*E. coli* and *B. subtilis*), we observed significant numbers of orthologous proteins in the "housekeeping" clusters ($P = 1.8 \times 10^{-2}$ to 1.5×10^{-11}; Supplemental Fig. S5). This suggests that a similar set of housekeeping proteins is consistently associated with ncRNAs in different species. These clusters generally had wider phylogenetic signatures (Supplemental Fig. S6B,D; Supplemental Table S4) than other ncRNA clusters. Other ncRNA clusters do not present KEGG-term enrichment or are enriched for other functions, and are associated with narrower signatures. For example, ncRNA-rich cluster #2 in *V. cholerae* is enriched in KEGG-term "chemotaxis" (Supplemental Fig S6E; Supplemental Table S4).

RNA-rich clusters in *E. coli* contain a high proportion of CDSs coding for RNA-related proteins and Hfq-bound mRNAs (Supplemental Table S5). A first striking observation is that many of the encoded proteins either (1) regulate sRNA expression (e.g., RpoE, YobF, CsrA), (2) are regulated by ncRNA (e.g., OmpR), (3) are part of ribonucleoprotein complexes (e.g., RnpA, ribosomal proteins), or (4) bind RNA (e.g., CspA). Overall, 64% of the ORFs in the two major ncRNA clusters in *E. coli* are RNA related. Interestingly, the Hfq protein, a central player in RNA-based regulations in many bacterial species, is present in ncRNA-rich clusters in all four species analyzed. In a recent study, Sittka et al. (2008) sequenced 727 mRNAs that co-immunoprecipitated with Hfq in *Salmonella*. These mRNAs are thought to involve a majority of ncRNA targets that recruit Hfq through the ncRNA intermediate. We searched for *E. coli* orthologs of these potential Hfq targets and found 601 CDSs. Although these Hfq-related CDSs represent only 18% of total *E. coli* CDSs, they represent almost half of the CDSs in ncRNA clusters #1 and #2 (respectively, 45/99 CDSs and 15/33 CDSs; $P = 5.6 \times 10^{-6}$ and 7.4×10^{-3}). There is thus a significant bias toward the presence of Hfq-bound mRNAs in the ncRNA-rich clusters of *E. coli*.

Experimental validation of candidate ncRNAs

We chose the Gram positive bacterium *Staphylococcus aureus* to challenge the NAPP method. *S. aureus* is a human commensal and an opportunistic pathogen that is widely responsible for nosocomial infections. Using NAPP to identify novel ncRNAs could have medical relevance, especially as only few ncRNAs were experimentally confirmed in this species (Pichon and Felden 2005). NAPP performed on strain N315 discerned 189 new ncRNA candidates in the ncRNA-rich clusters (Supplemental Table S6). We selected those that were positioned at more than 60 bp from a coding sequence, in order to avoid probing proximal elements in 5′ UTRs. Of the 48 sequences meeting this criterion, we randomly selected 24 for validation. To test for their presence, *S. aureus* was grown in rich media at 37°C. As ncRNAs are often expressed in specific growth conditions, total RNAs were extracted from samples collected at different growth phases. Northern blot experiments were performed using strand-specific-labeled oligonucleotides designed to probe the CNEs on each strand. Among the 24 CNEs tested, seven yield signals between 100 and 300 nt, showing the presence of corresponding small RNAs (named *rsaOA* to *rsaOG*; Fig. 5). Some of the labeled probes yield additional high molecular weight signals. Sequence analysis suggests that SA0347 (*metB*) is regulated by a T-box riboswitch and SA0667 (*queC*) by a preQ1 type riboswitch (Griffiths-Jones et al. 2005; Roth et al. 2007). Therefore, *rsaOB* and *rsaOC* are short transcripts likely resulting from premature transcription arrests of putatively longer transcripts that are also observed. Although these experiments suggest that nearly one-third of the ncRNAs predicted by NAPP may actually be produced by *S. aureus*, this proportion is likely underestimated, since we only assessed a limited set of conditions. For example, different stress conditions might induce expression of other, as yet undetected ncRNAs. This experimental approach demonstrates the efficiency of NAPP for the discovery of new small RNAs in bacteria.

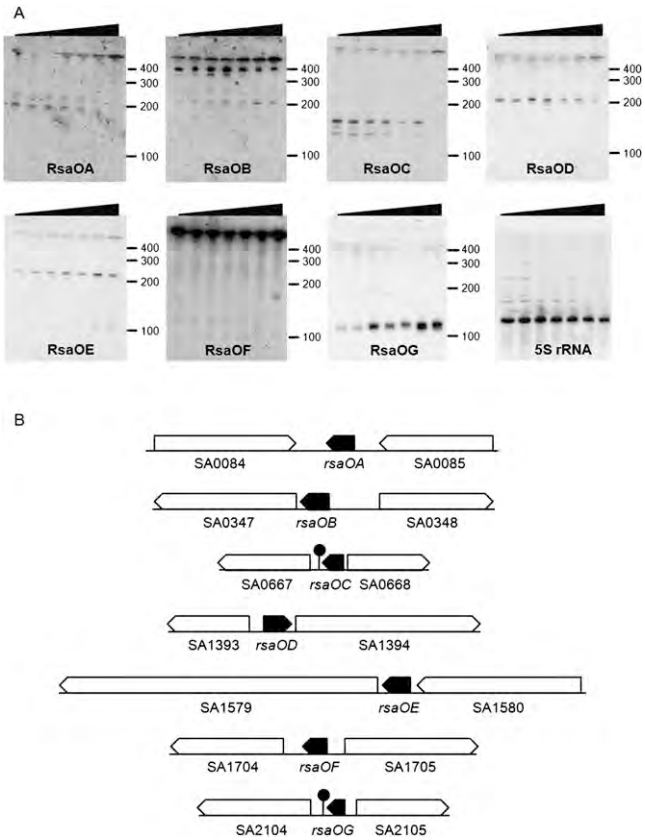


Figure 5. Identification of seven new *S. aureus* small RNAs. (*A*) Northern blot experiments showing the expression of NAPP-predicted small RNAs of N315 *S. aureus* in nutrient-rich medium. Total RNAs were prepared from cultures in exponential, early stationary, and late stationary phases. Black wedges indicate the increased OD_{600}: 0.2, 0.4, 0.8, 2, 3.5, 6, and 10 from *left* to *right*. Marker sizes are shown to the *left* of each blot. RNAs were detected using ^{32}P-labeled oligonucleotides designed to pair with the CNEs (cf. Supplemental Table S7). The name that we gave to each small RNA (*rsaOA* to *rsaOG*) is indicated. As the genome contains intergenic sequences similar to *rsaOF*, the RsaOF probe could hybridize to RNAs expressed from regions other than that indicated. 5S RNA was used as a control. (*B*) Genes flanking detected small RNAs. Arrows indicate the direction of transcription. Plain arrows correspond to small RNAs. Empty arrows correspond to open reading frames. Transcription terminators were predicted using Erpin (Gautheret and Lambert 2001) and RNAMotif (Macke et al. 2001).

Downloaded from genome.cshlp.org on September 8, 2010 - Published by Cold Spring Harbor Laboratory Press

Discussion

The assumption underlying phylogenetic profiling is that genes involved in a particular complex or pathway should co-occur in different species. As co-occurrence events are particularly significant when involving distant species (Cokus et al. 2007), and considering the rapid evolution of nucleic acid sequences, the application of phylogenetic profiling to the classification of nucleic acids was an uncertain endeavor. We provide here the first evidence that nucleic acid profiling performed on the full genome scale is functionally informative. Phylogenetic profiling efficiently clusters tRNAs, rRNAs, and other ncRNAs in several distinct groups, and concentrates a significant number of housekeeping and RNA-regulated (Hfq-bound) protein-coding genes in the ncRNA clusters. This approach opens fascinating perspectives for the analysis of RNA genes and RNA–protein interactions, and at the same time raises new questions as to why some sets of RNAs and CDSs should exhibit similar evolutionary profiles.

NAPP represents sequence constraints as a multidimensional value corresponding to conservation levels observed in multiple species. We propose that this vector representation depicts RNA evolution far more accurately than a one-dimensional value. Constraints on RNA sequences are subject to environmental and biological factors and can thus vary widely across species. The effectiveness of NAPP clustering shows that, in spite of these species-dependent evolutionary rates, sets of RNAs from the same genome display similar interspecies variations of conservation levels. In other words, we can witness constraints on RNA sequences that vary across the phylogenetic spectrum, but vary similarly from one RNA gene to another.

Exploiting these common evolutionary patterns, NAPP efficiently classifies bacterial RNAs in distinct clusters with no a priori knowledge of these molecules. A more precise functional classification that would distinguish ncRNAs involved, for instance, in iron homeostasis or other specific cellular functions will be harder to achieve. However, NAPP in its current form is already an efficient strategy for de novo ncRNA identification. Unlike other ncRNA prediction programs, NAPP neither requires similarities to existing gene families, nor involves analysis of sequence composition, RNA structure or covariation. It is remarkable that a clustering procedure that ignores these characteristics can perform similarly or better than dedicated RNA identification software. This approach thus opens a new path toward the identification of ncRNAs that escaped established screening methods.

Growth of the complete genome collection will result in denser sampling of species around genomes of interest, and hence to more informative phylogenetic profiles. As NAPP computing costs increase only linearly with the number of species, we should be able to fully exploit the rapidly expanding genome collection. Among possible outcomes of future functional analyses, we expect to gain insight into novel noncoding elements involved in specific adaptive responses, as well as in a large number of novel RNA–RNA and RNA–protein interactions.

Methods

Selection of conserved noncoding elements (CNEs)

We retrieved the 423 complete reference sequence (RefSeq) genomes from the NCBI genome database (Pruitt et al. 2007) as of February 2007. We extracted noncoding regions of the reference genome based on NCBI's coding sequence annotations and aligned these to the 422 remaining complete genome sequences using Washington University's WU-BLAST (parameters: −p BLASTN −W 5 −e 0.01). We computed the conservation rate of each intergenic position as the number of genomes in which this position was involved in a BLASTN hit. To reduce the impact of near-identical genomes in the collection, we weighted the conservation rates using interspecies phylogenetic distances derived from 16S rRNA. We obtained the 16S rRNA sequences from the RDP website (Cole et al. 2005), performed multiple alignments of the sequences with MUSCLE (Edgar 2004), and computed the distance matrix using DNAdist from the PHYLIP V3.6 package (Felsenstein 1989). The final conservation index (CI) for any intergenic position of the reference species *Sr* is computed as follows:

$$CI = \sum_{k\in[1;422]} \frac{Dist(Sr,Sk) * X_{r\to k}}{\sum_{j\in[1;422]} Dist(Sr,Sj)}$$

With *Dist*(*Sr*,*Sk*) the distance between the 16S rRNAs of species *Sr* and *Sk*, and *X* a Boolean representing the presence or absence of a nucleotide from species *Sr* in the local alignment with species *Sk*. The denominator is the sum of all distances between the 16S rRNA of *Sr* and the 16S rRNAs of other species *Sj*. We defined CNEs as segments with a CI greater than 0.004 over a minimal length of 15 nt. We merged consecutive CNEs separated by 35 nt or less. These parameters were determined empirically based on the observation of known ncRNAs in the *E. coli* and *B. subtilis* genomes.

Phylogenetic profiles and profile clustering

CNEs and protein-coding genes (CDSs from start to stop codon) were submitted to the same procedure. We aligned each CNE/CDS from a reference organism against the 422 remaining complete genomes using WU-BLAST (parameters: −p BLASTN −W 5 −e 0.01). The highest bit score obtained against each genome was normalized by the BLASTN score of the CNE/CDS sequence against itself. The phylogenetic profile of a CNE/CDS corresponds to this vector of 422 normalized scores. To clusterize profiles by similarity, we used the K-means procedure from the R package (Ihaka and Gentleman 1996) using Pearson distance and $k = 35$ clusters. We reached the number of 35 clusters by gradually increasing the cluster number until empty groups began to appear. Because of the stochastic component in K-means, each K-means run generates a different clustering. We performed multiple K-means runs for each species studied and observed no significant variation in cluster quality with respect to ncRNA contents. Known ncRNAs were always grouped into one or two clusters with highly significant enrichments. Detailed contents of ncRNA-enriched clusters for an expanding list of species are available for download at http://rna.igmors.u-psud.fr/download.

Functional enrichment

We obtained the coordinates of noncoding RNAs from the RFAM 8.0 database (Griffiths-Jones et al. 2005) and the coordinates of *E. coli* terminators, promoters, and transcription factor binding sites (TFBS) from RegulonDB 6.1 (Gama-Castro et al. 2008). We classified a CNE as a known ncRNA, terminator, or TFBS when it overlapped the corresponding element. On average, Rfam sequences covered 84% of the CNE length in *E. coli* and *B. subtilis*. Functional terms for *E. coli* and *B. subtilis* proteins were obtained from the KEGG database, January 2008 version (Kanehisa et al. 2002). We disregarded KEGG pathways that did not present sufficient functional relationships among their members, as suggested by Snitkin et al. (2006). For enrichment in Hfq-bound mRNAs, we extracted the list of 727 Hfq-associated mRNAs in

Downloaded from genome.cshlp.org on September 8, 2010 - Published by Cold Spring Harbor Laboratory Press

Salmonella typhimurium from Sitka et al. (2008) and obtained *E. coli* orthologs from the COG database (Tatusov et al. 2003). To measure the enrichment of cluster C in property P, we used a two-sided Fisher test (P-value < 10^{-2}), comparing the number of elements with property P in cluster C, the number of elements with property P in all clusters, and the total numbers of elements in cluster C and in all clusters.

Comparison with RNAz and SIPHT

Multiple alignments were generated for RNAz analysis as follows. For each CNE over 50 nt, we collected the best WU-BLAST hits (parameters: −p BLASTN −W 5 −e 0.01) in each of the 422 species. We discarded hits shorter than 75% of the query size and performed multiple alignments of the CNE and BLASTN hits using Muscle (Edgar 2004). Multiple alignments containing four or more distinct sequences were submitted to RNAz (512 alignments for *E. coli*, 653 for *B. subtilis*, 653 for *S. aureus*, and 456 for *V. cholerae*). Overall, only 10%–15% of CNEs produced a multiple alignment of sufficient quality for RNAz analysis. RNAz was run with default options (any alignment <50% identity is discarded). When a positive ncRNA prediction was scored by RNAz in one of the scanning windows, we tagged the whole alignment as ncRNA. Predictions of the SIPHT program were obtained directly from the authors' web server: http://bio.cs.wisc.edu/sRNA.

Experimental validation

We used N315, a methicillin resistant *S. aureus* strain that caused hospital-acquired infections in 1982 and was fully sequenced (Kuroda et al. 2001). Overnight staphylococcal cultures in BHI medium were diluted 1000-fold in the same medium and grown at 37°C. Total RNAs were extracted at the indicated optical density as previously described (Oh and So 2003; Bohn et al. 2007). Northern blot experiments were performed on samples separated by denaturing polyacrylamide gel electrophoresis. RNAs were transferred to nylon membranes using a semidry electrophoretic transfer unit. Oligonucleotide probes (see Supplemental Table S7) were γ32-dCTP labeled using the T4 polynucleotide kinase according to manufacturer's instructions (Fermentas). Hybridizations were performed overnight at 52°C. Radioactivity was detected using phosphor screen imaging system (Typhoon, GE-Healthcare).

Acknowledgments

This work was supported in part by the Institut Fédératif de Recherche 115 (IFR 115) and by grant ANR-06-MIME-016-01 from Agence Nationale pour la Recherche.

References

Altuvia, S. 2007. Identification of bacterial small noncoding RNAs: Experimental approaches. *Curr. Opin. Microbiol.* **10:** 257–261.

Barrick, J.E. and Breaker, R.R. 2007. The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome Biol.* **8:** R239. doi: 10.1186/gb-2007-8-11-r239.

Bohn, C., Rigoulay, C., and Bouloc, P. 2007. No detectable effect of RNA-binding protein Hfq absence in *Staphylococcus aureus*. *BMC Microbiol.* **7:** 10. doi: 10.1186/1471-2180-7-10.

Cokus, S., Mizutani, S., and Pellegrini, M. 2007. An improved method for identifying functionally linked proteins using phylogenetic profiles. *BMC Bioinformatics* (Suppl. 4) **8:** S7. doi: 10.1186/1471-2105-8-S4-S7.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam, S.A., McGarrell, D.M., Garrity, G.M., and Tiedje, J.M. 2005. The Ribosomal Database Project (RDP-II): Sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.* **33:** D294–D296.

del Val, C., Rivas, E., Torres-Quesada, O., Toro, N., and Jimenez-Zurdo, J.I. 2007. Identification of differentially expressed small non-coding RNAs in the legume endosymbiont *Sinorhizobium meliloti* by comparative genomics. *Mol. Microbiol.* **66:** 1080–1091.

Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32:** 1792–1797.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95:** 14863–14868.

Enault, F., Suhre, K., Abergel, C., Poirot, O., and Claverie, J.M. 2003. Annotation of bacterial genomes using improved phylogenomic profiles. *Bioinformatics* (Suppl. 1) **19:** i105–i107.

Felsenstein, J. 1989. PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics* **5:** 164–166.

Gama-Castro, S., Jimenez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Penaloza-Spinola, M.I., Contreras-Moreira, B., Segura-Salazar, J., Muniz-Rascado, L., Martinez-Flores, I., Salgado, H., et al. 2008. RegulonDB (version 6.0): Gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.* **36:** D120–D124.

Gautheret, D. and Lambert, A. 2001. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol.* **313:** 1003–1011.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., and Bateman, A. 2005. Rfam: Annotating noncoding RNAs in complete genomes. *Nucleic Acids Res.* **33:** D121–D124.

Ihaka, R. and Gentleman, R. 1996. R: A language for data analysis and graphics. *J. Comput. Graph. Statist.* **5:** 299–314.

Janky, R. and van Helden, J. 2008. Evaluation of phylogenetic footprint discovery for predicting bacterial *cis*-regulatory elements and revealing their evolution. *BMC Bioinformatics* **9:** 37. doi: 10.1186/1471-2105-9-37.

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30:** 42–46.

Kuroda, M., Ohta, T., Uchiyama, I., Baba, T., Yuzawa, H., Kobayashi, I., Cui, L., Oguchi, A., Aoki, K., Nagai, Y., et al. 2001. Whole genome sequencing of methicillin-resistant *Staphylococcus aureus*. *Lancet* **357:** 1225–1240.

Livny, J., Teonadi, H., Livny, M., and Waldor, M.K. 2008. High-throughput, kingdom-wide prediction and annotation of bacterial noncoding RNAs. *PLoS One* **3:** e3197. doi: 10.1371/journal.pone.0003197.

Macke, T.J., Ecker, D.J., Gutell, R.R., Gautheret, D., Case, D.A., and Sampath, R. 2001. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.* **29:** 4724–4735.

Missal, K., Rose, D., and Stadler, P.F. 2005. Noncoding RNAs in *Ciona intestinalis*. *Bioinformatics* (Suppl. 2) **21:** ii77–ii78.

Oh, E.T. and So, J.-S. 2003. A rapid method for RNA preparation from Gram-positive bacteria. *J. Microbiol. Methods* **52:** 395–398.

Panek, J., Bobek, J., Mikulik, K., Basler, M., and Vohradsky, J. 2008. Biocomputational prediction of small noncoding RNAs in *Streptomyces*. *BMC Genomics* **9:** 217. doi: 10.1186/1471-2164-9-217.

Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E.S., Kent, J., Miller, W., and Haussler, D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.* **2:** e33. doi: 10.1371/journal.pcbi.0020033.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci.* **96:** 4285–4288.

Pichon, C. and Felden, B. 2005. Small RNA genes expressed from *Staphylococcus aureus* genomic and pathogenicity islands with specific expression among pathogenic strains. *Proc. Natl. Acad. Sci.* **102:** 14249–14254.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. 2007. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35:** D61–D65.

Ragan, M.A. and Gaasterland, T. 1998. Microbial genescapes: A prokaryotic view of the yeast genome. *Microb. Comp. Genomics* **3:** 219–235.

Rivas, E. and Eddy, S.R. 2001. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2:** 8. doi: 10.1186/1471-2105-2-8.

Roth, A., Winkler, W.C., Regulski, E.E., Lee, B.W.K., Lim, J., Jona, I., Barrick, J.E., Ritwik, A., Kim, J.N., Welz, R., et al. 2007. A riboswitch selective for the queuosine precursor preQ1 contains an unusually small aptamer domain. *Nat. Struct. Mol. Biol.* **14:** 308–317.

Sittka, A., Lucchini, S., Papenfort, K., Sharma, C.M., Rolle, K., Binnewies, T.T., Hinton, J.C.D., and Vogel, J. 2008. Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet.* **4:** e1000163. doi: 10.1371/journal.pgen.1000163.

Snitkin, E.S., Gustafson, A.M., Mellor, J., Wu, J., and DeLisi, C. 2006. Comparative assessment of performance and genome dependence

Downloaded from genome.cshlp.org on September 8, 2010 - Published by Cold Spring Harbor Laboratory Press

among phylogenetic profiling methods. *BMC Bioinformatics* **7:** 420. doi: 10.1186/1471-2105-7-420.

Sonnleitner, E., Sorger-Domenigg, T., Madej, M.J., Findeiss, S., Hackermuller, J., Huttenhofer, A., Stadler, P.F., Blasi, U., and Moll, I. 2008. Detection of small RNAs in *Pseudomonas aeruginosa* by RNomics and structure-based bioinformatic tools. *Microbiology* **154:** 3175–3187.

Srinivasan, B.S., Caberoy, N.B., Suen, G., Taylor, R.G., Shah, R., Tengra, F., Goldman, B.S., Garza, A.G., and Welch, R.D. 2005. Functional genome annotation through phylogenomic mapping. *Nat. Biotechnol.* **23:** 691–698.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., et al. 2003. The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* **4:** 41. doi: 10.1186/1471-2105-4-41.

Tjaden, B., Saxena, R.M., Stolyar, S., Haynor, D.R., Kolker, E., and Rosenow, C. 2002. Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic Acids Res.* **30:** 3732–3738.

Torarinsson, E., Sawera, M., Havgaard, J.H., Fredholm, M., and Gorodkin, J. 2006. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.* **16:** 885–889.

Washietl, S., Hofacker, I.L., and Stadler, P.F. 2005. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci.* **102:** 2454–2459.

Washietl, S., Pedersen, J.S., Korbel, J.O., Stocsits, C., Gruber, A.R., Hackermuller, J., Hertel, J., Lindemeyer, M., Reiche, K., Tanzer, A., et al. 2007. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res.* **17:** 852–864.

Zhang, Y., Zhang, Z., Ling, L., Shi, B., and Chen, R. 2004. Conservation analysis of small RNA genes in *Escherichia coli*. *Bioinformatics* **20:** 599–603.

Received December 2, 2008; accepted in revised form February 9, 2009.

RsaOG, a new staphylococcal family of highly transcribed non-coding RNA

Antonin Marchais, Chantal Bohn, Philippe Bouloc and Daniel Gautheret.

Published in RNA Biology, 2010.[124]

Resume

Cet article décrit une famille d'ARNnc agissant probablement en *trans* et nommée RsaOG. Cette famille a été découverte dans le cadre de l'analyse du génome de *S. aureus* par la méthode NAPP. Cet ARNnc est spécifique du genre Staphylococcus où il est très conservé. Nous délimitons par une expérience de 5'RACE le site de démarrage de la transcription du transcrit chez *S. aureus* et la présence d'un terminateur stable nous permet de borner son extrémité 3'. RsaOG a pour particularité de présenter deux régions très conservées qui semblent former un pseudo-noeud impliquant un segment simple brin et une boucle interne de la tige principale. De plus nous prédisons à l'aide de l'outil RNAup[125] quatre gènes dont les transcrits pourraient être la cible de cet ARNnc. Comme RsaOG n'est retrouvé que dans le genre Staphylococcus, il est vraisemblable qu'il soit impliqué dans des processus spécifiques à ce genre et par conséquent, il pourrait à l'avenir, sachant que ces bactéries constituent un important problème de santé publique, permettre de mieux comprendre les mécanismes d'infection, ou même devenir une cible thérapeutique.

RNA Biology 7:2, 116-119; March/April 2010; © 2010 Landes Bioscience

RsaOG, a new staphylococcal family of highly transcribed non-coding RNA

Antonin Marchais, Chantal Bohn, Philippe Bouloc and Daniel Gautheret*
Institut de Génétique et Microbiologie; CNRS/UMR 8621; Université Paris Sud 11; Orsay, France

©2010 Landes Bioscience. Do not distribute.

The expression of trans-acting small RNAs in firmicutes has been poorly documented to date. This gap is being filled quickly in the genus Staphylococcus, which is both a model firmicute and an important human pathogen. Here we analyse RsaOG, a novel small RNA family specific to Staphylococcus and highly transcribed. This well conserved element, first discovered in a computational screen, was precisely mapped in the genome by RACE mapping and the identification of a putative transcriptional promoter. The proposed secondary structure presents two highly conserved unpaired sequences, part of which can form a pseudoknot. We suggest a possible involvement of the remaining conserved single stranded region in trans regulatory interactions.

Key words: sRNA, small RNA, ncRNA, firmicutes

Submitted: 11/04/09

Revised: 12/09/09

Accepted: 12/09/09

Previously published online: www.landesbioscience.com/journals/rnabiology/article/10925

*Correspondence to: Daniel Gautheret; Email: daniel.gautheret@u-psud.fr

Introduction

Methicillin-resistant *Staphylococcus aureus* (MRSA) is a critical health problem as a cause of hospital-acquired infection. In this context, the characterization of novel functional loci in this organism is interesting for the identification of potential drug targets or diagnosis/virulence markers. Among functional elements in the *S. aureus* genome, those present in so-called "intergenic" regions (IGR) remain the least characterized. IGRs host different classes of non-coding RNA (ncRNA) elements, including trans-acting small RNA (sRNA) genes and cis-regulatory elements such as attenuators and riboswitches. Currently, the majority of known non-coding RNAs in Staphylococcal strains as well as in other Gram positive bacteria are of the second category. Known sRNAs in *S. aureus* include RNAIII[1] and Pathogenicity Island specific ncRNAs.[2] The ncRNA RsaOG (acronym for RNA *S. aureus* Orsay G) was initially discovered in a bioinformatics ncRNA screen of the *S. aureus* genome[3] and confirmed in a second study.[4] Here, we show that it is the founding member of a new trans-acting ncRNA family conserved in all Staphylococcaceae. We define the *rsaOG* gene boundaries and analyze its RNA sequence alignment, structure and phylogeny.

Results

We identified RsaOG using a phylogenetic clustering analysis of non-coding elements in the *S. aureus* genome.[3] A northern blot experiment indicated that the *rsaOG* gene was highly expressed. RsaOG was detected at all growth stages but with a progressive increase in post-exponential and stationary phase. Northern blot indicated an RsaOG size between 130 and 150 nucleotides.

The transcriptional start of *rsaOG* in *S. aureus* N315 was determined by a 5' rapid amplification of cDNA ends (5' RACE) mapping at position 2368061 on the genetic map, in agreement with the recently reported location.[4] This position is likely a transcriptional start since (i) a 5' RACE performed in the absence of TAP treatment shows a highly reduced level of first strand synthesized cDNA (**Fig. 1B**) and (ii) a potential -10 and -35 transcription promoter, conserved in other Staphylococcus is present upstream of the identified transcriptional start (**Fig. 1A**). Analysis of the *rsaOG* region reveals the presence of a characteristic Rho-independent terminator. The 5' RACE experiments and computational

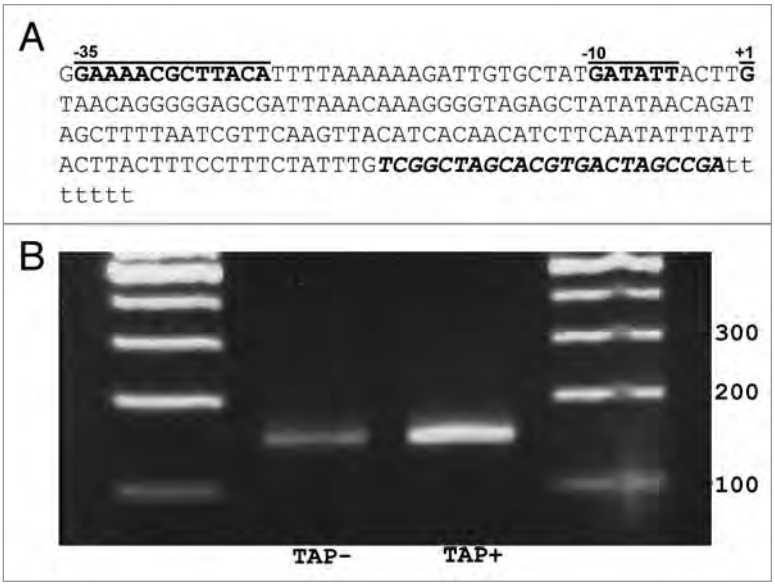


Figure 1. (A) The *S. aureus rsaoG* sequence annotated with transcriptional promoter signals (-10,-35) and 5' RACE validated TSS (+1). The transcriptional terminator hairpin is shown in italics. (B) Northern blot for quantification of first strand synthetized cDNA during reverse-transciption step with (TAP^+) and without (TAP^-) TAP treatement. Marker sizes are shown to the left and right of the blot.

©2010 Landes Bioscience. Do not distribute.

terminator prediction indicate that the *S. aureus N315* RsaOG transcript is 146 nucleotides long.

We searched for RsaOG homologs using the Blastn[5] and Erpin[6] programs in all sequenced bacteria and detected their presence only in the Staphylococcus genus, from *S. aureus* to the most distant sequenced genome *Staphylococcus carnosus*. In all cases, *rsaOG* is flanked by the same protein-coding genes, a *suhB* paralog and a *deoR*-like transcription repressor (SA2104 and SA2105 in N315, respectively). Since *rsaOG* and SA2104 are present only in the Staphylococcus genus, we can speculate a common history for both genes, although we cannot definitely relate these genes in an evolutionary or functional manner. However, the general congruence of the *rsaOG* and SA2104 phylogenetic trees with the Staphylococcus 16S phylogeny[7] (**Fig. 2B**) indicates that *rsaOG* and SA2104 were neighbors in the staphylococcal common ancestor.

The proximity (~150 nt) and similar orientation of *rsaOG* and SA2104 raised the possibility that *rsaOG* was a cis-acting element of SA2104. However, this hypothesis is not supported by experimental data: no additional band corresponding to transcripts longer than RsaOG, a characteristic of riboswitches or attenuators detected by northern blot experiments.[3] In addition, a microarray study suggests a σ^B-dependent regulation of SA2104 and the presence of a σ^B-dependent promoter located downstream of RsaOG.[8] These observations strongly support that SA2104 is transcribed independently of *rsaOG*.

The multiple alignment of putative RsaOGs (**Fig. 2A**) shows a size variation between 143 and 154 nt and the presence of two regions of higher conservation. The first conserved region starting at +22 contains the AG-rich motif CAAAGGGGUA. A second occurrence of this motif appears further downstream as a perfect repeat in *Staphylococcus saprophyticus* and in a degenerated form in the other species. The other conserved region starting at +77 is UC-rich and about 47 nt-long.

We predicted the RsaOG secondary structure using Mfold[9] in conjunction with MCfold[10] for pseudoknot prediction. Our secondary structure model (**Fig. 2C**) is supported by 10 compensatory mutations in the major hairpin, and leaves the conserved elements mostly unpaired, suggesting a possible involvement of these elements in other RNA-RNA interactions. The AG-rich motif, located in a secondary structure bulge, is able to form a pseudoknot pairing with the end of the second conserved region. This arrangement suggests that the remaining part of the longer conserved region could be involved in interactions with potential RsaOG targets because such non-structured conserved elements in sRNAs are often involved in trans-pairing.[3] Taken together, these results suggest that the pseudoknot could, by constraining the structure, protect this conserved unpaired region and nucleate or facilitate pairing to the target.

We sought possible targets of RsaOG in all upstream regions of protein-coding genes in Staphylococcus species (see Methods). We found four potential targets in *S. aureus* genes that were also predicted in at least three other species. These targets were in genes *gmk* (guanylate kinase), *miab* (dimethylallyl adenosine tRNA methylthiotransferase), *SA2393* (2-hydropantoate 2-reductase) and *SA2053* (probable glucose uptake protein). For all target genes, the predicted interaction involved the conserved region of RsaOG and the Shine-Dalgarno sequence, which lends support to a possible trans-acting regulation.

Discussion

Until recently, regulation of gene expression by sRNA was considered an exception in firmicutes. Recent and ongoing studies are now changing this view.[3,4] We described here a new Staphylococcal RNA, RsaOG, that is likely involved in trans-acting regulatory functions. The fact that RsaOG is expressed during all growth stages suggests a persistent repression or activation function for this gene (or a fine tuning role involving a second factor). The staphylococcus-specific conservation suggests that *rsaOG* could be involved in a pathway that is specific to this genus, such as toxin production or invasion, thus strengthening the medical interest of this non-coding RNA. We provide a seed alignment and core secondary structure of RsaOG that will be useful for future genome searches and mutagenesis studies.

Methods

RACE mapping was performed as in Argaman et al.[12] Homology searches were performed against the NCBI nr databases

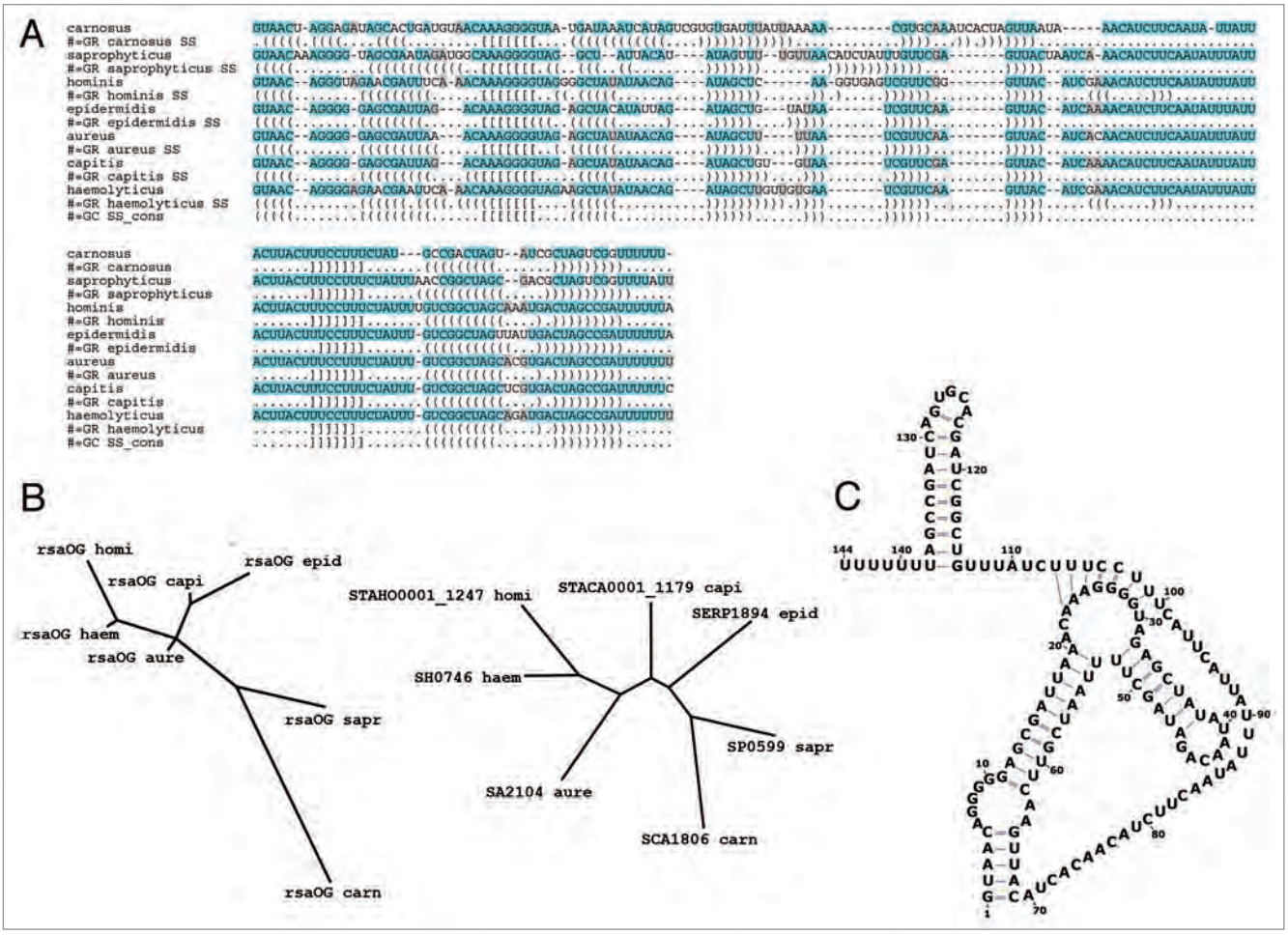


Figure 2. (A) Multiple alignment of the rsaOG orthologs. Nucleotide conservation is shown with a shaded background. Predicted secondary structure base pairs are shown with parentheses and the pseudoknot with square brackets. In all species, the pseudoknot involves a CAAAGGGGUA conserved motif. (B) Unrooted trees of RsaOG and SA2104 (aur: *S. aureus*; sapr: *S. saprophyticus*; epid: *S. epidermidis*; homi: *S. hominis*; capi: *S. capitis*; haem: *S. haemolyticus*). (C) Varna[16] view of the *S. aureus* RsaOG structure with its pseudoknot.

using Blastn[5] and Erpin[6] for non coding elements and Blastp for coding genes. The "best reciprocal hit" strategy was used to discriminate *suhB* paralogs and orthologs. Multiple Alignments of non-coding RNAs and proteins were generated with the MUSCLE software.[13] RsaOG targets were sought in regions from -100 to +30 around start codons of annotated protein-coding genes, using the RNAup[11] program with parameters -Xf -b -w 25.

We predicted the secondary structure of each RNA using Mfold[9] and, independently, MCfold,[10] as we suspected the presence of a pseudoknot and MCfold is able to predict pseudoknots. The MC-fold predicted pseudoknot was compatible with the Mfold prediction. We used the Ralee Emacs plugin[14] to edit the structure alignment in order to merge the Mfold secondary structures and the pseudoknot. The phylogenetic trees of RsaOG and SA2104 were produced with phyML.[15]

Acknowledgements

This work was supported in part by the Institut Fédératif de Recherche 115 (IFR 115) and by grant ANR-06-MIME-016-01 from Agence Nationale pour la Recherche.

Note

Supplementary materials can be found at:
www.landesbioscience.com/supplement/MarchaisRNA7-2-Sup.pdf
www.landesbioscience.com/supplement/MarchaisRNA7-2-Sup.txt

References

1. Novick RP. Mobile genetic elements and bacterial toxinoses: the superantigen-encoding pathogenicity islands of *Staphylococcus aureus*. Plasmid 2003; 49:93-105.
2. Pichon C, Felden B. Small RNA genes expressed from *Staphylococcus aureus* genomic and pathogenicity islands with specific expression among pathogenic strains. Proc Natl Acad Sci USA 2005; 102:14249-54.
3. Marchais A, Naville M, Bohn C, Bouloc P, Gautheret D. Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles. Genome Res 2009; 19:1084-92.
4. Geissmann T, Chevalier C, Cros MJ, Boisset S, Fechter P, Noirot C, et al. A search for small noncoding RNAs in *Staphylococcus aureus* reveals a conserved sequence motif for regulation. Nucleic Acids Res 2009.
5. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403-10.
6. Gautheret D, Lambert A. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. J Mol Biol 2001; 313:1003-11.
7. Ghebremedhin B, Layer F, Konig W, Konig B. Genetic classification and distinguishing of Staphylococcus species based on different partial gap, 16S rRNA, hsp60, rpoB, sodA and tuf gene sequences. J Clin Microbiol 2008; 46:1019-25.
8. Bischoff M, Dunman P, Kormanec J, Macapagal D, Murphy E, Mounts W, et al. Microarray-based analysis of the *Staphylococcus aureus* sigmaB regulon. J Bacteriol 2004; 186:4085-99.

9. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic acids research 2003; 31:3406-15.
10. Parisien M, Major F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature 2008; 452:51-5.
11. Muckstein U, Tafer H, Hackermuller J, Bernhart SH, Stadler PF, Hofacker IL. Thermodynamics of RNA-RNA binding. Bioinformatics 2006; 22:1177-82.
12. Argaman L, Hershberg R, Vogel J, Bejerano G, Wagner EG, Margalit H, Altuvia S. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. Curr Biol 2001; 11:941-50.
13. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004; 32:1792-7.
14. Griffiths-Jones S. RALEE—RNA ALignment editor in Emacs. Bioinformatics 2005; 21:257-9.
15. Guindon S, Gaswel O. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol 2003; 52: 696-704.
16. Darty K, Denise A, Ponty Y. VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics 2009; 25:1974-5.

©2010 Landes Bioscience.
Do not distribute.

A SPORULATION-SPECIFIC, SMALL NONCODING RNA HIGHLY CONSERVED IN ENDOSPORE FORMERS

Antonin Marchais, Sandra Duperrier, Sylvain Durand, Daniel Gautheret, and Patrick Stragier.
In preparation.

RESUME

Dans cet article, nous présentons les premiers éléments indiquant que la méthode NAPP a la capacité de prédire la fonction d'un ARNnc bactérien. En recherchant les séquences uniquement conservées dans les firmicutes formant des endospores et en utilisant *B. subtilis* comme espèce de référence, nous nous sommes intéressés à un sous-groupe constitué d'une centaine d'éléments comportant des ARNnc connus, des gènes impliqués pour plus de la moitié dans la sporulation, et pour finir quelques CNE non caractérisés. Un des ARNnc connus a particulièrement retenu notre attention pour deux raisons. Premièrement, sa description dans la banque RFAM était certainement fausse puisque son alignement de séquence et structure secondaire indiquait un terminateur Rho-indépendant et donc une transcription sur le brin opposé. La seconde observation qui a attiré notre attention était la présence d'un signal en amont de cet élément non-codant correspondant à un promoteur sous le contrôle des facteurs de transcriptions spécifique de la sporulation, σ^F et σ^G. Le laboratoire de P. Stragier a vérifié par Northern blot que cet ARNnc était transcrit et que son expression était dépendante des facteurs σ^F et σ^G en fusionnant son promoteur avec le gène rapporteur *lacZ*. L'ARN a été nommé CsfG pour « Controlled by sporulation factor ». Le laboratoire de P. Stragier a également pu détecter un phénotype associé à la disruption de *csfg* dans *B. subtilis*. En absence de cet ARNnc, des études de compétition avec une souche sauvage semblent indiquer que les bactéries mutantes présentent une moindre capacité de survie estimée à 30% après chaque cycle de germination/croissance/sporulation. Ces résultats récents ne sont pas incorporés dans la version de l'article présenté dans ma thèse. CsfG constitue le premier ARNnc agissant en *trans* dont la fonction a pu être prédite, à ma connaissance, par une méthode *in-silico* sans que l'on ne recherche ses cibles. Ce travail est en train d'être complété par une recherche des cibles de CsfG avant son envoi pour publication.

A sporulation-specific, small noncoding RNA highly conserved in endospore formers

Antonin Marchais[1], Sandra Duperrier[2], Sylvain Durand[2], Daniel Gautheret[1], and Patrick Stragier[2*]

[1] Université Paris-Sud, CNRS-UMR8621, Institut de Génétique et Microbiologie, Bâtiment 400, 91405 Orsay Cedex, France

[2] CNRS-UPR9073, Université Paris-Diderot, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, 75005 Paris, France.

* *Corresponding author*. E-mail address: stragier@ibpc.fr

Running title: sporulation specific sRNA

Keywords: small RNA, sporulation

Abstract

Endospore formation is a characteristic shared by some *Bacilli* and *Clostridia* that involves the creation of two cell types, the forespore and the mother cell. Hundreds of protein-encoding genes have been shown to be transcribed in a cell-specific fashion during this developmental process in *Bacillus subtilis*. We have used a phylogenetic profiling procedure to identify clusters of *B. subtilis* coding and noncoding sequences that co-occur in other endospore formers. One such cluster shows a strong bias for sporulation-related genes (42 % among 156 genes) and is enriched in potential noncoding RNAs. We have studied one RNA candidate, encoded in the *ylbG-ylbH* interval. *In vivo* analysis using a transcriptional fusion to the *Escherichia coli lacZ* gene demonstrates that this region of the chromosome contains a gene, *csfG*, encoding a 147-nucleotide RNA that is transcribed only during sporulation, specifically in the forespore. Despite being nonessential for the production of spores in *B. subtilis csfG* is present in many endospore formers, mostly *Bacilli* and some *Clostridia*, whereas it is absent from bacteria that do not produce endospores. All *csfG* RNAs contain a strongly conserved, pyrimidine-rich, central motif that overlaps a potential stem-loop structure. The remarkable conservation of this sequence in widely divergent bacteria suggests that it plays a conserved physiological role, presumably by interacting with an unidentified target in the forespore, where it contributes to the acquisition of the spore properties.

Introduction

Endospore formation is a primitive developmental process observed in some *Bacilli* and *Clostridia* when subjected to nutrient starvation. These bacteria belong to the phylum of Firmicutes that also contains many bacteria unable to sporulate, among which members of the *Listeria*, *Staphylococcus*, *Streptococcus*, and *Lactobacillus* orders. The morphological process of sporulation appears to be extremely similar in all endospore formers, starting with a polar division that creates two cells of different sizes and different fates, the smaller forespore and the larger mother cell. Engulfment of the forespore by the mother cell is followed by the synthesis of peptidoglycan and protein layers surrounding the forespore and contributing to its resistance properties after its release in the environment by lysis of the mother cell.[1] Sporulation has been extensively studied in *Bacillus subtilis* because of the amenability of this bacterium to classical and reverse genetics. Transcriptome experiments have identified several hundreds of protein-encoding genes that are expressed at some stage of sporulation in one or the other cell.[2-4] It is now well established that the developmental process is initiated by accumulation of the Spo0A~P regulatory protein and the σ^H transition sigma factor, whereas a cascade of four sigma factors, σ^F followed by σ^G in the forespore, and σ^E followed by σ^K in the mother cell, is responsible for the temporal and spatial activation of sporulation-specific genes after septation.[5] Analysis of the complete genomes of more than 50 endospore formers, including many *Clostridia*,[6] indicates the conserved presence of the major regulatory players identified in *B. subtilis*, in accordance with an early suggestion that all modern endospore formers derive from a single sporulating ancestor.[7]

Until recently, only proteins were considered to be involved in the control of sporulation. However, with the identification of many small noncoding RNAs (sRNAs) in a wide variety of bacteria, where they modulate gene expression of

multiple regulons, including complex processes such as biofilm formation,[8] morphological development,[9] or host-pathogen interaction,[10] it becomes essential to investigate the presence and the role of sRNAs during sporulation. This question has been addressed by Silvaggi *et al.*[11] who combined microarray transcriptional profiling of intergenic regions in *B. subtilis* with a computational screen for potential RNA secondary structures conserved in other endospore formers. Importantly for what follows, Silvaggi and co-workers eliminated the candidates for which a role as a *cis*-regulatory RNA element had been previously reported. They identified one sRNA that accumulates early during sporulation before septation and for which no ortholog was indicated, a second sRNA synthesized under the control of σ^K in the mother cell and also present in *Bacillus anthracis*, and finally a complex pattern of sRNAs synthesized from an intergenic region under the control of σ^G in the forespore and σ^K in the mother cell and conserved in *Bacillus halodurans*. Disrupting the intergenic regions encoding these sRNAs did not lead to any obvious sporulation defect.

Another study[12] focused exclusively on *B. subtilis* intergenic regions larger than 500 bp and did not detect any sporulation related sRNA. A comprehensive, high-resolution tiling array analysis has been recently reported[13] and, although performed on cells growing exponentially in rich or minimal medium, it provides useful hints about genome regions potentially transcribed only during sporulation because of their conspicuous silence in both growing conditions in the thorough published data set. Here we have built upon our previously published phylogenetic study of *B. subtilis* intergenic and coding regions[14] to identify potential sporulation-related sRNAs. We report the characterization of one such sRNA, present exclusively in endospore formers, expressed in the forespore, and containing a highly-conserved pyrimidine motif.

Results and Discussion

A cluster enriched in sporulation genes

We reasoned that sporulation-related sRNAs would be encoded by genes co-occuring with protein-encoding genes known to be involved in sporulation and present in several endospore formers. We used the NAPP (Nucleic Acids Phylogenetic Profiling) procedure to identify clusters of *B. subtilis* coding and non-coding sequences that co-occur in other species.[14] Briefly, NAPP involves collecting all conserved noncoding elements (CNEs) and coding sequences (CDS) in a reference species and seeking homologous sequences in other available species. CNEs and CDSs are then clustered based on the similarity of their occurrence profiles. When using *B. subtilis* as the reference species, CNEs corresponding to actual sRNA genes or riboswitches were significantly over-represented in two clusters.[14]

Table 1. A *B.subtilis* cluster enriched in sporulation genes.

Entry into sporulation (Spo0A~P and/or σ^H) [10][a]	Early mother cell (σ^E) [27]	Vegetative genes affecting sporulation [4]			
abrB	*asnO*	*hpr = scoC*	*pyrD*	*yqfU*	
kinE	*cotJC*	*menH*	*radC*	*yqzB = ccpN*	
qcrB	*cypA*	*spoIIJ*	*rbfA*	*yrhF*	
sigF	*kamA*	*ylbF*	*rbtT = ypzK*	*ysbB*	
spo0A	*mtpA*		*rnc*	*ysxB*	
spo0F	*prkA*		*rnmV*	*ytjP*	
spoIIAB	*spmB*		*rocF = argI*	*ytoI*	
spoIIE	*spoIIIAA*		*speA*	*yugJ*	
ymaH = htq	*spoIIID*		*speD*	*yusL = fadM*	
ypgR	*spoIVA*		*speE*	*yutE*	
	spoIVCB		*tagO*	*yutM*	
Early forespore (σ^F) [8]	*spoVD*		*topB*	*yvfS*	
	sqhC		*udk*	*yvfV*	
rnfd	*usd*		*yabA*	*yvgM*	
pgk	*yabP*		*yabB*	*yvoF*	
sigG	*yhaM*		*yabN*	*yvyE*	
yhbA	*yhbH*		*ybxG*	*ywhB*	
yjbA	*yhfN*		*ycbL*	*ywlG*	
yqhH	*yiaK*		*yceD*	*yxdJ*	
ytfJ	*yloS = thiN*		*yceF*		
yutC	*ylxY*		*ydfO*		
	yngE		*yfkE*		
Late forespore (σ^G) [11]	*yqfY = ispG*		*yflG = mapB*		
	yqiQ = pxpB		*yhaO*		
sleB	*ytpB*		*yjlC*		
spoIVB	*ywjF = fadF*		*yjlD*		
spoVAE	*ywlC*		*ykfA*		
spoVAF			*ykqB = ktrC*		
spoVT			*ykrT = mtnK*		
sspA	**Late mother cell (σ^K) [6]**		*ykrW*		
sspB			*ykrZ = mtnD*		
sspC	*dapG*		*ykuI*		
sspF	*gerE*		*ylkM = queF*		
sspI	*ykoY*		*yjlxP*		
tepA	*ypbH = mecB*		*yneT*		
	ytlA		*yoaS*		
	ytlD		*yobN*		
			ypbD		
			ypvA		
			yqeT		
			yqfF		

Other genes [90]
ackA
acoA
addB
ahpF
ahrC
araB
bcd
bkdAB
codV
dhbB
fer
fin
folB
folK
ftsX
fur
groES
hemY
hemZ
intA
intC
luxS
mntR
motA
mraW
murD
mutSB
paiA
paiB
pps

CNE [19][b]
rpoB-rpoC
rpmA-rplB
rplB-rpsS
yfmA-yflT
yjbH-yizD
metE S-box riboswitch
mtnKA S-box riboswitch
*ylbG-ylbH**
*polC-ylxS**
pbpX-ymdA
yneE-yneF
yoaZ-peiP
alaS T-box riboswitch
gapB-speD
phoR-polA
thiI-sspA
yugH-yugI
rho-rpmEA
*yybO-yybP**

[a] Protein encoding genes are grouped according to their transcription dependencies and their number is shown for each subset. Alternate names are indicated.
[b] The conserved noncoding elements (CNE) of unknown function are labeled by their bracketing genes. Asterisks point to previously identified CNEs.

Here we further analyzed phylogenetic and functional biases in these two RNA-rich clusters. We have identified a subcluster of genes co-occuring in many endospore formers, mainly *Bacilli* and some *Clostridia*, and highly enriched in sporulation-related genes. Out of 156 protein-encoding genes, 66 (i.e. 42%) are essential for sporulation (*bona fide spo* genes) or known to be transcribed at some stage of this developmental process (Table 1). The cluster also contains 19 CNEs, among which three riboswitches (for *alaS*, *metE*, and the *mtnKA* operon), two genetic intervals previously reported as potentially expressing a non-coding RNA (next to *ylbH* and to *yybP*),[15] and one region already identified as producing several sRNAs during sporulation (the *polC-ylxS* interval).[11] The latter result confirms the existence of a specific functional bias among the 175 members of the cluster and suggests that sporulation related sRNAs might be encoded by some of the other CNEs. The presence in the cluster of *ymaH*, encoding the Hfq RNA chaperone, provides an additional clue for a functional relationship between some of the 156 protein-encoding genes of the cluster and some of the sRNAs encoded by the 19 CNEs.

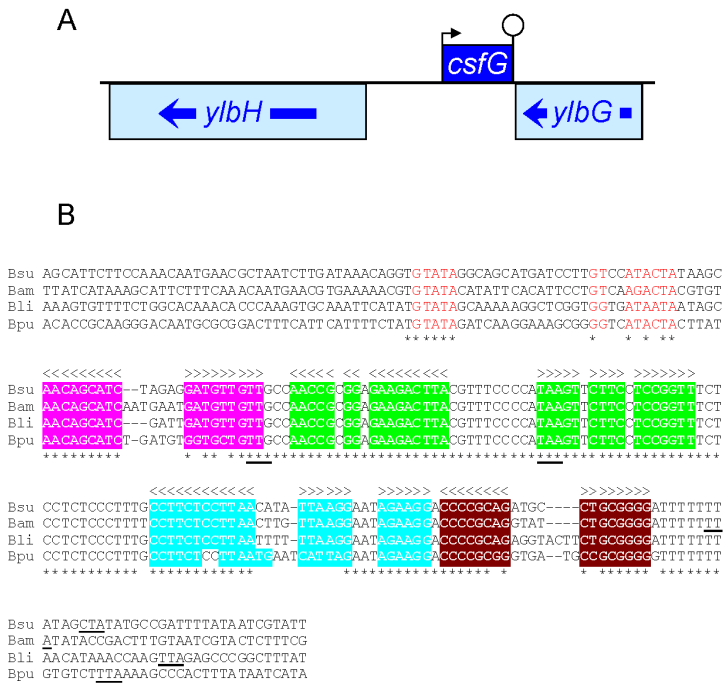


Figure 1. An sRNA gene in the *ylbH-ylbG* interval.
(a) The protein coding sequences of the *B. subtilis ylbG* and *ylbH* genes are shown as boxes (drawn to scale) with arrows showing their transcription orientation. The location of the putative *csfG* gene on the opposite strand is indicated by a box bracketed by its transcription initiation (arrow) and termination (lollipop) signals.
(b) The sequence immediately adjacent to *ylbG* is shown for *B. subtilis*, *Bacillus amyloliquefaciens*, *Bacillus licheniformis*, *and Bacillus pumilus*. Asterisks point to nucleotides identical in the four sequences. Arrowheads indicate potential base pairing and the four corresponding stem-loop structures are highlighted by a specific colored background. Nucleotides that constitute a sigma^F-controlled promoter are shown in red. For clarity the overlapping sigma^G-dependent promoter is not shown. The start and stop codons of a conserved short open reading frame are underlined under the alignment. The stop codon of *ylbG* (located downstream and in convergent orientation) are underlined under each sequence.

A forespore-specific sRNA

Our attention was drawn to the CNE corresponding to the *ylbG-ylbH* interval (Fig. 1a). This CNE overlaps an annotated *cis*-acting RNA named "*ylbH* leader" (RFAM entry RF00516)[15] and was therefore excluded from the analysis of Silvaggi *et al.*[11] However, we observed that this CNE is most likely expressed from the reverse strand as a *trans*-acting RNA, since it is followed on this strand by a run of Ts that could be part of a strong rho-independent transcription terminator. Moreover, aligning this region with the orthologous regions from three closely related *Bacilli* reveals an impressive sequence conservation that starts immediately downstream of a putative σ^F-controlled promoter and stops

immediately after the putative transcription terminator (Fig. 1b). A consensus, fully-conserved -35 recognition motif (GTATA), is followed at the optimal 15-bp distance by a sequence closely related to the consensus -10 recognition motif, Gg--A-AHTR, where H is A or C or T, and R is A or G.[4] Some σ^F-controlled promoters are also recognized by σ^G because of an overlap of sequence recognition by the two sigma factors. This is apparently the case here since the conserved putative promoter motifs fit well with the consensus for σ^G-dependent promoters, tGHATA (17-18 bp) MAWAMTA, where M is A or C, and W is A or T.[4]

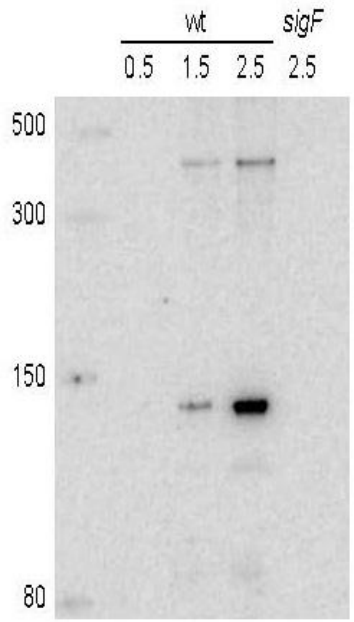


Figure 2. Northern blot detection of CsfG.
RNA was extracted from *B subtilis* cells (either wild type or deleted for the *sigF* gene) grown in DSM medium and harvested at various times after entry into sporulation. Hours of sporulation are indicated. Total RNA was separated by electrophoresis on an 8% polyacrylamide gel, transferred to a nylon membrane, and CsfG RNA was detected by hybridization with a complementary oligonucleotide. Samples were run alongside ssRNA markers (NEB) with the indicated sizes.

If this promoter is actually functional it would lead to the synthesis of a 147-nt RNA starting with an A residue at position 1569347 in the revised *B. subtilis* genome sequence (NCBI accession number NC_000964) and ending with a run of seven Us at position 15692001, in the opposite orientation to the adjacent *ylbG* and *ylbH* genes (Fig. 1a). To confirm the existence of this RNA we carried out a Northern blot analysis on total RNA extracted from cells harvested at early stages of sporulation. In our growth conditions polar septation takes place one hour after the entry into stationary phase. As shown in Fig. 2 an RNA of the expected size is present 30 minutes after septation (lane 1.5), but absent 30 minutes earlier (lane 0.5). This small RNA accumulates during the next hour (lane 2.5) but is not synthesized in the absence of σ^F. The probe hybridizes with a less abundant, larger RNA (*ca.* 400-nt), that follows the same temporal pattern. Since no promoter sequence can be identified at the corresponding position upstream, on the strand complementary to *ylbH*, it is likely that this larger RNA results from readthrough transcription and therefore ends up around the beginning of the *ylbG* gene. In any case it is a very minor species as compared to the smaller RNA.

A 11-codon reading frame is present in the four RNAs encoded by the orthologous DNA regions shown in Fig. 1b, starting with a UUG triplet at position 22 in the *B. subtilis* RNA. However, the corresponding ribosome binding site (GAGG) is conspicuously absent in the sequences from the three other *Bacilli*, making it unlikely that this small open reading frame is actually translated. Moreover, this reading frame is missing in orthologs of this RNA from more distant organisms discussed hereafter. Potential base pairing is identical in these *Bacilli* RNAs and comprises four stem-loop structures identified by specific colors in Fig. 1b. We note that the second loop is 100 % conserved between the four *Bacilli,* whereas the three other loops are highly variable. Also, 14 out of 15 nucleotides between the second and the third stem-loop are fully conserved.

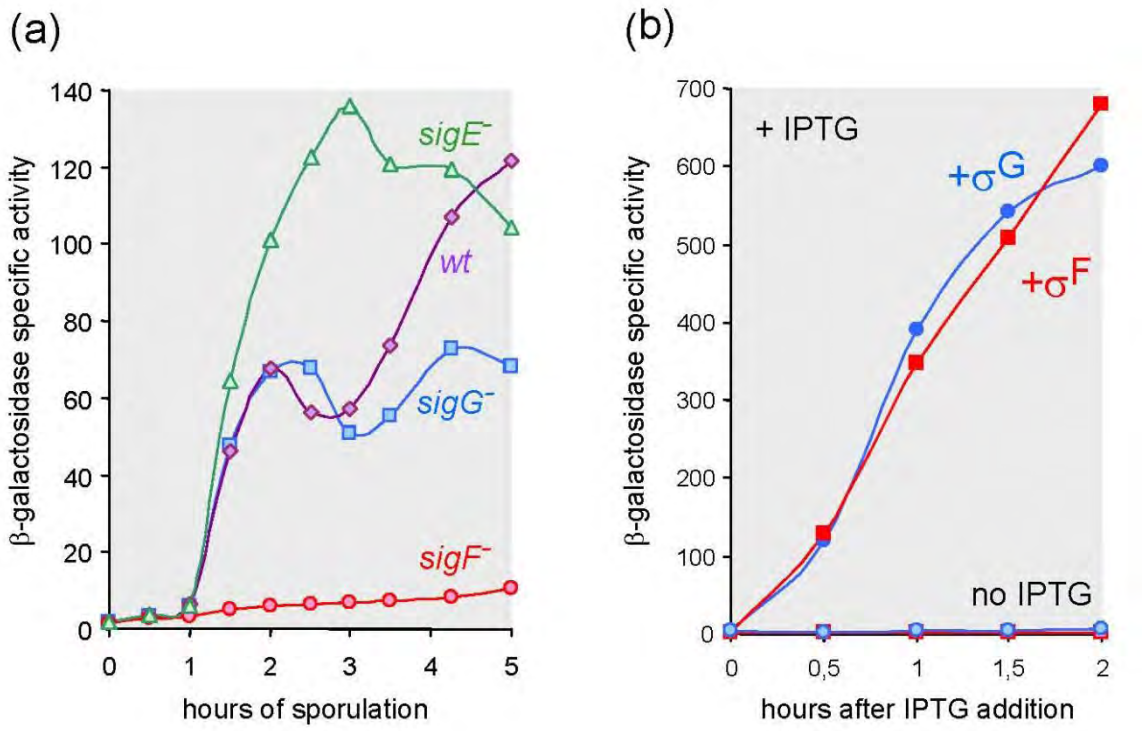


Figure 3. Forespore-specific transcription of the *csfG* gene.
(a) Time course expression of a transcriptional *csfG-lacZ* fusion integrated into the *B. subtilis* chromosome in various mutant backgrounds. Cells were grown in DSM.
(b) Induction of the same fusion in cells growing exponentially in 2xYT medium and harboring a plasmid allowing synthesis of sigma^F or sigma^G in the presence of IPTG. A control experiment was carried out in the absence of IPTG.

In order to assess quantitatively the synthesis of this sRNA in *B. subtilis* we fused its promoter region to the *Escherichia coli lacZ* gene and integrated the resulting transcriptional fusion into the *B. subtilis* chromosome at the nonessential *amyE* locus. The cloned fragment contains 132 bp upstream of the predicted transcription start as well as the first 35 bp from the transcribed region. As shown in Fig. 3a, β-galactosidase synthesis starts 1-hour after the initiation of the sporulation process and is abolished in the absence of σ^F. A second wave of β-galactosidase synthesis is observed 2-hours later in a wild-type strain but is missing in the absence of σ^G. Finally, in a mutant strain where the absence of σ^E leads to the formation of two forespore compartments, the level of β-galactosidase accumulation is roughly double of what is found in a wild-type strain. Altogether, these results fit perfectly with the presence of a promoter recognized by both σ^F and σ^G, and activated exclusively during sporulation, in the forespore.

This was further confirmed by inducing expression of the *lacZ* fusion in exponentially growing cells that were engineered to produce σ^F or σ^G in response to the addition of IPTG. The results shown in Fig. 3b indicate that both sigma

factors are equally efficient in activating transcription of the fusion. From now on, the gene encoding this sRNA will be called *csfG* (controlled by sigma-F and sigma-G).

Most of the *csfG* gene was replaced by a kanamycin-resistance cassette, eliminating 9 bp from the promoter region and 123 bp from the RNA encoding region. A *B. subtilis* strain harboring this mutation produces as much heat-resistant spores as the wild type strain and does not show any obvious growth defect (data not shown). We did not investigate more subtle phenotypes, such as germination rate.

A highly conserved motif

As expected from its belonging to a cluster of genes co-occuring in several endospore formers and strongly enriched in sporulation genes *csfG* is present in a wide range of *Bacilli* and *Clostridia*. It is always found immediately upstream of and diverging from a *ylbH* ortholog. We identified a *csfG* ortholog in all the *Bacillaceae* that we analyzed (20), in *Alicyclobacillaceae* and in *Paenibacillaceae*, but not in *Planococcaceae* (e.g. *Lysinibacillus sphaericus*). We also found a *csfG* ortholog in various *Clostridia*, all of them (but *Dethiobacter alkaliphilus*) being thermophiles. Conversely, *csfG* is absent from bacteria that do not form endospores.

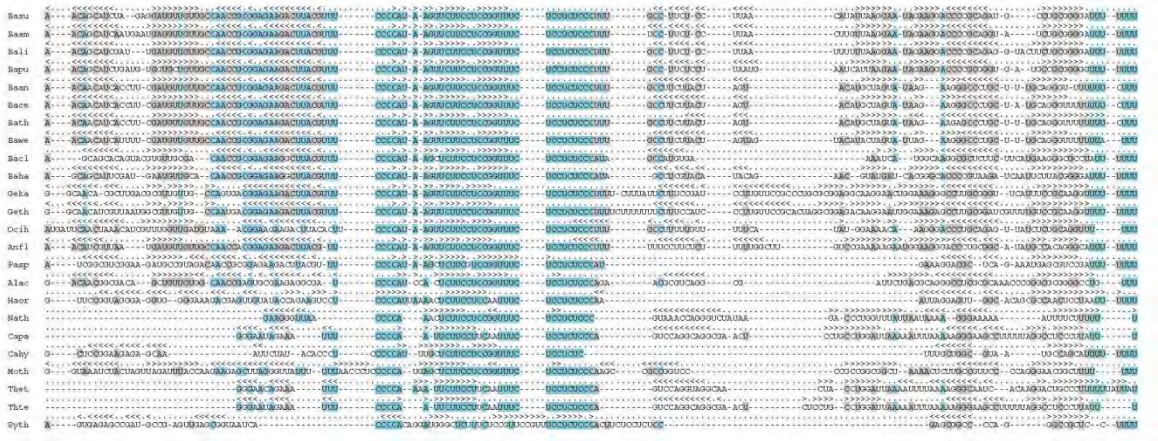


Figure 4. Multiple structural alignment of 24 CsfG orthologs.
Bacterial names are indicated by the their first letters and correspond to *B. subtilis, B. amyloliquefaciens, B. licheniformis, B. pumilus, Bacillus anthracis, Bacillus cereus, Bacillus thuringiensis, Bacillus weihenstephanensis, Bacillus clausii, Bacillus halodurans, Geobacillus kaustophilus, Geobacillus thermodenitrificans, Oceanobacillus iheyensis, Anoxybacillus flavithermus, Paenibacillus sp. JDR-2, Alicyclobacillus acidocaldarius, Halothermotrix orenii, Natranaerobius thermophilus, Carboxydibrachium pacificum, Carboxydothermus hydrogenoformans, Moorella thermoacetica, Thermoanaerobacter ethanolicus, Thermoanaerobacter tengcongensis, Symbiobacterium thermophilum*. Text is highlighted in a turquoise/blue/grey gradient to indicate regions of high/medium/low conservation. Arrowheads indicate potential base pairing.

A multiple sequence alignment of 24 CsfG orthologs is shown in Fig. 4 where potential base pairing is indicated. Whereas CsfG orthologs from closely related *Bacilli* are conserved throughout their whole length (Fig. 1b) this extended alignment pinpoints the striking conservation of a central sequence that constitutes the second half of a stem-loop structure and its adjacent unpaired region. All CsfG sequences are predicted to also contain a rho-independent transcription terminator. The remaining predicted secondary structures are variable, with fewer stem loops in sequences originating from *Clostridia* in accordance with their shorter size. Examples of the predicted secondary structures of a few CsfG RNAs are shown in Fig. 5 where the conserved central motif is highlighted.

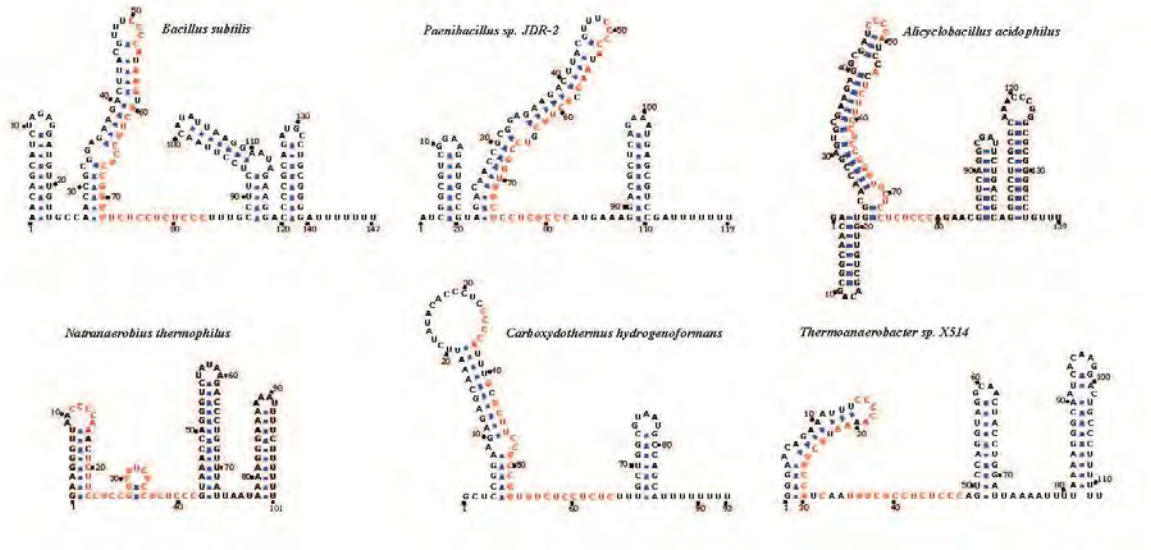


Figure 5. Predicted secondary structure of representative CsfG RNAs.
A VARNA view (20) of the predicted secondary structure of CsfG RNAs from six distant endospore formers is shown. Highly conserved nucleotides are indicated in red.

All CsfG sequences from Bacilli share an extended conserved motif consisting of UCCCCAU, followed at 4 nt by UCUUC, and one nt further by a strictly conserved 18-residue, pyrimidine-rich sequence. A shorter subset of this motif is present in the CsfG sequences from *Clostridia*. The conserved CCCCA block is predicted to be in a loop, the UCUUC block to be included in a stem, and the long UC-rich sequence to be at least partially unpaired. The remarkable conservation of this sequence and its specific arrangement suggest that this part of CsfG could be interacting with an RNA target, following a strand displacement mechanism similar to what has been described for DsrA in *E. coli*.[16] It could also be the recognition motif for an RNA binding protein as previously described for some other highly conserved RNAs.[8]

An elusive physiological role

CsfG is synthesized exclusively in the forespore all along the sporulation process, first by σ^F then by σ^G. Its presence in all *Bacillaceae* and in several other endospore formers, including bacteria that diverged from *B. subtilis* billions of years ago, indicates that CsfG was present in the earliest endospore formers and

that its preservation is somehow beneficial for the organism. If CsfG plays a negative role on expression of some targets, as it is often the case with *trans*-acting sRNAs, we can only speculate that decreasing the synthesis of some vegetative protein(s) may contribute to a more robust and efficient sporulation program.

Interestingly, in several *Clostridia* where *csfG* is missing, it is replaced upstream of *ylbH* by *gpr*, a gene that is also transcribed exclusively in the forespore by σ^F- and σ^G-associated RNA polymerase. In some *Clostridia* containing *csfG* the *gpr* gene is located only a few kb downstream of *csfG*, next to a small cluster of σ^G-controlled *ssp* genes. These genes encode small proteins stored in the mature spore where they become the substrate of the Gpr protease during germination. These genomic relationships might reflect an ancient genetic linkage and they hint at a possible functional involvement of CsfG in germination.

Whatever its role it is noteworthy that the CsfG sRNA was initially identified by a computational screen for elements showing phylogenetic co-occurrence in endospore formers. Making a spore involves a large number of genes, but only a subset of them lead to a clearcut sporulation defect when disrupted.[4] Moreover, despite the conservation among endospore formers of a core of sporulation genes, mostly the main regulators and their principal targets, the presence or absence of a large number of sporulation-expressed genes depend on the specific habitat of the bacterium.[7] Phylogenetic profiles used in the NAPP algorithm are particularly fit to capture such complex patterns as shown by the identification of CsfG.

Materials and Methods

Computational procedures

The complete NAPP procedure is described in Marchais *et al.*[14] In short, coding sequences (CDS) and conserved non-coding elements (CNE) from a species of interest (here *B. subtilis*) are sought systematically in 420 bacterial genomes using Blastn,[17] producing a profile of normalized Blast scores. We then measure pairwise Pearson's distances between profiles and we cluster profiles using either K-means or hierarchical clustering. The —sporulation" cluster was identified by visual inspection of the hierarchical clustering representation, as part of a larger cluster enriched in known noncoding RNAs identified by Marchais *et al.*[14] Searches for CsfG homologs were performed against the NCBI nr databases using Blastn. The multiple structural alignment of CsfG homologs was performed using the LocARNA web server[18] and further manually edited using the Ralee Emacs plugin[19] to correct local errors. Drawing and editing of RNA secondary structure was performed with the VARNA software.[20] The CsfG entry was submitted to the RFAM database.[21] A pre-release is available at: http://fr.wikipedia.org/wiki/Utilisateur:Amarchais/CsfG_RNA.

Experimental procedures

Bacterial growth conditions and genetic methods were as described in a recent work.[22] All *B. subtilis* strains are derivatives of JH642. The *sig* mutants are from our laboratory collection. Sporulation was induced by exhaustion in Difco sporulation medium (DSM) in the absence of antibiotics. Induction of σ^F or σ^G activity during exponential growth was carried out in 2xYT medium, in strains harboring a plasmid with the *sigF* or the *sigG* sequence under the control of the *spac* promoter, by addition of 1-mM isopropyl β-D-thiogalactoside (IPTG) at OD_{570} = 0.25. β-galactosidase activity is expressed as nanomoles of 2-nitrophenyl-β-D-galactopyranoside hydrolysed per minute per milligram of protein.

The *csfG* region was cloned by PCR using oligonucleotides (sequences available on request) that created an *Eco*RI site 130-bp upstream of the predicted

csfG transcription start. The natural *Sac*II site located at position 32 in CsfG was used to create a transcriptional fusion with *lacZ* after cloning into the pDG1661 vector that allows integration by a double recombination event at the *amyE* locus.[23] Similarly, a *Nco*I site and a *Pst*I site were introduced by PCR at position -9 and +123 respectively, according to the predicted *csfG* transcription start. A kanamycin resistance cassette was cloned between these two restriction sites and transferred onto the chromosome by a double recombination event that led to inactivation of *csfG*.

RNA was prepared according to a published protocol.[24] Total RNA (10 µg) was separated on a denaturing 8% polyacrylamide-8 M urea gel and transferred to a Hybond-N^+ Membrane (GE Healthcare) for 4 h at 60 V in 0.5X TBE. The membrane was pre-hybridized overnight in Ultrahyb buffer (Ambion) and probed with a ^{32}P-labelled oligonucleotide (GGAGAAGGCAAAGGGAGAGGAGAAACCG GAGG). Hybridization and wash steps were as described previously.[25] Briefly, the membrane was hybridized 4 h at 42°C in Ultrahyb buffer, washed once in 2X SSC, 0.01% SDS for 10 minutes at 42 °C and 5 times at room temperature in 0.2X SSC, 0.01% SDS. The signal was revealed using a Typhoon Trio (GE Healthcare).

Acknowledgements

We are grateful to C. Condon for his advice and encouragement. This work was supported by CNRS (IFR 115 and UPR9073).

References

1. Errington, J. (2003). Regulation of endospore formation in *Bacillus subtilis*. *Nature Rev. Microbiol.* **1**, 117-126.

2. Eichenberger, P., Fujita, M., Jensen, S. T., Conlon, E. M., Rudner, D. Z., Wang, S. T. *et al.* (2004). The program of gene transcription for a single differentiating cell type during sporulation in *Bacillus subtilis*. *PLoS Biol.* **2,** 1664-1683.

3. Steil, L., Serrano, M., Henriques, A. O. & Völker, U. (2005). Genome-wide analysis of temporally regulated and compartment-specific gene expression in sporulating cells of *Bacillus subtilis*. *Microbiology,* **151,** 339–420.

4. Wang, S. T., Setlow, B., Conlon, E. M., Lyon, J. L., Imamura, D., Sato, T., Setlow, P., Losick, R. & Eichenberger, P. (2006). The forespore line of gene expression in *Bacillus subtilis*. *J. Mol. Biol.* **358,** 16-37.

5. Hilbert, D. W. & Piggot, P. J. (2004). Compartmentalization of gene expression during *Bacillus subtilis* spore formation. *Microbiol. Mol. Biol. Rev.* **68,** 234-262.

6. Paredes, C. J., Alsaker, K. V. & Papoutsakis, E. T. (2005). A comparative genomic view of clostridial sporulation and physiology. *Nature Rev. Microbiol.* **3**, 969-978.

7. Stragier, P. (2001). A gene odyssey: exploring the genomes of endospore-forming bacteria. In Bacillus subtilis *and its Closest Relatives: From Genes to Cells* (Sonenshein, A. L., Hoch, J. A. & Losick, R., eds), pp. 519-526. ASM Press, Washington.

8. Babitzke, P. & Romeo, T. (2007). CsrB sRNA family: sequestration of RNA-binding regulatory proteins. *Curr. Opin. Microbiol.* **10,** 156-163.

9. Yu, Y.-T. N., Yuan, X. & Velicer, G. J. (2010). Adaptive evolution of an sRNA that controls *Myxococcus* development. *Science,* **328,** 993.

10. Toledo-Arana, A., Repoila, F. & Cossart, P. (2007). Small noncoding RNAs controlling pathogenesis. *Curr. Opin. Microbiol.* **10,** 182-188.

11. Silvaggi, J. M., Perkins, J. B. & Losick, R. (2006). Genes for small, noncoding RNAs under sporulation control in *Bacillus subtilis*. *J. Bacteriol.* **188,** 532-541.

12. Saito, S., Kakeshita, H. & Nakamura, K. (2009). Novel small RNA-encoding genes in the intergenic regions of *Bacillus subtilis*. *Gene,* **428,** 2-8.

13. Rasmussen, S., Nielsen, H. B. & Jarmer, H. (2009). The transcriptionally active regions in the genome of *Bacillus subtilis*. *Mol. Microbiol.* **73,** 1043-1057.

14. Marchais, A., Naville, M., Bohn, C., Bouloc, P. & Gautheret, D. (2009). Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles. *Genome Res.* **19,** 1084-1092.

15. Barrick, J. E., Corbino, K. A., Winkler, W. C., Nahvi, A., Mandal, M., Collins, J. *et al.* (2004). New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc. Natl. Acad. Sci. USA,* **101,** 6421-6426.

16. Lease, R. A., Cusick, M. E. & Belfort, M. (1998). Riboregulation in *Escherichia coli*: DsrA RNA acts by RNA:RNA interactions at multiple loci. *Proc. Natl. Acad. Sci. USA,* **95,** 12456-12461.

17. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215,** 403-410.

18. Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F. & Backofen, R. (2007). Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.* **3,** e65.

19. Griffiths-Jones, S. (2005). RALEE—RNA ALignment Editor in Emacs. *Bioinformatics,* **21,** 257-259.

20. Darty, K., Denise, A. & Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics,* **25,** 1974-1975.

21. Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., Wilkinson, A. C., Finn, R. D., Griffiths-Jones, S., Eddy, S. R. & Bateman, A. (2009). Rfam: updates to the RNA families database. *Nucleic Acids Res.* **37,** D136-140.

22. Karmazyn-Campelli, C., Rhayat, L., Carballido-López, R., Duperrier, S., Frandsen, N. & Stragier, P. (2008). How the early sporulaiton sigma factor σ^F delays the switch to late development in *Bacillus subtilis*. *Mol. Microbiol.* **67,** 1169-1180.

23. Guérout-Fleury, A.-M., Frandsen, N. & Stragier, P. (1996). Plasmids for ectopic integration in *Bacillus subtilis*. *Gene,* **180,** 57-61.

24. Bechhofer, D. H., Oussenko, I., Yao, S., Mathy, N. & Condon, C. (2008). Analysis of mRNA decay in *Bacillus subtilis*. In *Methods in Enzymology* (Maquat, E. & Arraiano, C. M., eds), Vol. 447, pp. 259-276. Academic Press, New York.

25. Opdyke, J. A., Kang, J.-G. & Storz, G. (2004). GadY, a small-RNA regulator of acid response genes in *Escherichia coli* . *J. Bacteriol.* **186,** 6698-6705.

ARTICLES EN CO-AUTEUR

Afin de préciser ma contribution à ces deux articles présentés en annexe, j'en insère ci-après les résumés.

EXPERIMENTAL DISCOVERY OF SMALL RNAS IN STAPHYLOCOCCUS AUREUS REVEALS A RIBOREGULATOR OF CENTRAL METABOLISM

Chantal Bohn, Candice Rigoulay, Svetlana Chabelskaya, Cynthia M. Sharma, Antonin Marchais, Patricia Skorski, Elise Borezée-Durant, Romain Barbet, Eric Jacquet, Annick Jacq, Daniel Gautheret, Brice Felden, Jörg Vogel and Philippe Bouloc.
Published in Nucleic Acid Research, 2010.[87]

Résumé

Ce travail a été réalisé en collaboration avec l'équipe Signalisation et Réseaux de régulations bactériens de Philippe Bouloc à l'Institut de génétique et de microbiologie et le laboratoire de Jorg Vogel du « Institute for Molecular Infection Biology » de Berlin. Il est consacré à la recherche de nouveaux ARNnc chez *S. aureus* et à la description plus fine de l'un d'entre eux, RsaE. A la différence des travaux précédents, cette recherche s'appuie sur des données de séquençage massif pour la partie concernant la détection des ARNnc et sur des données de MicroArray pour étudier les régulations post-transcriptionnelles induites par RsaE. Ce travail a permis de valider expérimentalement 12 nouveaux ARNnc chez *S. aureus* et de caractériser de nombreuses cibles pour RsaE qui implique cet ARNnc dans la modulation de la voie métabolique du methenyl tetrahydrofolate (THF) et du cycle de Krebs. J'ai réalisé la prédiction *in silico* des cibles de RsaE qui, couplée aux résultats de MicroArray a permis d'identifier les cibles les plus probables. A cet effet, j'ai mis au point et réalisé les calculs de probabilité permettant de confirmer le recouvrement significatif entre la liste des gènes régulés dans les expériences de MicroArray et ceux prédits comme des cibles putatives.

ARNOLD: A WEB TOOL FOR THE PREDICTION OF RHO-INDEPENDENT TRANSCRIPTION TERMINATORS

Magali Naville, Adrien Ghuillot-Gaudeffroy, Antonin Marchais and Daniel Gautheret.

In press, RNA Biology, 2011.

Résumé

Ce travail décrit la réalisation d'un site web mettant à disposition un outil en ligne de prédiction des terminateurs Rho-indépendants dans les séquences fournies par l'utilisateur. Il implique uniquement des personnes de notre équipe et notamment Magali Naville qui a développé cet outil dans le cadre de ses recherches sur les atténuateurs transcriptionnels[126]. Cet outil de prédiction combine les résultats de recherche de terminateurs Rho-indépendants produits par les logiciels Erpin[127] et RNAmotif[128]. Ma participation, en dehors de la résolution de problèmes techniques inhérents au développement d'une page web dynamique, a été de réaliser les études statistiques permettant de définir et comparer la sensibilité et la spécificité de cet outil à celles du logiciel TranstermHP[129]. J'ai choisi pour calculer la spécificité de chacune de ces méthodes un jeu de séquence sans terminateur différent de celui utilisé couramment pour ce type d'étude. En effet, au lieu de calculer une spécificité sur des séquences intra-cistroniques, qui maintenant sont connus pour contenir ce type de structure, j'ai générés des séquences aléatoires de composition en di-nucléotides similaires à un jeu de 425 séquences de terminateur de *B. subtilis* validés expérimentalement[130]. Pour la sensibilité nous avons directement utilisé les séquences des terminateurs validés expérimentalement chez *B. subtilis*.

DISCUSSION

Mes travaux de doctorat se sont révélés fructueux à plusieurs titres. Du point de vue bioinformatique, nous sommes les premiers à avoir démontré que le profilage phylogénétique appliqué aux séquences d'acide nucléique pouvait produire une classification pertinente des séquences intergéniques conservées et, comme pour les séquences d'acide aminé, une indication sur leur fonction. Du point de vue du biologiste, nous pouvons nous réjouir d'avoir décrit un certain nombre de nouveaux ARNnc chez *S. aureus* et *B. subtilis* et d'avoir pu aider à la détermination de nouvelles cibles pour l'ARNnc RsaE. Toutefois, ce travail de doctorat se caractérise par l'émergence de nouveaux questionnements qui se révèlent encore plus nombreux et complexes à solutionner qu'il ne l'était à l'origine du projet. A cela s'ajoute la courte durée de la thèse qui ne permet pas d'obtenir tous les résultats voulus et laisse donc des perspectives nombreuses à la poursuite éventuelle du travail. Les articles insérés dans ma thèse possèdent tous un chapitre de discussion des résultats que je ne souhaite pas paraphraser ici. Je préfère aborder dans cette discussion mes travaux en cours qui ont pour objectif de poursuivre l'analyse des ARNnc décrits chez *S. aureus* et *B. subtilis* et d'optimiser la méthode NAPP. Dans un deuxième temps, je résumerai certaines questions restées en suspens à la suite de mon travail et qui me permettront de développer une conclusion plus générale sur les ARNnc bactériens et leur intégration dans les systèmes biologiques.

Parmi les ARNnc détectés par NAPP et validés expérimentalement chez *S. aureus*, nous nous sommes intéressés plus spécifiquement à RsaOG car il présente une conservation spécifique au genre Staphylococcus et un fort taux de transcription. Actuellement nous poursuivons nos investigations sur RsaOG en collaboration avec Philippe Bouloc et Chantal Bohn de l'institut de génétique et de microbiologie. Nous travaillons sur la confirmation des cibles que nous avons prédite, mais aussi sur une hypothèse que j'ai formulée en croisant des observations structurales de conservation de la boucle interne et de présence putative d'un pseudo-nœud avec le rôle putatif de cet ARN dans la régulation post-transcriptionnelle du gène *gmk*. En effet, cette séquence de boucle interne (Figure 10) ne participe jamais à l'appariement avec les cibles prédites mais présente, comme la région simple brin avec laquelle elle s'apparie en pseudo-nœud, une conservation presque totale.

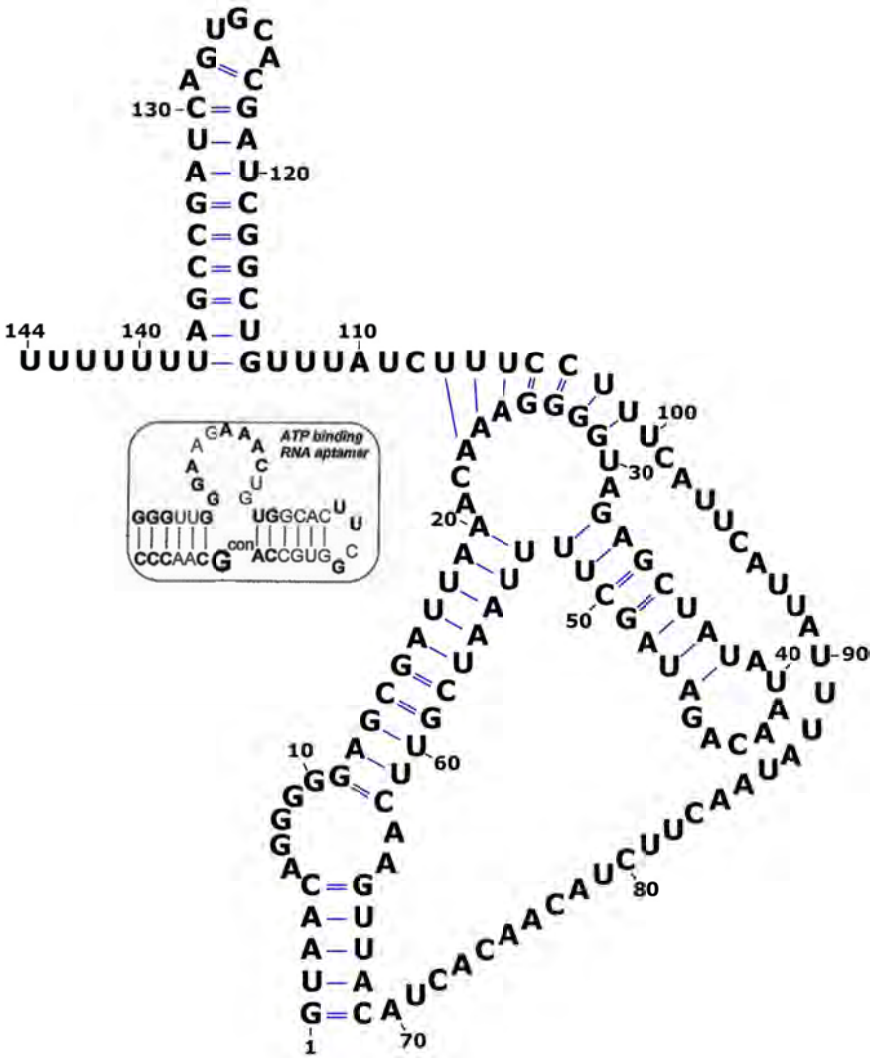


Figure 11. Structure secondaire de l'ARNnc RsaOG de *S. aureus* prédite par Mfold[131] et complétée du pseudo-nœud prédit par MCfold[132]. L'encadré est illustre la boucle interne de l'aptamère du petit ARN synthétique sélectionné par SELEX pour sa capacité de liaison à l'ATP[133]. Cette boucle interne est de même longueur et de même composition en nucléotide que la boucle interne de RsaOG impliquée dans la formation du pseudo-nœud. Dans l'ARNnc RsaOG de *S. saprophyticus*, cette boucle interne est dupliquée intégralement.

En me demandant si cette boucle avait la possibilité d'interagir avec un ligand comme le font les riboswitchs, j'ai réalisé au cours d'un séminaire de Jean Lehmann (IGM), où il était question de l'interaction d'ARN synthétique avec l'ATP, que la séquence de la boucle de RsaOG était très similaire et avait exactement la même composition en nucléotide que la boucle d'un ARN synthétique obtenu par SELEX et impliquée dans l'interaction avec l'ATP (Figure 11). Si l'on ajoute l'observation selon laquelle la cible prédite pour former l'interaction la plus stable avec RsaOG est *gmk*, une guanylate kinase directement impliquée dans la catalyse de la réaction ATP + GMP → ADP + AMP, on peut imaginer un modèle dans lequel RsaOG jouerait un rôle de

capteur pour l'un des éléments de cette réaction. Pour tester cette hypothèse nous sommes en train de réaliser des études de microcalorimétrie qui permettent d'observer de légères variations de chaleur induites par des interactions moléculaires.

D'autres ARNnc détectés par NAPP chez *S. aureus* ont retenu notre attention. On peut citer RsaOA qui est le résultat d'une fusion certainement récente entre la région 5' d'un transposon et une partie du gène *usp* (universal stress protein). Cet ARNnc semble avoir beaucoup divergé dans certaines souches de *S. aureus* ou le transposon en question semble plus actif. J'ai donc fait l'hypothèse que cet ARNnc pouvait être impliqué dans le contrôle de la dissémination du transposon. Cependant des études de fonds sont bien sûr nécessaires avant de pouvoir construire un modèle solide quant au rôle et mécanisme d'action de RsaOA. Le séquençage en cours d'une cinquantaine de souche de *S. aureus* par le Broad Institute pourra probablement dans les mois à venir nous permettre de réaliser une étude phylogénétique préalable intéressante de RsaOA et du transposon dont il provient. On pourra ainsi tenter de savoir si l'activité de ce dernier est associée à la présence de RsaOA.

Chez *B. subtilis*, notre stratégie d'utilisation de NAPP pour détecter et prédire la fonction nous a permis de valider expérimentalement l'ARNnc CsfG. L'implication de cet ARN dans la sporulation reste pour le moment putative mais des expériences conjointes de recherches de cible *in silico* et *in vivo* sont en cours. Une des hypothèses les plus raisonnables serait qu'à l'instar d'autres ARNnc des firmicutes[87], CsfG module négativement la traduction de gènes non-essentiels à la sporulation aboutissant à une meilleure gestion des ressources de la bactérie. Si tel était le cas, nous validerions de manière très convaincante la capacité de NAPP à prédire la fonction d'un ARNnc.

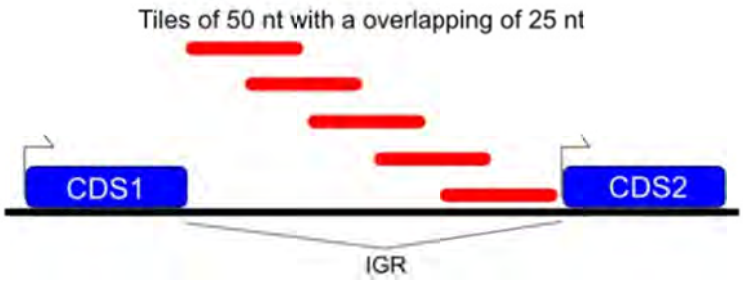


Figure 12. Représentation schématique de l'extraction de tuiles de 50 nucléotides recouvrantes sur 25 nucléotides dans une région située entre deux séquences codantes. Le recouvrement important des tuiles entre elles est selon nous nécessaire pour qu'un maximum de tuiles soit représentatif d'éléments génomiques sous la contrainte d'une pression de sélection homogène. (CDS : séquence codante ; IGR : région intergénique)

Nous avons développé NAPP avec l'objectif d'identifier un nouveau type de signature des ARNnc dans les génomes bactériens avec un minimum d'a priori. Or, l'extraction des CNE dans les génomes, qui est la première étape du pipeline NAPP, est un a priori important puisque l'on recherche des séquences conservées avec un seuil de conservation définit empiriquement. Des ARNnc peu conservés peuvent donc échapper à nos critères. Comme nous l'a fait remarquer Alain Denise au cours d'une séance du club bioinformatique de l'IGM, il suffit, pour outrepasser ce biais, d'utiliser à la place des CNE des tuiles chevauchantes obtenues en découpant les régions intergéniques sans se préoccuper de leur conservation (Figure 12). Nous avons donc implémenté ce protocole. Les résultats produits par NAPP en utilisant des tuiles de 50 nucléotides chevauchantes de 25 nucléotides, nous ont permis d'obtenir plus de groupes enrichis en ARNnc, avec pour les deux meilleurs groupes, des enrichissements plus importants que dans notre étude initiale. Aussi, nous avons maintenant intégré ce tuilage dans le pipeline de NAPP.

Pour quelle raison l'utilisation de tuiles plutôt que des CNE produit-elle un meilleur enrichissement en ARNnc et un plus grand nombre de groupes enrichis ? Il est probable que les tuiles aient une meilleure capacité à représenter une région conservée de manière hétérogène. En effet, vu leur petite taille, les tuiles ont statistiquement plus de chance de correspondre à de petites régions sous une pression de sélection homogène. A contrario de nombreux CNE avaient tendance à recouvrir une région de conservation variable. Le profil phylogénétique construit à partir de ce type de CNE représentait donc une somme de profils phylogénétiques hétérogènes. Une autre observation, plus inattendue, et certainement due à ce même phénomène, est la classification des tuiles recouvrant un même ARNnc dans des groupes différents. Ce dernier résultat suggère que les ARNnc soient, comme les protéines, formés de différent domaines

fonctionnels. Ces domaines pourraient être spécifiques d'interactions avec des protéines comme HFQ ou des hélicases à ARN, ou encore d'appariement avec des cibles ARNm. Chacun de ces domaines serait alors classé phylogénétiquement avec les gènes codants ou tuiles qui intéragissent avec eux. Une analyse statistique plus détaillée du clustering des tuiles chez plusieurs espèces est nécessaire pour confirmer ces hypothèses.

Un autre développement en cours du projet NAPP est la nouvelle exécution du pipeline avec un jeu récent de 1070 génomes bactériens séquencés et stockés sur le site du NCBI. NAPP est conçu pour que le temps d'analyse d'un génome n'augmente que linéairement avec le nombre de génomes utilisés dans le profil. Une conséquence de l'ajout de nouveaux génomes est l'amélioration de la qualité des profils phylogénétiques produits. En effet, en prenant en compte des espèces intermédiaires qui permettent de mieux couvrir l'histoire évolutive des Bactéries, nous devrions encore gagner en résolution et, avec les nombreuses souches séquencées pour certaines espèces, arriver à détecter des ARNnc spécifiques à des espèces, voire des souches.

La présence de différents groupes enrichis en ARNnc suggère qu'il existe plusieurs niveaux d'intervention des ARNnc dans les réseaux de régulations. En effet, ces groupes de co-évolution sont plus ou moins conservés dans les espèces. Ceci rappelle les observations réalisées *in vivo* selon lesquelles certains ARNnc interagissent avec des régulateur centraux comme *rpos* ou *fur* alors que d'autres interviennent dans des réseaux spécifiques comme le quorum-sensing ou la sporulation. Cependant, quelque-soit l'espèce de référence utilisée nous retrouvons toujours un ou deux groupes réunissant la majorité des ARNnc, des gènes de ménage liés aux ARN et quelques régulateurs majeurs des processus cellulaires. Cette observation suggère que ces ARNnc sont contraints par la pression de sélection à évoluer d'une manière similaire indépendamment de l'espèce de référence étudiée. Ceci nous a conduit à réfléchir à une méthode de détection des ARNnc plus simple techniquement et indépendante d'une espèce de référence. Le laboratoire travaille actuellement à une réinterprétation des profils phylogénétiques afin d'identifier d'éventuelles signatures évolutives des ARNnc caractérisées par une certaine vitesse d'évolution ou « decay rate ».

Avant de conclure, je souhaite revenir sur une des grandes problématiques émergentes de ces dernières années dans la communauté « ARN ». Avec les méthodes de séquençage massif, la

détection de signaux de transcription dans les génomes est devenue de plus en plus triviale et cette avancée technologique a entrainé logiquement un glissement de la recherche vers l'étape suivant la caractérisation, c'est-à-dire, pour les ARNnc agissant en *trans*, la détermination des cibles ARNm. Nous aurions pu espérer avec NAPP obtenir des groupes de « co-évolution » dans lesquels les ARNnc agissant en *trans* se classeraient avec leurs cibles. Or, des études statistiques, au cours desquelles nous avons confronté les données expérimentales de cibles validées à nos groupes enrichis en ARNnc agissant en *trans*, ne nous ont pas permis d'observer de corrélation entre les profils phylogénétique des ARNnc et ceux de leurs cibles. Plusieurs phénomènes expliquent sans doute ce résultat négatif. L'un d'eux peut être le suivant. Au cours de mes travaux de recherche de cibles pour RsaE, CsfG et d'autres ARNnc chez *E. coli*, je me suis aperçu que si je réalisais cette recherche dans le génome d'une espèce proche en utilisant l'ARNnc homologue, souvent très conservé, les cibles variaient très rapidement. Ces observations très préliminaires peuvent s'expliquer par la faiblesse de l'appariement ARNnc-cible qui implique fréquemment de nombreux mésappariements et pourrait donc conférer une grande flexibilité au système. En effet, l'ARNnc agissant en *trans* possède généralement une région de « compromis » très conservée puisqu' elle interagit avec plusieurs cibles selon différents appariements et se trouve donc très contrainte par la nécessité de satisfaire toutes ces interactions. A contrario, la contrainte moindre exercée sur les cibles rend le système ARNnc/cible plus adaptable à une modification de l'environnement de la bactérie, car des mutations ponctuelles peuvent facilement entrainer la perte ou la création de cibles. A la lumière de ce raisonnement, les cibles pourraient ne pas évoluer de la même façon que les ARNnc les régulant. NAPP, qui exploite la co-évolution des séquences génomiques, ne serait donc pas applicable à la recherche de cible. Cependant, la variabilité des cibles ne signifie pas obligatoirement l'absence de signatures évolutives discriminantes pour ces éléments. En ce moment même, Alban Ott, un doctorant en première année dans notre équipe, s'intéresse à la détermination d'une signature en 5' des ARNm ciblés par des ARNnc.

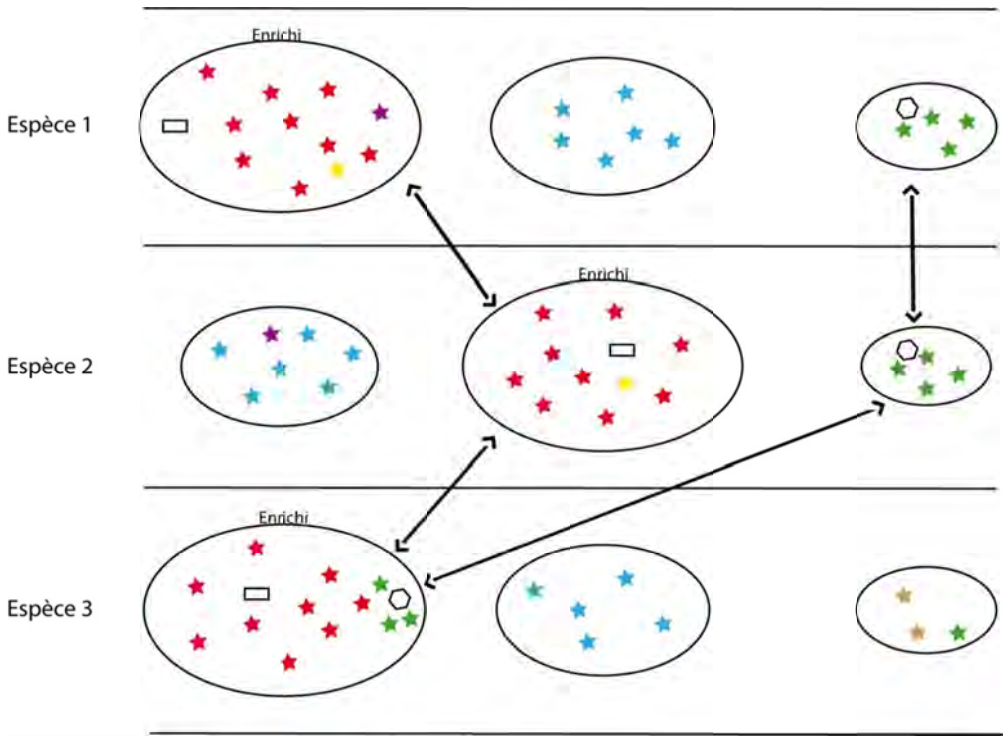


Figure 13. Représentation schématique, pour trois espèces, de trois « cluster » contenant respectivement des éléments génomiques avec des profils similaires. Les étoiles illustrent ces éléments génomiques. Des couleurs identiques correspondent à une similarité des séquences. Les groupes avec le plus d'étoiles rouges sont enrichis en ARNnc. Le rectangle et l'hexagone vide représentent des ARNnc. En comparant le contenu en éléments des groupes entre les espèces nous pouvons espérer construire un modèle décrivant au mieux le scénario évolutif à l'origine de ces groupes. De plus cette méthode pourrait nous permettre de prédire un plus grand nombre d'ARNnc. En effet, l'hexagone qui se trouve dans les espèces 1 et 2 serait imprédictible a priori avec NAPP. Toutefois en comparant ces groupes contenant uniquement des étoiles vertes avec les groupes de l'espèce 3 nous pourrions observer que les étoiles vertes co-évoluent dans cette espèce avec les autres éléments du groupe enrichi.

Je souhaite conclure ma thèse sur la perspective qui me semble la plus excitante et qui concerne l'utilisation des données produites par NAPP sur un grand jeu de génomes. Toutes ces données de co-évolution des séquences d'acides nucléiques dans un millier d'espèces bactériennes pourraient permettre, je pense, de mieux comprendre, l'intégration des ARNnc dans les réseaux de régulation et les relations qu'ils entretiennent avec leur différents partenaires, mais aussi de déterminer les chemins évolutifs empruntés par ces ARN et les gènes qui y sont associés. En effet, la comparaison des différents groupes enrichis en ARNnc entre de multiples espèces pourrait être une bonne stratégie pour retracer les évènements évolutifs (Figure 13). De nombreux développements techniques seront probablement nécessaires pour contourner le problème lié à la quantité de données à analyser et la mise en place de méthodes de comparaison qui, idéalement, ne devraient pas impliquer de recherche d'orthologie entre les groupes de différentes espèces. Des analyses de voisinage dans des graphes représentant ces groupes seraient en effet préférables afin

de conforter l'idée sous-jacente du profilage phylogénétique, selon laquelle un orthologue fonctionnel est une séquence qui co-évolue avec un groupe de gènes interdépendants dans le maintien d'un processus.

RÉFÉRENCES

1 **Caspersson, T. & Schultz, J. Pentose Nucleotides in the Cytoplasm of Growing Tissues. *Nature* 143, 602-602 (1939).**

2 **Schultz, J., Caspersson, T. & Aquilonius, L. The Genetic Control of Nucleolar Composition. *Proc Natl Acad Sci U S A* 26, 515-523 (1940).**

3 **Caspersson, T. & Schultz, J. Ribonucleic Acids in Both Nucleus and Cytoplasm, and the Function of the Nucleolus. *Proc Natl Acad Sci U S A* 26, 507-515 (1940).**

4 **Albert, C. The constitution of protoplasm. *Science* 97, 451-456 (1943).**

5 **Levene, P. A., and Bass, L. W. Nucleic acids. *Chemical Catalog Company, New York* (1931).**

6 **Avery, O. T., Macleod, C. M. & McCarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med* 79, 137-158 (1944).**

7 **Hershey, A. D. & Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* 36, 39-56 (1952).**

8 **Dounce, A. L. [Duplicating mechanism for peptide chain and nucleic acid synthesis.]. *Enzymologia* 15, 251-258 (1952).**

9 **Boivin, A. & Vendrely, R. Sur le rôle possible de des deux acides nucléiques dans la cellule vivante. *Experientia* 3, 32-34 (1947).**

10 **Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171, 737-738 (1953).**

11 **Hoagland, M. B. An enzymic mechanism for amino acid activation in animal tissues. *Biochim Biophys Acta* 16, 288-289 (1955).**

12 **Hoagland, M. B., Zamecnik, P. C. & Stephenson, M. L. Intermediate reactions in protein biosynthesis. *Biochim Biophys Acta* 24, 215-216 (1957).**

13 **Schweet, R. S., Bovard, F. C., Allen, E. & Glassman, E. The Incorporation of Amino Acids into Ribonucleic Acid. *Proc Natl Acad Sci U S A* 44, 173-177 (1958).**

14 **Berg, P., Bergmann, F. H. . A enzymatic mechanism for linking amino acids to RNA. *Proc Natl Acad Sci U S A* 44, 78-86 (1958).**

15 **Allen, E. H., Glassman, E., Cordes E. and Schweet, R. S. Incorporation of amino acids into ribonucleic acid. II. Amino Acid transfer ribonucleic acid. *J. Biol. Chem.* 235, 1068-1074 (1960).**

16 **Fraenkel-Conrat, H. The role of nucleic acid in the reconstitution of active tobacco mosaic virus *J. Am. Chem. Soc.* 78 (1956).**

17 **Fraenkel-Conrat, H., Singer, B. & Williams, R. C. Infectivity of viral nucleic acid. *Biochim Biophys Acta* 25, 87-96 (1957).**

18 **Fraenkel-Conrat, H. & Williams, R. C. Reconstitution of Active Tobacco Mosaic Virus from Its Inactive Protein and Nucleic Acid Components. *Proc Natl Acad Sci U S A* 41, 690-698 (1955).**

19 **Spirin, A. S., Belozerskii, A. N., Shugaeva, N. V. & Vaniushin, B. F. [Studies of species specificity of nucleic acids in bacteria.]. *Biokhimiia* 22, 744-754 (1957).**

20 **Volkin, E. & Astrachan, L. Intracellular distribution of labeled ribonucleic acid after phage infection of Escherichia coli. *Virology* 2, 433-437, doi: (1956).**

21 **Volkin, E. & Astrachan, L. Phosphorus incorporation in Escherichia coli ribo-nucleic acid after infection with bacteriophage T2. *Virology* 2, 149-161, doi: (1956).**

22 **Volkin, E., Astrachan, L. & Countryman, J. L. Metabolism of RNA phosphorus in Escherichia coli infected with bacteriophage T7. *Virology* 6, 545-555, doi: (1958).**

23 **Volkin, E. The Function of Rna in T2-Infected Bacteria. *Proc Natl Acad Sci U S A* 46, 1336-1349 (1960).**

24 **Ycas, M. & Vincent, W. S. A Ribonucleic Acid Fraction from Yeast Related in Composition to Desoxyribonucleic Acid. *Proc Natl Acad Sci U S A* 46, 804-811 (1960).**

25 **Hall, B. D., Spiegelman, S. Sequence complementarity of T2-DNA and T2-specific RNA. *Proc Natl Acad Sci U S A* 47, 137-163 (1961).**

26 **Hayashi, M. & Spiegelman, S. The selective synthesis of informational RNA in bacteria. *Proc Natl Acad Sci U S A* 47, 1564-1580 (1961).**

27 **Yankofsky, S. A. & Spiegelman, S. The identification of the ribosomal RNA cistron by sequence complementarity. I. Specificity of complex formation. *Proc Natl Acad Sci U S A* 48, 1069-1078 (1962).**

28 **Yankofsky, S. A. & Spiegelman, S. The identification of the ribosomal RNA cistron by sequence complementarity. II. Saturation of and competitive interaction at the RNA cistron. *Proc Natl Acad Sci U S A* 48, 1466-1472 (1962).**

29 **Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3, 318-356 (1961).**

30 **Brenner, S., Jacob, F. & Meselson, M. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* 190, 576-581 (1961).**

31 **Takanami, M. Transfer of amino acids from soluble ribonucleic aci to ribosome. III. Further studies on the interaction between ribosome and soluble ribonucleic aci. *Biochim Biophys Acta* 61, 432-444 (1962).**

32 **Takanami, M. Transfer of amino acids from soluble ribonucleic acid to ribosome. II. Transfer of soluble ribonucleic acid to ribosome. *Biochim Biophys Acta* 55, 132-138 (1962).**

33 **Weiss, S. B., and Gladstone, L. A mammalian system for the incorporation of cytidine triphospahate into ribonucleotide acid. *J. Am. Chem. Soc.* 81, 4118-4119 (1959).**

34 **Hurwitz, J., Bresler, A. and Diringer, R. The enzymatic incroporation of ribonucleotides into polyribonucleotides and the effect of DNA. *Biochem Biophys Res Commun* 3, 15-19 (1960).**

35 **Weiss, S. B. Enzymatic Incorporation of Ribonucleoside Triphosphates into the Interpolynucleotide Linkages of Ribonucleic Acid. *Proc Natl Acad Sci U S A* 46, 1020-1030 (1960).**

36 **Roberts, R. B., Britten, R. J., and McCarthy, B. J. in *Molecular Genetics* 291-352 (Academic Press, New York and London, 1963).**

37 **Holley, R. W. Structure of an alanine transfer ribonucleic acid. *JAMA* 194, 868-871 (1965).**

38 **Holley, R. W. *et al.* Structure of a Ribonucleic Acid. *Science* 147, 1462-1465 (1965).**

39 **Madison, J. T., Everett, G. A. & Kung, H. Nucleotide sequence of a yeast tyrosine transfer RNA. *Science* 153, 531-534 (1966).**

40 Madison, J. T., Everett, G. A. & Kung, H. K. On the nucleotide sequence of yeast tyrosine transfer RNA. *Cold Spring Harb Symp Quant Biol* 31, 409-416 (1966).

41 Mirzabekov, A. D., Grunberger, D., Holy, A., Bayev, A. A. & Sorm, F. Recognition of synonym codons by purified yeast tRNA-val fractions and fragments from tRNA1-val. *Biochim Biophys Acta* 145, 845-847 (1967).

42 RajBhandary, U. L., Stuart, A., Faulkner, R. D., Chang, S. H. & Khorana, H. G. Nucleotide sequence studies on yeast phenylalanine sRNA. *Cold Spring Harb Symp Quant Biol* 31, 425-434 (1966).

43 Zachau, H. G., Dutting, D. & Feldmann, H. Nucleotide sequences of two serine-specific transfer ribonucleic acids. *Angew Chem Int Ed Engl* 5, 422, doi: (1977).

49 Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* 12, 1-8, doi: (1977).

50 Gilbert, W. Why genes in pieces? *Nature* 271, 501 (1978).

51 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74, 5463-5467 (1977).

52 Crick, F. H. Biochemical activities of nucleic acids. The present position of the coding problem. *Brookhaven Symp Biol* 12, 35-39 (1959).

53 Cech, T. R., Zaug, A. J. & Grabowski, P. J. In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* 27, 487-496, doi: (1981).

54 Stark, B. C., Kole, R., Bowman, E. J. & Altman, S. Ribonuclease P: an enzyme with an essential RNA component. *Proc Natl Acad Sci U S A* 75, 3717-3721 (1978).

55 Kwek, K. Y. *et al.* U1 snRNA associates with TFIIH and regulates transcriptional initiation. *Nat Struct Biol* 9, 800-805, doi: (2002).

56 Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75, 843-854, doi: (1993).

57 Hamilton, A. J. & Baulcombe, D. C. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* 286, 950-952, doi: (1999).

58 Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* 289, 905-920, doi: (2000).

59 Schluenzen, F. *et al.* Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell* 102, 615-623, doi: (2000).

60 Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* 289, 920-930, doi: (2000).

61 Woese, C. *The genetic code: The molecular basis for genetic expression.* (1967).

62 Gilbert, W. Origin of life: the rna world. *Nature* 319, doi:10.1038/nrmicro1491 (2006).

64 Moore, S. D. & Sauer, R. T. The tmRNA system for translational surveillance and ribosome rescue. *Annu Rev Biochem* 76, 101-124, doi: (2004).

66 Opdyke, J. A., Kang, J. G. & Storz, G. GadY, a small-RNA regulator of acid response genes in Escherichia coli. *J Bacteriol* 186, 6698-6705, doi:10.1128/JB.186.20.6698-6705.2004 (2004).

67 Vogel, J. An RNA trap helps bacteria get the most out of chitosugars. *Mol Microbiol* 73, 737-741, doi:10.1111/j.1365-2958.2009.06806.x (2009).

68 Waters, L. S. & Storz, G. Regulatory RNAs in bacteria. *Cell* 136, 615-628, doi:10.1016/j.cell.2009.01.043 (2009).

69 Sudarsan, N. *et al.* Tandem riboswitch architectures exhibit complex gene control functions. *Science* 314, 300-304, doi:10.1126/science.1130716 (2006).

70 Winkler, W. C., Nahvi, A., Roth, A., Collins, J. A. & Breaker, R. R. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* 428, 263-267, doi:10.1038/nature02362 (2004)

71 Nudler, E. & Mironov, A. S. The riboswitch control of bacterial metabolism. *Trends Biochem Sci* 29, 11-17, doi: (2004).

72 Naville, M. & Gautheret, D. Transcription attenuation in bacteria: theme and variations. *Brief Funct Genomics* 9, 178-189 (2010).

73 Brantl, S. Regulatory mechanisms employed by cis-encoded antisense RNAs. *Curr Opin Microbiol* 10, 102-109, doi:10.1016/j.mib.2007.03.012 (2007).

74 Fozo, E. M., Hemm, M. R. & Storz, G. Small toxic proteins and the antisense RNAs that repress them. *Microbiol Mol Biol Rev* 72, 579-589, Table of Contents, doi:10.1128/MMBR.00025-08 (2008).

75 Tramonti, A., De Canio, M. & De Biase, D. GadX/GadW-dependent regulation of the Escherichia coli acid fitness island: transcriptional control at the gadY-gadW divergent promoters and identification of four novel 42 bp GadX/GadW-specific binding sites. *Mol Microbiol* 70, 965-982, doi:10.1111/j.1365-2958.2008.06458.x (2008).

76 Sharma, C. M. *et al.* The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature* 464, 250-255, doi:10.1038/nature08756 (2010).

77 Filiatrault, M. J. *et al.* Transcriptome analysis of Pseudomonas syringae identifies new genes, noncoding RNAs, and antisense activity. *J Bacteriol* 192, 2359-2372, doi:10.1016/j.tig.2005.05.008 (2005).

79 Sharma, C. M., Darfeuille, F., Plantinga, T. H. & Vogel, J. A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites. *Genes Dev* 21, 2804-2817, doi:10.1101/gad.447207 (2007).

80 Pfeiffer, V., Papenfort, K., Lucchini, S., Hinton, J. C. & Vogel, J. Coding sequence targeting by MicC RNA reveals bacterial mRNA silencing downstream of translational initiation. *Nat Struct Mol Biol* 16, 840-846, doi:10.1038/nsmb.1631 (2009).

81 Repoila, F., Majdalani, N. & Gottesman, S. Small non-coding RNAs, co-ordinators of adaptation processes in Escherichia coli: the RpoS paradigm. *Mol Microbiol* 48, 855-861, doi: (2003).

82 Bejerano-Sagie, M. & Xavier, K. B. The role of small RNAs in quorum sensing. *Curr Opin Microbiol* 10, 189-198, doi:10.1016/j.mib.2007.03.009 (2007).

83 Masse, E., Salvail, H., Desnoyers, G. & Arguin, M. Small RNAs controlling iron metabolism. *Curr Opin Microbiol* 10, 140-145, doi:10.1016/j.mib.2007.03.013 (2007).

84 Aiba, H. Mechanism of RNA silencing by Hfq-binding small RNAs. *Curr Opin Microbiol* 10, 134-139, doi:10.1016/j.mib.2007.03.010 (2007).

85 Sittka, A. *et al.* Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet* 4, e1000163, doi:10.1093/nar/gkq462 (2010).

88 Douchin, V., Bohn, C. & Bouloc, P. Down-regulation of porins by a small RNA bypasses the essentiality of the regulated intramembrane proteolysis protease RseP in Escherichia coli. *J Biol Chem* 281, 12253-12259, doi:10.1074/jbc.M600819200 (2006).

89 Masse, E., Vanderpool, C. K. & Gottesman, S. Effect of RyhB small RNA on global iron use in Escherichia coli. *J Bacteriol* 187, 6962-6971, doi:10.1128/JB.187.20.6962-6971.2005 (2005).

90 De Lay, N. & Gottesman, S. The Crp-activated small noncoding regulatory RNA CyaR (RyeE) links nutritional status to group behavior. *J Bacteriol* 191, 461-476, doi:10.1128/JB.01157-08 (2009).

91 Gonzalez-Flecha, B. & Demple, B. Role for the oxyS gene in regulation of intracellular hydrogen peroxide in Escherichia coli. *J Bacteriol* 181, 3833-3836 (1999).

92 Svenningsen, S. L., Waters, C. M. & Bassler, B. L. A negative feedback loop involving small RNAs accelerates Vibrio cholerae's transition out of quorum-sensing mode. *Genes Dev* 22, 226-238, doi:10.1101/gad.1629908 (2008).

93 Boisset, S. *et al.* Staphylococcus aureus RNAIII coordinately represses the synthesis of virulence factors and the transcription regulator Rot by an antisense mechanism. *Genes Dev* 21, 1353-1366, doi:10.1101/gad.423507 (2007).

94 Babitzke, P. & Romeo, T. CsrB sRNA family: sequestration of RNA-binding regulatory proteins. *Curr Opin Microbiol* 10, 156-163, doi:10.1016/j.mib.2007.03.007 (2007).

95 Gorke, B. & Vogel, J. Noncoding RNA control of the making and breaking of sugars. *Genes Dev* 22, 2914-2925, doi:10.1101/gad.1717808 (2008).

96 Sorek, R., Kunin, V. & Hugenholtz, P. CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* 6, 181-186, doi:10.1038/nrmicro1793 (2008).

97 Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709-1712, doi:10.1126/science.1138140 (2007).

98 Cramton, S. E., Schnell, N. F., Gotz, F. & Bruckner, R. Identification of a new repetitive element in Staphylococcus aureus. *Infect Immun* 68, 2344-2348 (2000).

99 Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G. & Gottesman, S. Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev* 15, 1637-1651, doi:10.1016/j.tibtech.2009.09.006 (2010).

100 Rivas, E. & Eddy, S. R. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* 2, 8 (2001).

101 Rivas, E. & Eddy, S. R. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* 16, 583-605 (2000).

102 Gorodkin, J. *et al.* De novo prediction of structured RNAs from genomic sequences. *Trends Biotechnol* 28, 9-19, doi:S0167-7799(09)00180-2

103 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* 215, 403-410 (1990).

104 Washietl, S., Hofacker, I. L. & Stadler, P. F. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A* 102, 2454-2459, doi:10.1073/pnas.0409169102 (2005).

105 Song, D. *et al.* Computational prediction of novel non-coding RNAs in Arabidopsis thaliana. *BMC Bioinformatics* 10 Suppl 1, S36, doi:10.1186/1471-2105-10-S1-S36 (2009).

106 Washietl, S. *et al.* Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res* 17, 852-864, doi:10.1101/gr.5650707 (2007).

107 del Val, C., Rivas, E., Torres-Quesada, O., Toro, N. & Jimenez-Zurdo, J. I. Identification of differentially expressed small non-coding RNAs in the legume endosymbiont Sinorhizobium meliloti by comparative genomics. *Mol Microbiol* 66, 1080-1091, doi:10.1111/j.1365-2958.2007.05978.x (2007).

108 Knudsen, B. & Hein, J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res* 31, 3423-3428 (2003).

109 Pedersen, J. S. *et al.* Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* 2, e33, doi:10.1371/journal.pcbi.0030193 (2007).

113 Mathews, D. H. & Turner, D. H. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol* 317, 191-203, doi: (2002).

114 Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F. & Backofen, R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol* 3, e65, doi:10.1371/journal.pcbi.0030065 (2007).

115 Sankoff, D. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.* 45, 810-825 (1985).

116 Yao, Z., Weinberg, Z. & Ruzzo, W. L. CMfinder--a covariance model based RNA motif finding algorithm. *Bioinformatics* 22, 445-452, doi:10.1093/bioinformatics/btk008 (2006).

117 The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306, 636-640, doi:10.1126/science.1105136 (2004).

118 Torarinsson, E. *et al.* Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. *Genome Res* 18, 242-251, doi:10.1101/gr.6887408 (2008).

119 Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* 14, 708-715, doi: (2004).

120 Livny, J., Fogel, M. A., Davis, B. M. & Waldor, M. K. sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes. *Nucleic Acids Res* 33, 4096-4105, doi:10.1093/nar/gki715 (2005).

121 Livny, J., Teonadi, H., Livny, M. & Waldor, M. K. High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs. *PLoS One* 3, e3197, doi:10.1073/pnas.0503838102 (2005).

123 Marchais, A., Naville, M., Bohn, C., Bouloc, P. & Gautheret, D. Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles. *Genome Res* 19, 1084-1092, doi:10.1101/gr.089714.108 (2009).

124 Marchais, A., Bohn, C., Bouloc, P. & Gautheret, D. RsaOG, a new staphylococcal family of highly transcribed non-coding RNA. *RNA Biol* 7, 116-119, doi: (2010).

125 Muckstein, U. *et al.* Thermodynamics of RNA-RNA binding. *Bioinformatics* 22, 1177-1182, doi:10.1093/bioinformatics/btl024 (2006).

126 Naville, M. & Gautheret, D. Premature terminator analysis sheds light on a hidden world of bacterial transcriptional attenuation. *Genome Biol* 11, R97, doi:10.1186/gb-2010-11-9-r97 (2010).

127 Gautheret, D. & Lambert, A. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *Journal of molecular biology* 313, 1003-1011 (2001).

128 **Macke, T. J. *et al.* RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res* 29, 4724-4735 (2001).**

129 **Kingsford, C. L., Ayanbule, K. & Salzberg, S. L. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol* 8, R22, doi:10.1186/gb-2007-8-2-r22 (2007).**

130 **de Hoon, M. J., Makita, Y., Nakai, K. & Miyano, S. Prediction of transcriptional terminators in Bacillus subtilis and related species. *PLoS Comput Biol* 1, e25, doi: (2003).**

ANNEXES

Nucleic Acids Research Advance Access published May 28, 2010

Nucleic Acids Research, 2010, 1–17
doi:10.1093/nar/gkq462

Experimental discovery of small RNAs in *Staphylococcus aureus* reveals a riboregulator of central metabolism

Chantal Bohn[1], Candice Rigoulay[1], Svetlana Chabelskaya[2], Cynthia M. Sharma[3], Antonin Marchais[1], Patricia Skorski[1], Elise Borezée-Durant[4], Romain Barbet[5], Eric Jacquet[5], Annick Jacq[1], Daniel Gautheret[1], Brice Felden[2], Jörg Vogel[3] and Philippe Bouloc[1,*]

[1]Institut de Génétique et Microbiologie, CNRS/UMR 8621, IFR115, Centre scientifique d'Orsay, Université Paris-Sud, bâtiment 400, 91405 Orsay Cedex, [2]Université de Rennes I, Inserm U835, UPRES EA2311, Biochimie Pharmaceutique, 2 avenue du Prof. Léon Bernard 35043 Rennes, France, [3]Max Planck Institute for Infection Biology, Charitéplatz 1, 10117 Berlin, Germany, [4]Institut Micalis, UMR1319, Institut National de la Recherche Agronomique, Domaine de Vilvert, 78352 Jouy en Josas Cedex and [5]Centre de recherche de Gif, Institut de Chimie des Substances Naturelles, UPR2301 CNRS, IMAGIF, qPCR-Platform, 91198 Gif sur Yvette Cedex, France

Received January 31, 2010; Revised April 5, 2010; Accepted May 11, 2010

ABSTRACT

Using an experimental approach, we investigated the RNome of the pathogen *Staphylococcus aureus* to identify 30 small RNAs (sRNAs) including 14 that are newly confirmed. Among the latter, 10 are encoded in intergenic regions, three are generated by premature transcription termination associated with riboswitch activities, and one is expressed from the complementary strand of a transposase gene. The expression of four sRNAs increases during the transition from exponential to stationary phase. We focused our study on RsaE, an sRNA that is highly conserved in the bacillales order and is deleterious when over-expressed. We show that RsaE interacts *in vitro* with the 5′ region of *opp3A* mRNA, encoding an ABC transporter component, to prevent formation of the ribosomal initiation complex. A previous report showed that RsaE targets *opp3B* which is co-transcribed with *opp3A*. Thus, our results identify an unusual case of riboregulation where the same sRNA controls an operon mRNA by targeting two of its cistrons. A combination of biocomputational and transcriptional analyses revealed a remarkably coordinated RsaE-dependent downregulation of numerous metabolic enzymes involved in the citrate cycle and the folate-dependent one-carbon metabolism. As we observed that RsaE accumulates transiently in late exponential growth, we propose that RsaE functions to ensure a coordinate downregulation of the central metabolism when carbon sources become scarce.

INTRODUCTION

Regulatory RNAs in prokaryotes are frequently referred to as small RNA (sRNAs). The sRNAs are typically between 50 and 300 nts in length. Most sRNAs with identified regulatory functions pair with mRNAs to modulate their translation, stability or transcription (1). The sRNAs encoded opposite to mRNA genes are referred to as '*cis*-encoded sRNAs', since they often regulate the associated mRNA expressed from the complementary coding strand by extended perfect base-pairing (2). In contrast, sRNAs expressed from intergenic regions (IGRs) often act on target mRNAs transcribed from genetically unlinked regions. Base-pairing between these *trans*-encoded sRNAs and mRNA targets is usually imperfect and limited to short sequences. In many bacteria, including the Gram-negative bacterium *Escherichia coli*, the activity of *trans*-encoded sRNAs requires the RNA-chaperone Hfq (3). However, the Gram-positive bacterium *Staphylococcus aureus* might not require Hfq for sRNA activity since an *hfq* deletion mutant of this

*To whom correspondence should be addressed. Tel: +33 1 69 15 70 16; Fax: +33 1 69 15 66 78; Email: philippe@bouloc.eu

© The Author(s) 2010. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

organism has no discernable phenotype (4). A third class of sRNAs can exert a regulatory function by directly interacting with proteins (5–7). Lastly, some mRNAs possess structured 5′ untranslated regions (UTR) that, under specific growth conditions, can adopt alternative conformations to affect gene expression of the downstream genes in *cis*. Binding of molecules such as metabolites or tRNAs to these so-called riboswitches induces RNA structure changes of the transcribed UTR, causing premature transcriptional termination or translational inhibition (8). Some riboswitches are known to produce sRNAs (9), which might independently function on *trans*-encoded target mRNAs by base-pairing mechanisms (10).

Several methods have been used to identify sRNAs, including computational predictions, direct labeling, DNA microarrays and shotgun sequencing (11,12). *Escherichia coli* and *Salmonella* each encode an estimated 100 sRNAs (11). These respond to diverse environmental factors such as glucose availability, iron limitation, envelope stress, or transition to anaerobiosis, to facilitate adjustments, respectively, in carbon metabolism (13,14), of levels of iron-requiring enzymes (15–17), amounts of metabolic transporters (18,19) or quantities of enzymes required for anaerobic growth (20,21). Other regulatory functions of sRNAs include quorum sensing and virulence (22–26).

Staphylococcus aureus is a human commensal of skin and nares and an opportunistic pathogen responsible for a wide variety of human infections including superficial skin and wound infections to deep abscesses (endocarditis and meningitis), septicemia or toxin-associated syndromes (27). It is a leading cause of nosocomial and community acquired diseases and many strains acquire antibiotic resistance (28). *Staphylococcus aureus* virulence gene expression is regulated by the coordinated action of numerous regulators. One of the best characterized regulators is an sRNA called RNAIII (514 nt), which modulates virulence gene expression (29). Its 3′ domains can mediate translational repression of *S. aureus* virulence genes directly by interacting with *spa* mRNA, which encodes the surface-anchored Protein A, or indirectly, by interacting with *rot* mRNA, which encodes Rot, a transcriptional regulator of several exoproteins (30–32).

Staphylococcus aureus has recently emerged as a model organism for the study of bacterial sRNAs (33). A pioneering study based on a computational analysis of IGRs led to the identification of 12 sRNAs that were experimentally confirmed, among which seven were expressed from pathogenicity islands (34). Thirty-four new sRNAs, including some that likely originate from putative riboswitches were recently identified (35–37). Finally, one deep sequencing and two global transcriptome studies on *S. aureus*, performed using different growth conditions suggested the existence of numerous non-coding small stable RNAs; however, most of them were not confirmed by alternative methods (37–39).

As sRNAs are potent regulators that can potentially affect viability, adaptability to different growth conditions, or virulence of this deadly pathogen, we undertook a search for new sRNAs in *S. aureus*. The characterization of one of these new sRNAs indicates that it coordinately downregulates multiple enzymes of two major metabolic pathways, likely to help *S. aureus* prepare for the entry into stationary phase.

MATERIALS AND METHODS

Bacterial strains, plasmids and growth conditions

The RNomics approach was performed on strain N315 (27). Transcriptome analysis was performed on strain RN6390 (40), a laboratory derivative of the NCTC8325 clinical strain (41) that was shown to be virulent in different animal models (42,43). Note that for simplicity, we only used the N315 gene name nomenclature. Engineered plasmids were constructed in *E. coli* TG1, and transferred to RN4220 (*rsbU*, *agr*), which is transformable with exogenous DNA (44), and subsequently to RN6390.

Staphylococcus aureus strains were routinely grown in BHI medium with aeration at 37°C. *Escherichia coli* TG1 used for cloning experiments was grown in LB broth at 37°C. Chloramphenicol was added to media as needed at the following concentrations: 5 μg/ml for *S. aureus* and 10 μg/ml for *E. coli*. To determine the effect of sRNA induction, staphylococcal cultures were prepared as follows: 10-fold serial dilutions of exponential cultures were spotted on BHI plates containing or not 1 μM anhydrotetracycline (aTc) and incubated at 37°C for 24 h.

pAT12 contains the *repA*, *repC*, *tetR* genes and the *xyl*/*tetO* promoter (45). The pAT12 derivative containing *rsaA*, *rsaH*, *rsaE* or expressing RNA sequences antisense of RsaA, RsaH and RsaE (named pAT12-RsaA, pAT12-RsaH, pAT12-RsaE, pAT12-ASRsaA, pAT12-ASRsaH and pAT12-ASRsaE, respectively) were constructed as follows: DNA sequences of interest were amplified from N315 genomic DNA by PCR, using primers indicated in Supplementary Data S1. The resulting products were digested by EcoRI and NotI, ligated to EcoRI-NotI-treated pAT12. Expression of the cloned sequences on constructed plasmids was induced by adding anhydrotetracycline (aTc) to the media, and confirmed by northern blot analysis.

DNA preparation

Lysis of *S. aureus* cell suspensions was achieved by treatment with 100 μg/ml of lysostaphin (Sigma-Aldrich, St. Louis, MO, USA) for 1 h at 37°C, followed by standard methods for DNA preparation. DNA transformation of *E. coli* and for *S. aureus* was performed as described (46,47).

RNA manipulation

For northern blot analysis and 454 pyrosequencing, overnight cultures of staphylococcal strains grown in BHI medium were diluted 1000-fold in the same medium and grown at 37°C; samples were withdrawn at mid-or post-exponential or stationary phase cultures. To study the effect of stress conditions, cultures at $OD_{600} = 0.2$ were either shifted to 45°C, or supplemented with NaCl 1.5 M or H_2O_2 10 mM for 2 h. Total RNAs were extracted

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

as described (48) or using NucleoSpin RNAII (Macherey Nagel, Düren, Germany).

Northern blots were performed as described (35,46,49). Samples were separated by either PAGE or agarose gels and probed with γ^{32}-P 5′ labeled oligonucleotides (described in Supplementary Data S1).

Rapid amplification of cDNA ends (RACE) mapping was performed as described (50,51).

Synthetic RNAs were generated using the T7 MEGAshortscript kit (Ambion, Foster City, CA, USA) according to manufacturer's instructions.

RNA gel mobility shift assays were performed as described (52). [α^{32}-P]UTP was added to the transcription mix to produce radioactive transcripts. The transcripts were denatured in 50 mM HEPES (pH 6.9), 50 mM NaCl, 5 mM KCl and 1 mM $MgCl_2$ for 3 min at 85°C, followed by refolding for 10 min at 30°C. Labeled RsaE (0.04 pmol) was incubated with increasing amounts of unlabeled synthetic RNAs or yeast tRNAs in 10 mM Tris–HCl (pH 7.5), 6 mM NaCl, 10 mM EDTA and 5 mM dithiothreitol for 10 min at 30°C. Samples were supplemented with 10% glycerol (final concentration) and loaded on a native 5% polyacrylamide gel containing 5% glycerol.

Toeprint experiments were performed as described (53,54) with the following modifications. Annealing mixtures contained 0.2 pmol unlabeled *opp3A* mRNA and 1 pmol labeled SA0849 mRNA (primer 404) in a buffer containing 10 mM Tris–acetate (pH 7.5), 60 mM NH_4Cl and 1 mM DTT. The effect of RsaE was tested by adding various quantity of RsaE (25–300 pmol) to the annealing mixtures prior to the addition of purified *E. coli* 70S ribosomes. Ribosomes were activated for 15 min at 37°C and diluted in the reaction buffer in the presence of 1 mM $MgCl_2$. For each sample, 4 pmol of 70S were added, followed by 5 min incubation and adjustment of $MgCl_2$ to 10 mM. Ten picomoles of uncharged $tRNA^{fMet}$ were then added and samples were incubated 15 min. cDNA was synthesized using two units of AMV RT (New England Biolabs, Ipswich, MA, USA) for 15 min at 37°C. Reactions were stopped by addition of 10 µl loading buffer II (Ambion, Foster City, CA, USA). The cDNAs were loaded and separated by Urea 8% PAGE. The toeprint position on *opp3A* mRNA was determined by DNA sequencing.

cDNA library from sRNAs, 454 pyrosequencing and analysis of data

RNA preparations (cf. above) were treated twice with DNase (DNA-free, Ambion, Foster City, CA, USA) according to manufacturer's instructions. For cDNA library construction, *S. aureus* RNAs from six different growth conditions (see 'Results' section) were mixed in a equimolar ratio. In addition, we used a new treatment protocol to enrich for primary bacterial transcripts (most mRNAs and sRNAs) (73). For this purpose, the pooled RNA was treated with Terminator 5′-phosphate dependent exonuclease (Epicentre, Madison, WI, USA). From this RNA, a cDNA library was prepared and analyzed on a Roche GS20 sequencer (F. Hoffmann-La Roche, Basel, Switzerland) as described (55). The cDNA library was constructed by Vertis Biotechnology AG (http://www.vertis-biotech.com/). In brief, the Terminator exonuclease-treated RNA was first treated with tobacco acid pyrophosphatase (TAP). The RNA samples were subsequently poly(A)-tailed using poly(A) polymerase followed by ligation of an RNA adapter to the 5′-monophosphate of the sRNAs. First-strand cDNA synthesis was then performed using an oligo(dT)-adapter primer and MMLV H- reverse transcriptase. The resulting cDNAs were then PCR-amplified and the specific barcode sequence CAGC was attached to the 5′ ends of the cDNAs.

DNA sequences obtained from a multiplex 454 run on a Roche GS20 sequencer that contained the specific barcode tag CAGC were extracted. The 5′ end linker and polyA-tails were clipped, yielding 38 052 insert sequences (reads). 543 reads with lengths <14 nt were excluded from the study. The remaining 37 507 reads (≥14 nt) were blasted with WU-blast (http://blast.wustl.edu/, chosen parameters were $E = 0.0001$, $m = 1$, $n = -3$, $Q = 3$, $R = 3$) against the *S. aureus* N315 genome sequence (NC_002745.fna), and 36 021 reads that could be mapped were kept for further analysis. Those with blast hits that corresponded to rRNA, tRNA, tmRNA, RNaseP, SRP RNA and known mRNA were removed. 202 reads (Supplementary Data S2) corresponded to sequences putatively issued from IGRs, and were assembled with cd-hit-est (similarity threshold $c = 0.92$) into 160 clusters (56). Further blast analysis of these 160 remaining clusters revealed additional tRNA or mRNA sequences that were excluded. The assembled sequences were manually analyzed to identify the sRNAs presented in this study. For example, at this last step, reads issued from sequences adjacent to open reading frames were considered as likely mRNA UTRs and discarded.

Transcriptome analysis

Overnight cultures grown in BHI supplemented with chloramphenicol were diluted 1000-fold in the same medium at 37°C. At $OD_{600} = 0.6$, aTc (1 µM) was added to the medium and 5 min later, the cultures were sampled. Total RNAs were extracted and treated twice with DNase I. RNA quality was controlled with a 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA). Two conditions were tested, RN6390 containing pAT12-RsaE and pAT12-AS-RsaE. Trancriptome experiments were performed using *S. aureus* Affymetrix DNA chips (Santa Clara, CA, USA) by Cogenics (Newton, MA, USA) according to the standard manufacturer's instructions.

Quantitative real-time PCR (qRT–PCR)

Overnight cultures grown in BHI supplemented with chloramphenicol were diluted 1000-fold in the same medium at 37°C. At $OD_{600} = 0.3$, aTc (1 µM) was added to the medium and cultures were sampled at 0, 5, 10, 20 and 30 min. RNAs were prepared as described (48). Genes for normalization were selected, qRT–PCR experiments were performed and data were analyzed as described (57). The geometric mean of 3 genes (*recA*, *gyrA* and *ftsZ*) chosen among 13 was used to normalize the samples.

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

RESULTS

Screening for new *S. aureus* sRNAs

As compared to enteric bacteria, comparatively few sRNAs have been experimentally confirmed to date in *S. aureus* (34–37,58). Our goal was to identify new sRNAs without any *a priori* restrictions in the selection; we therefore used an RNomics approach based on pyrosequencing of cDNA libraries (12). The study was performed on N315, a methicillin-resistant *S. aureus* strain that was isolated in a case of a hospital-acquired infection in 1982 (27). As sRNAs are often expressed in response to specific signals, total RNA was extracted from six different conditions. N315 was inoculated in rich medium (BHI) under aeration at 37°C and samples were withdrawn in (i) exponential phase, (ii) post-exponential phase, (iii) stationary phase and after 2 h of exposure to (iv) osmotic, (v) thermal or (vi) oxidative stress (cf. experimental procedures). For cDNA library preparation the RNAs extracted from these six growth conditions were pooled in an equimolar ratio, treated with Terminator 5′-phosphate dependent exonuclease to deplete processed sRNAs, thus enriching primary transcripts, and converted to cDNA libraries to be analyzed by 454 pyrosequencing as described (55,59,73).

A total of 36 021 cDNA sequences (referred to as reads) > 14 nt matched to the N315 *S. aureus* reference DNA sequence. The vast majority of these reads corresponded to rRNAs (33 094, mainly 5S) or tRNAs (2548), and were not analyzed further. Likewise, 173 reads matching small mRNAs or probable mRNA degradation products were discarded. A total of 126 reads that did not correspond to the above-described RNAs were retained for further analysis (Supplementary Data S2). Of these, 50 were likely issued from mRNA UTR regions, rather than sRNAs *per se*, and 73 were putatively derived from IGRs. In addition, seven sequences mapped to the complementary strand of coding sequences (Supplementary Data S2), suggesting the expression of *cis*-encoded RNAs. The 71 sequences of interest were clustered in 32 IGRs. None of the sequences matched RNAIII, as expected since the N315 strain expresses little if any RNAIII. 4.5S RNA, tmRNA, RNAse P RNA and 6S RNA were found in *S. aureus via* a computational screening (34); we identified all of these (corresponding to 54 reads, Table 1). In addition, six of the seven previously reported sRNAs of *S. aureus* pathogenicity islands, named SprA through SprG (34), were recovered by our RNomics approach (Table 1); the exception was SprE, which has turned out to be an rRNA fragment and not an sRNA (S.C. and B.F., unpublished results). In addition to already known sRNAs, our global RNomics search suggested the existence of new *trans*-encoded sRNAs [of which two were suggested, but not confirmed, by previous studies (34,38,39)]. We confirmed 16 of these putative sRNAs by northern blot (two other tested regions could not be confirmed). We named them RsaOH to RsaOW [for RNA *S. aureus* Orsay (35)] and the corresponding genes *rsaOH* to *rsaOW* (Table 1). While the present work was being completed, the sRNAs RsaOJ, RsaOK and RsaON were independently reported; we will refer to them using the recently published names, RsaA, RsaH and RsaE, respectively (36).

Expression of *S. aureus* sRNA genes encoded in IGRs

Expression of the 15 identified RNAs was followed as a function of the growth phase. An overnight culture of N315 was diluted into a fresh rich medium and grown aerobically at 37°C; samples were withdrawn at different stages of growth and sRNA amounts were evaluated by northern blot analysis (Figure 1). We also examined the expression of six sRNAs in the six growth and stress conditions used for the RNomics approach (Supplementary Data S4). The studied sRNAs presented below are ordered according to their position on the N315 genomic map. Conservation of the identified sRNAs in other species is indicated in Table 1.

RsaOH is about 100-nt long. Its gene, *rsaOH*, is located between SA0065 and *kdpE* genes encoding a putative C4-dicarboxylate transporter and a putative transcriptional regulator belonging to the OmpR family of two-component response regulators, respectively. The *rsaOH* detected sequence is adjacent to the 3′-end of *kdpE* (Supplementary Data S3). One may speculate that RsaOH interacts with the 3′-end of the *kdpD-kdpE* mRNA and therefore could regulate its stability. RsaOH increases with the growth phase but strongly decreases in the stationary phase (Figure 1).

RsaOI is about 240-nt long; it is the longest sRNA that we identified, and contains no expected ORF. Its gene, *rsaOI*, is located between *proP*, encoding a putative transporter and partially (38 nt) overlaps in antisense orientation with the 5′ end of SA0532 encoding a *Staphylococcus*-specific hypothetical protein (Supplementary Data S3). It is therefore possible that RsaOI affects transcription or translation of SA0532. In contrast to RsaOH, RsaOI expression drastically increases in stationary phase (Figure 1).

RsaA (RsaOJ) is 142-nt long (36). Its gene, *rsaA*, is located between SA0543 encoding a conserved putative membrane protein and SA0544 encoding a putative heme peroxidase. Like RsaOI, RsaA, is detected in all growth phases and drastically increases in stationary phase (Figure 1). RsaA did not accumulate upon thermal, salt or oxidative stress (Supplementary Data S4).

RsaH (RsaOK) is 128-nt long (36). Its gene is located between SA0724, encoding a hypothetical protein similar to a cell-division inhibitor, and SA0725, encoding a protein of unknown function. RsaH could encode a putative 23 amino-acid polypeptide (MCKTLHDTNEGKNLTPFSSGPVR). In the corresponding IGR of *S. aureus* strain RF122 the SAB0725 ORF of a putative 52 amino-acid protein with no significant similarity to proteins from other organisms has been annotated. However, in the majority of *S. aureus* strains, including N315, the coding sequence is interrupted by a stop codon leading to the putative 23 amino-acid peptide. RsaH was previously reported to be induced in stationary phase (36). We found that in N315, RsaH is barely detectable in early exponential phase, but accumulates during

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

Table 1. sRNAs identified in this RNomics study

RNA name[a]	IGR	Flanking genes	Strand Orientation	Nb of reads[b]	Alternative names	Conservation[c]
RsaOH	72446/73508	SA0065/*kdpE*	< > <	1	IGR59 (34)	*S. aureus*, *S. epidermidis*, *S. haemolyticus*
4.5S RNA (34)	501358/502001	SA0434/SA0435	> > >	23	SRP RNA	Conserved
RsaOI	626797/627130	*proP*/SA0532	> < >	4	Sau-6477 (37)	*S. aureus*
RsaA (36)	636758/637419	SA0543/SA0544	< > <	2	RsaOJ (this work), Sau-64 (37)	*S. aureus*, *S. epidermidis*
RsaC (36)	679620/680738	SA0586/SA0587	< < <	1		*S. aureus*
RsaH (36)	829363/830064	SA0724/SA0725	< > >	2	RsaOK (this work)	Staphylococcal strains
tmRNA (34)	843706/844543	smpB/SA0737	> > >	27	SsrA	Conserved
RsaOL	856992/857482	*cspC*/SA0748	> < <	1	Sau-07 (37)	*S. aureus*, *S. epidermidis*
RsaOM	875349/875597	SAS023/SA0769	> > >	1	Predicted SAM-I riboswitch	Conserved
RsaE (36)	975283/975742	SA0859/SA0860	> > <	3	RsaON (this work) Sau-20 (37)	Bacillales order
Sau-02 (37)	1006365/1007020	SAS028/SAS029	> < >	1		*S. aureus*, *S. epidermidis*, phages, plasmids
RsaOO	1012811/1013050	SA0891/SA0892	> < <	1		*S. aureus*
RsaOP	1089149/1089566	SA0961/SA0962	> > >	1		Staphylococcal strains
RsaOQ	1372074/1372564	SA1198/SA1199	> > >	1	Predicted T-box	Conserved
RNaseP RNA (34)	1483578–1484012	SA1277/SA1278	< < <	2		*S. aureus*, phages
6S RNA (34)	1660651–1660795	SAS050/*aspS*	< < <	2	Srr80 (38,39)	Conserved
SprA (34)	1856223/1856978	SA1623/tnp IS232	< > <	1		Staphylococcal strains
SprB (34)	1866662–1867133	SA1633/SA1634	> < <	3		*S. aureus*
SprC (34)	1871168–1872530	*luckE*/SA1639	< < >	1		*S. aureus*, *S. epidermidis*, *S. haemolyticus*
SprD (34)	2006879–2007560	SA1754/SA1755	< < >	3		*S. aureus*
RsaOR	2008572–2009085	SA1756–57/*sak*	< < <	1	Srr6 (38,39, SprX)	*S. aureus*, phages
SprF3	2010790–2011000	SA1956/SAS069	> > >	1	RsaOS (this work)	Staphylococcal strains
Sau-30 (37)	2294189/2294725	cbiO/rplQ	< < <	2		Staphylococcal strains
RsaOG (35)	2367825/2368208	SA2104/SA2105	< < >	2	RsaI (36)	Staphylococcal strains
SprA2 (34)	2490680–2491385	SA2217/SA2218	< < >	1	WAN014FZW (38,39)	Staphylococcal strains
RsaOT	2544336–2544594	SA2267/SA2268	> < <	7	Srr43 (38,39)	*S. aureus*
RsaOU	2611619/2611918	*ptsG*/SA2327	< < <	2	Rat sequence	
RsaOV	2632123/2632361	SA2343/*copA*	> > >	1	Sau-40 (37)	*S. aureus*
RsaOW	Eight possible positions	Tnp for IS1181	< > –	4		
RsaOX		SA0062	*Cis*-encoded	1		

[a]Gene names corresponding to sRNAs confirmed by northern blot previously or in this study (cf. Figure 2 and Supplementary Data S4).
[b]Number of reads obtained by 454 sequencing matching the corresponding indicated IGR in given orientations (cf. Supplementary Data S2).
[c]Conservation determined by Blast analysis (default parameters) using the identified or predicted sRNA gene sequence. Conserved indicates that the sequence can be found in phylogenetically distantly related bacteria.

pre-stationary phase (Figure 1), and then strongly decreases in stationary phase.

RsaOL is estimated to be about 100-nt long, according to our northern blot probing. It is transcribed from a gene located in the IGR between *cspC* (encoding a cold shock protein from the CspA family) and SA0748 (encoding a poorly conserved putative protein of unknown function). RsaOL is expressed in all growth phases tested. Its quantity increases with the growth phase, except in stationary phase, where levels are slightly decreased (Figure 1).

RsaE (RsaON) is 93-nt long (36). The *rsaE* gene is located between SA0859, encoding a predicted dithiol-disulfide isomerase involved in polyketide biosynthesis, and SA0860 encoding an oligoendopeptidase. The RsaE sequence is present in the bacillales order and the synteny is conserved within the Firmicute phylum with the noticeable exception of the Listeriaceae family. The *rsaE* gene is identical in all *S. aureus* strains sequenced to date. RsaE is thus likely under a strong selective pressure for the conservation of its primary sequence: one possibility is that RsaE interacts with a specific protein requiring a defined nucleotide sequence for interaction and subsequent function. However, experiments based on the incubation of synthetic RsaE with protein extracts did not identify an RsaE-specific binding protein (data not shown). RsaE was also reported as induced in stationary phase (36). However, a detailed analysis reveals that in N315, although RsaE is detectable during all growth phases, it accumulates specifically during pre-stationary phase (OD_{600} = 6) and almost disappears in stationary phase (Figure 1). Its expression profile is similar to that of RsaH. In several strains, RNAIII is also induced in late exponential phase. Transition to stationary phase and possibly virulence activate the expression of several sRNAs, which make them candidates for being key regulators of transition adaptation processes. Furthermore, RsaE is reduced under heat-shock (Supplementary Figure S4).

RsaOO is about 180-nt long. Its gene, *rsaOO*, is located between SA0891 and SA0892, both of which encode membrane proteins. A second product, about 140-nt

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

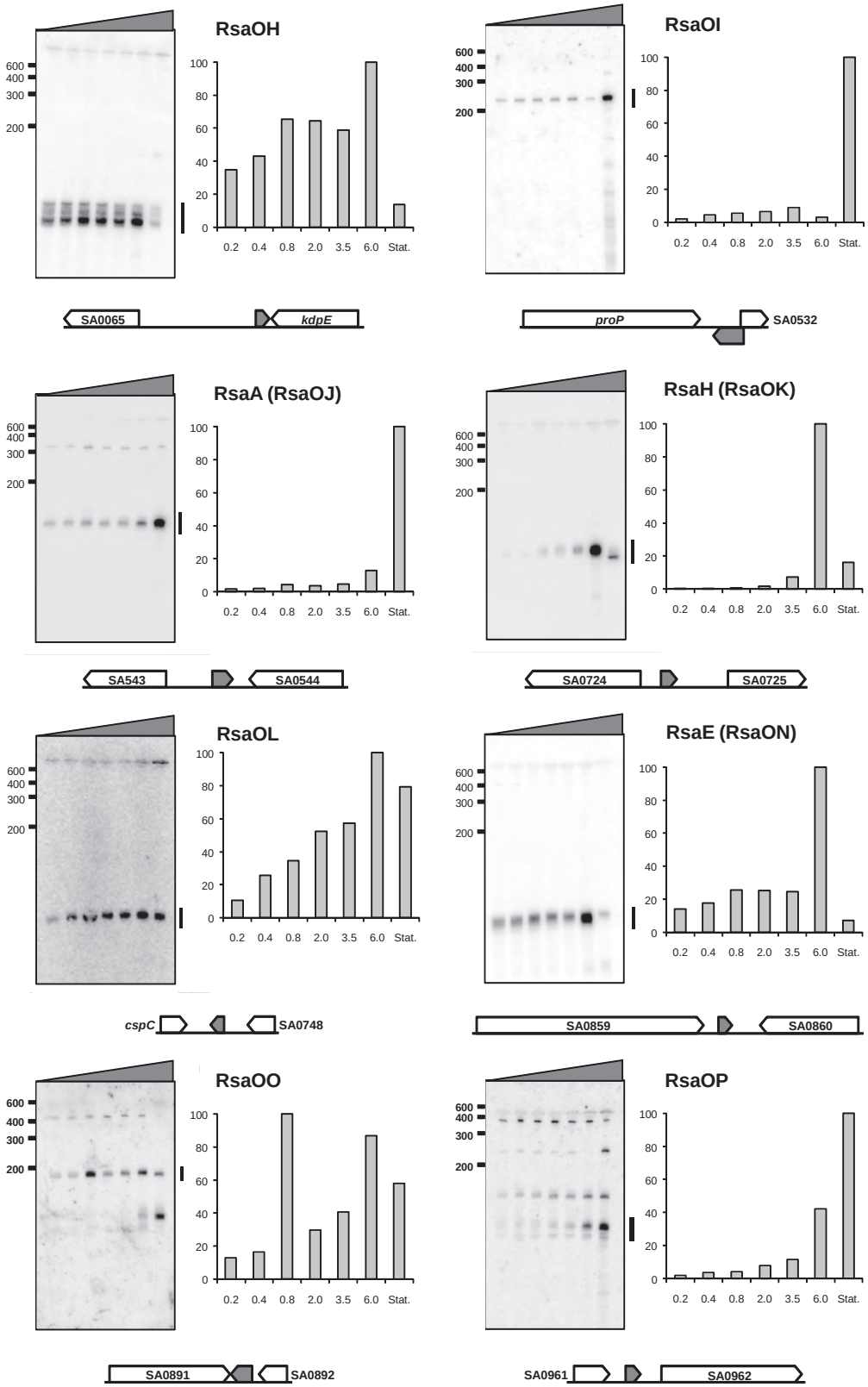


Figure 1. Identification and characterization of new N315 *S. aureus* sRNAs derived from the RNomics approach. Northern blots showing the expression of various sRNA candidates in nutrient-rich medium are presented. Total RNAs were prepared from cultures harvested at OD_{600}: 0.2, 0.4, 0.8, 2, 3.5, 6 and 10. Black wedges indicate the increasing OD_{600} values from left to right. All blots presented were performed using the same RNA extracts. Marker sizes are shown to the left of each blot. RNAs were detected using γ^{32}P-labeled oligonucleotides designed to pair with the predicted RNAs (cf. Supplementary Data S1). 5S rRNA was used as a loading control. Histograms indicate the relative expression of each sRNA in different growth phases (the intensity of each band was divided by the intensity of the 5S band from the same RNA extraction; and the maximum value for each histogram was set to 100). A bar on the right side of the blot indicates the band(s) selected for quantification. Flanking or overlapping genes (open arrows) of each sRNA (grey arrow) is presented to indicate genetic localization. Arrows indicate the direction of transcription and are proportional to known or expected transcript sizes.

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

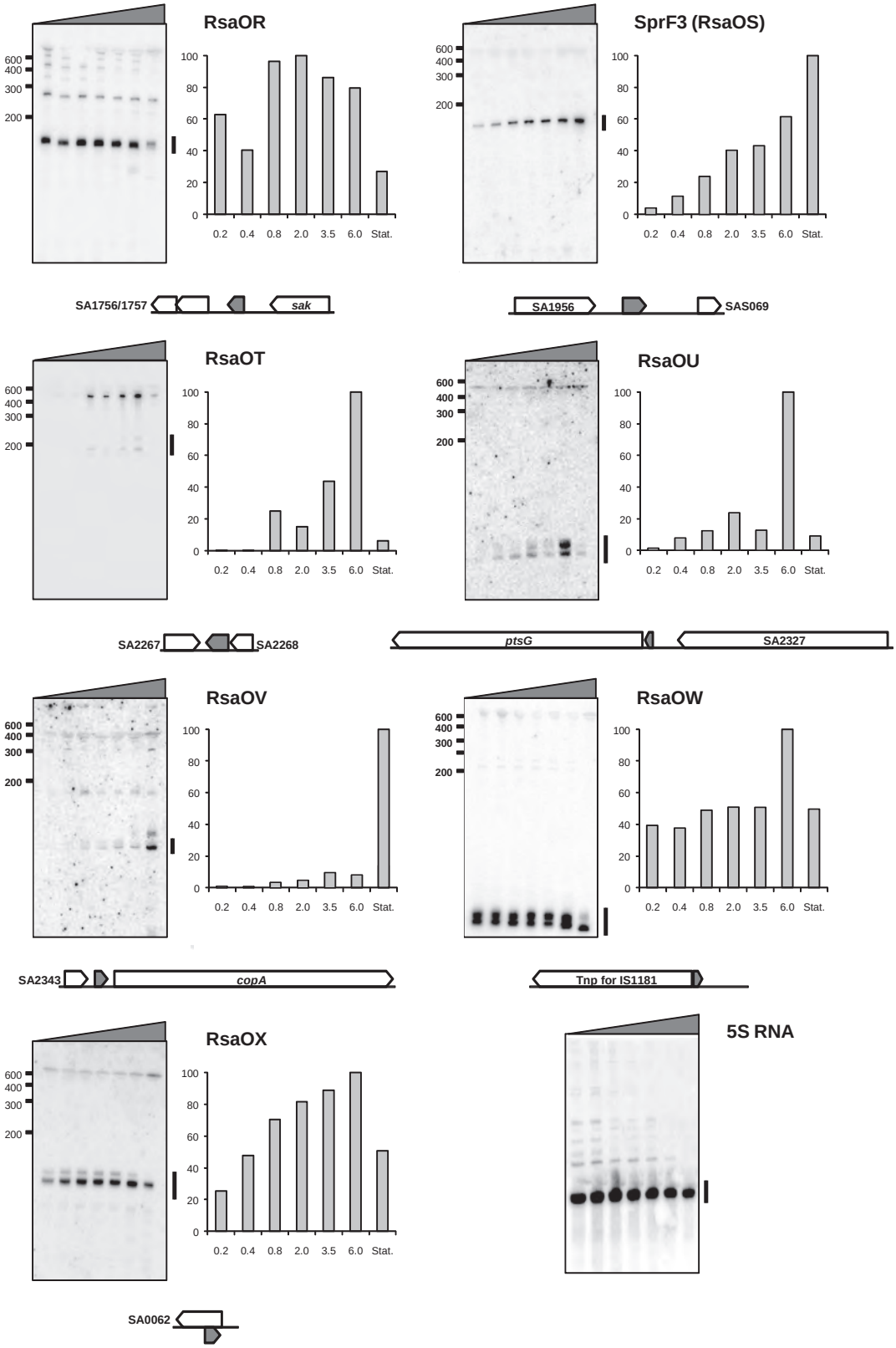


Figure 1. Continued.

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

long, accumulates in stationary phase, possibly issued from the processing of RsaOO (Figure 1).

The *rsaOP* gene is located between SA0961 and SA0962, which encode a protein of unknown function and an integral membrane protein (FtsW) involved in cell division, respectively. Several products were detected when probing for RsaOP; the major one is about 100-nt long and accumulates in stationary phase (Figure 1).

RsaOR (alias SprX) is about 160-nt long with a possible longer form. Its gene is located between SA1757, encoding a putative truncated amidase, and *sak*, encoding staphylokinase. The sequence comprising this region, including adjacent gene SA1757 is also found in several phages (tp310-3, 52A, phi12, tp310-2, 80apha, phiNM4, etc.). RsaOR is thus likely acquired via horizontal transfer and expressed by phages. RsaOR is expressed throughout growth, but amounts are reduced in stationary phase (Figure 1). It also accumulates upon salt stress (Supplementary Data S4). N315 RsaOR may encode a putative polypeptide of 40 amino acids (MHQLFTSLFT QACHWVFFLM IESIVFILLP RSIYDFSIPV). However, the initiation codon is absent in most *S. aureus* sequenced genomes and there is no obvious ribosome binding site sequence. Therefore, if expressed, this peptide would be expressed only in few *S. aureus* strains. We therefore favor the hypothesis that RsaOR is a regulatory sRNA.

The intergenic SA1760/SAS059 region contains *sprF* and on its complementary strand *sprG*. In N315, the corresponding RNAs SprF and SprG were proposed to have paralogs expressed from two other separated regions located on pathogenicity islands (34). A read obtained by 454 (Supplementary Data S3) matched with the SA1956/SAS069 IGR containing the putative *sprF3* (*rsaOS*) gene. Northern blot experiments show that the expression of *sprF3* increases gradually with the growth phase (Figure 1). 5′ RACE mapping experiments (Supplementary Data S3) indicate that the sequence of the first 47 nt of *sprF3* does not match *sprF* and *sprF2*, suggesting that targets of SprF3 could differ from those of SprF and SprF2.

RsaOT is about 180-nt long. It is poorly expressed and we could not identify its extremities. *rsaOT* expression is at a maximum during pre-stationary phase (Figure 1) and is also stimulated by the presence of H_2O_2 in the medium (Supplementary Data S4). As one of the RsaOT cDNA sequences obtained via our RNomic study is only at 6 nt downstream from the SA2268 stop codon, it is possible that the sRNA detected by northern blot is issued from 3′ end processing of SA2268.

RsaOV is about 100-nt long. Its gene, *rsaOV*, is located between SA2343 and *copA*, which encode a putative poorly conserved protein and a copper translocating ATPase, respectively. RsaOV accumulates in stationary phase (Figure 1).

The *rsaOW* gene is present in eight copies (*rsaOW1* to *rsa0W8*) on the N315 genome; it is located immediately upstream of the ORF encoding the IS1181 transposase (Supplementary Data S3). RsaOW is therefore perfectly complementary to the transposase 5′ UTR region (including its ribosome binding site). As RsaOW is detected in all growth conditions (Figure 1), one may hypothesize that its role is to control, likely lowering, the expression of the IS1181 transposase. RsaOW would therefore affect the *S. aureus* genome evolution.

Our results confirm the existence of several *trans*-encoded *S. aureus* sRNAs and show that their expression is, for the most part, dependent on growth conditions.

Expression of *cis*-encoded *S. aureus* sRNA genes

Our RNomics approach predicted seven candidates of *cis*-encoded sRNAs; we were able to detect one of them, RsaOX, by northern blot analysis (Figure 1). The others gave either no hybridization signal or signals corresponding to high-molecular weight RNAs (data not shown). RsaOX is about 120-nt long and is transcribed from the complementary strand of SA0062; it was expressed in all of the tested growth phases. As SA0062 encodes a putative transposase, the existence of RsaOX is reminiscent of the system described for the *E. coli* IS10 transposase, whose mRNA (called RNAin) is controlled by RNAout, its *cis*-encoded sRNA (60,61). SA0062 mRNA was recently proposed to be the target of $MazF_{SA}$, a sequence-specific endoribonuclease (62). One can postulate that in addition to $MazF_{SA}$, RsaOX could prevent the synthesis of the likely deleterious SA0062-encoded transposase and therefore might limit transpositions.

Identification and expression of sRNA for putative *S. aureus* riboswitches

The RNomics analysis allowed us to identify transcripts that likely result from premature transcriptional termination associated with riboswitch activities. RsaOM corresponds to the sequence upstream of SA0769 encoding a D-methionine transport system ATP-binding protein. Probing for this region revealed several small transcripts more expressed in the high salt stress condition (Supplementary Data S4). The SAS023-SA0769 IGR is predicted to contain a SAM riboswitch [also known as S-box leader or SAM-I riboswitch; (63)], which is found upstream of genes involved in methionine or cysteine biosynthesis (64). The effector, S-adenosyl-methionine, specifically binds to the nascent S-box RNA leading to a premature transcription arrest (65,66), which would explain the observed short transcripts (Supplementary Data S4). Surprisingly, the N315 predicted SAM riboswitch may encode a putative protein of 60 amino acids conserved in *S. aureus* genomes. However, this ORF does not exist in other species, suggesting that if expressed in *S. aureus*, its activity might not be related to the SAM riboswitch control.

RsaOQ corresponds to the upstream sequence of SA1199, which encodes component I of the anthranilate synthase involved in phenylalanine, tyrosine and tryptophan biosynthesis. Probing for this sequence revealed several bands with a different profile in stationary phase as compared to the exponential phase or stress conditions (Supplementary Figure S4). The SA1198-SA1199 IGR is predicted to contain two tandem T-boxes (63,67). These leader elements sense uncharged tRNA to regulate the expression of aminoacyl–tRNA synthetase genes and

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

amino-acid biosynthetic genes such as anthranilate synthase (68).

RsaOU is a tiny sRNA (estimated length is <70 nt) originating immediately upstream of *ptsG*, which encodes a component of the glucose-specific phosphotransferase system. RsaOU is highly expressed in pre-stationary phase (Figure 1). It is likely that *rsaOU* acts as a *cis*-element on *ptsG* expression. Indeed, the *ptsG* operon of *Staphylococcus carnosus* is regulated by a termination/antitermination mechanism associated with a sequence called RAT for ribonucleic antiterminator (69,70). HPr-dependent phosphorylation of GlcT facilitates antitermination (69,70). A similar regulation is likely to exist in *S. aureus*; in addition to *ptsG*, *glcA*, which encodes another enzyme from the PTS system, is preceded by a RAT sequence that generates an sRNA (36).

RsaE downregulates the one carbon pool by folate and central metabolic pathways

The *rsaA* (*rsaOJ*), *rsaE* (*rsaON*) and *rsaH* (*rsaOK*) genes (corresponding to RNAs highly expressed during specific growth phases) were cloned under the regulatory control of the $P_{xyl/tetO}$ promoter of pAT12, giving rise to pAT12-RsaA, pAT12-RsaE and pAT12-RsaH, respectively. These plasmids express the Tet repressor that binds to the *tetO* operator; sRNA expression was induced by addition of anhydrotetracycline (aTc) to the media. We also constructed derivatives, pAT12-ASRsaA, pAT12-ASRsaH and pAT12-ASRsaE expressing antisense RNAs of RsaA, RsaH and RsaE, respectively. RN6390 containing pAT12-RsaE, but not the other above described plasmids, was sensitive to aTc (100 nM) in a solid rich media at 30, 37 or 42°C (Figure 2 and data not shown). Accumulation of RsaE may cause toxicity by affecting essential functions; we therefore focused our interest on the functional characterization of RsaE.

Many sRNAs act *via* base-pairing to target mRNAs to affect their stability and/or translation. We hypothesized that RNA targets of RsaE would vary upon RsaE accumulation, and could thus be identified by transcriptome analysis. We therefore performed an mRNA analysis of strains expressing a plasmid-encoded RsaE (pAT12-RsaE) or RsaE-antisense (pAT12-ASRsaE), sampling RNA 5 min after induction of RsaE or its antisense sequence; this approach was designed to identify those mRNAs with which RsaE directly interacts, and is similar to the pulse-expression approaches that facilitated sRNA target discovery in *E. coli* and *Salmonella* (16,19). Preliminary northern blot analysis showed that 5 min of induction was sufficient for strong accumulation of either RsaE or its antisense RNA (data not shown). Induction was performed at OD_{600} = 0.6, when the expression of the chromosomally encoded RsaE is low (Figure 1). RsaE accumulation led to the downregulation of 25 genes; surprisingly, it also upregulated 39 genes at least 2-fold (Table 2 and Supplementary Data S5). To consolidate our results, the expression of 16-selected genes was studied *via* a kinetic of RsaE induction at time 0, 5, 10, 20 and 30 min by qRT–PCR (Table 2 and Supplementary Data S6). With the exception of one gene (*opp3A*, cf. Table 2 legend), the qRT–PCR results fully confirmed the transcriptome data. The mRNAs modulated either directly or indirectly by RsaE mainly encode membrane proteins or metabolic enzymes.

Strikingly, mRNAs that are downregulated by RsaE can be clustered around two essential pathways (Figure 3): (i) SA1366-SA1367 (*gcvP-gcvT*) and SA1553 (*fhs*) mRNAs are associated with production of methenyl tetrahydrofolate (THF) required for the

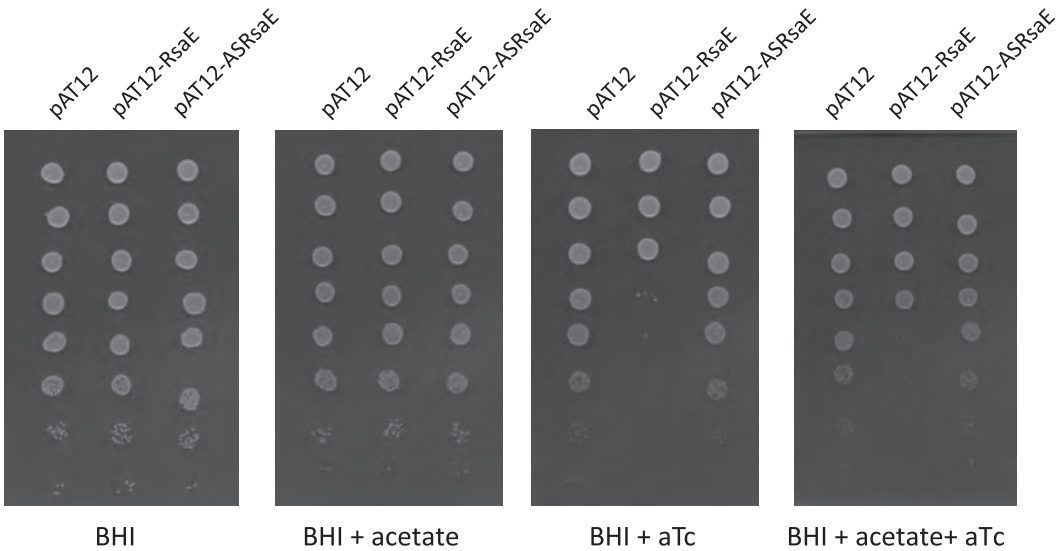


Figure 2. Accumulation of RsaE induces a growth defect that is partly alleviated by acetate. Overnight cultures were prepared in BHI from the strains listed below. Five-microliter spots of ten-fold serial dilutions of cultures were deposited from top (most concentrated) to bottom on BHI plates supplemented or not with aTc and acetate (100 mM) as indicated, and incubated at 37°C. Strains are RN6390 derivatives containing the indicated plasmids pAT12, pAT12-RsaE and pAT12-ASRsaE. pAT12-RsaE and pAT12-ASRsaE contain the *rsaE* gene and its antisense, respectively, both of which are under the transcriptional control of the P_{tet} promoter (inducible by aTc). Experiments were also performed at 30 and 42°C, and in LB medium, and gave similar results (data not shown).

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

Table 2. Variation of gene expressions upon 5 min of RsaE induction

Gene name[a]	Affim.[b]	qRT–PCR[c]	Δg[d]	Function
SA0845 (*opp3B*)	1.10	0.79	−25.47	Oligopeptide transport system
SA0849 (*opp3A*)[e]	2.18	0.76	−24.31	Oligopeptide transport system
SA0850 (*opp4A*)	4.38	6.06		Peptide/nickel transport system
SA0851 (*opp4D*)	3.15	4.59	−14.70	Peptide/nickel transport system
SA0873	0.48[f]		−21.13	Hypothetical protein
SA1088 (*sucC*)	0.54[f]		−18.24	Succinyl–CoA synthetase subunit beta
SA1089 (*sucD*)	0.47[f]		−15.51	Succinyl–CoA synthetase subunit alpha
SA1184 (*citB*)	0.47		−15.61	Aconitate hydratase
SA1365 (*gcvPB*)	0.48			Glycine dehydrogenase subunit 2
SA1366 (*gcvPA*)	0.36	0.48	−21.88	Glycine dehydrogenase subunit 1
SA1367 (*gcvT*)	0.32	0.36		Aminomethyltransferase
SA1432	0.43	0.46	−18.04	Hypothetical Mn^{2+} and Fe^{2+} transporter
SA1434	0.43			Acetyl–CoA carboxylase
SA1435	0.45			Acetyl–CoA carboxylase
SA1517 (*citC*)	0.40			Isocitrate dehydrogenase
SA1518 (*citZ*)	0.37	0.48	−14.80	Methylcitrate synthase
SA1553 (*fhs*)	0.30[f]		−14.08	Formate–tetrahydrofolate ligase
SA1858 (*ilvD*)	3.45	1.90		Dihydroxy-acid dehydratase
SA1859 (*ilvB*)	2.75			Acetolactate synthase large subunit
SA1860 (*ilvH*)	3.12			Acetolactate synthase 1 regulatory subunit
SA1861 (*ilvC*)	2.17			Ketol-acid reductoisomerase
SA1862 (*leuA*)	2.77			2-Isopropylmalate synthase
SA1863 (*leuB*)	2.56			3-Isopropylmalate dehydrogenase
SA1865 (*leuD*)	3.11			Isopropylmalate isomerase small subunit
SA1968 (*arg*)	0.35	0.44	−18.81	Arginase

RN6390 strains containing either pAT12-RsaE or p12-ASRsaE were grown at 37°C in BHI to OD600: 0.6. aTc was added to the medium and after 5 min, samples were removed and total RNA were extracted.
[a]SA0845 to SA849, SA1088-SA1089, SA1367-SA1365, SA1435 to SA1432, SA1518-SA1517, SA1858 to SA1865, likely organized in six distinct operons, respectively (genes for the same operon are expected to have a RsaE-dependent coordinated regulation).
[b]Gene expression ratios between strains containing pAT12-RsaE versus p12-ASRsaE obtained by Affimetrix microarray analysis (see also Supplementary Data S6).
[c]Gene expression ratios between strains containing pAT12-RsaE versus p12-ASRsaE obtained by qRT–PCR analysis. Experiments were also performed with samples taken at 0, 10, 20 and 30 min after the aTc addition to the media (Supplementary Data S7).
[d]Free energy of pairing between RsaE and the 5′ region of corresponding mRNAs as calculated by RNAup software (71).
[e]Discrepancy was observed between transcriptome and qRT–PCR experiments in the case of *opp3A* mRNA is possibly due to cross-hybridization of opp4A cDNA with the transcriptome opp3A probes.
[f]Results supported by published transcriptome results (36).

one-carbon shuffle, and (ii) SA1088-SA1089 (*sucC-sucD*), SA1184 (*citB*) and SA1517-SA1518 (*citC-citZ*) mRNA, which produce enzymes of the Krebs cycle. The growth inhibitory effect of RsaE overproduction is consistent with the observed downregulation of these genes; furthermore, addition of acetate to the media, which should allow acetyl–CoA formation for partial bypass of the TCA cycle, partly suppressed the toxicity (Figure 2).

Genes upregulated by RsaE include the operon for valine, leucine and isoleucine biosynthesis. In addition, several genes, likely encoding membrane proteins involved in peptide transport, are upregulated. Among them, two encode subunits of an oligopeptide permease (Opp) [SA0850 (*opp4A*), SA0851 (*opp4D*)].

The RsaE accumulation appears to correlate with an adaptation response aimed at reducing central metabolic pathways and increasing the amino acid pool.

RsaE binds the *opp3A* mRNA ribosome-binding site

To identify the primary RsaE targets among the putative candidates revealed by the transcriptome analysis, we performed a computational screen using RNAup. This software determines pairing energies between sRNAs and putative mRNA targets (71). Pairing energies between RsaE and the 5′ region (−150/+20) of all *S. aureus* CDSs were evaluated; one of the best scores was obtained with SA0849 (ΔG:−21 kcal/M) encoding Opp3A. The transcript of the two immediate *opp3A*-downstream genes strongly accumulates upon RsaE induction (Table 2 and Supplementary Data S5 and S6). These *opp* genes encode components of ABC transporters possibly involved in peptide/nickel transport (72). The combination of computational and transcriptome data indicates that RsaE could affect *opp* regulation by pairing with the SA0849 mRNA.

The interaction of RsaE with different mRNAs predicted to be RsaE targets (i.e. by trancriptome results and computational analyses using RNAup) was tested by gel retardation assays. The candidates correspond to (i) *opp3A* (SA0849) and *opp4D* (SA0851) encoding components of oligopeptide transport systems (72), (ii) SA1432 and SA0645 encoding putative manganese and anion transporters, respectively, (iii) *sucC* (SA1088) and *sucD* (SA1089) encoding the succinyl–CoA synthetase alpha and beta subunits, respectively, (iv) *gcvP* (SA1366) and *gcvT* (SA1367), encoding enzymes from the glycine cleavage pathway, (v) *acsA* (SA1554) encoding the acetyl–CoA synthetase and (vi) *arg* (SA1968) encoding

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

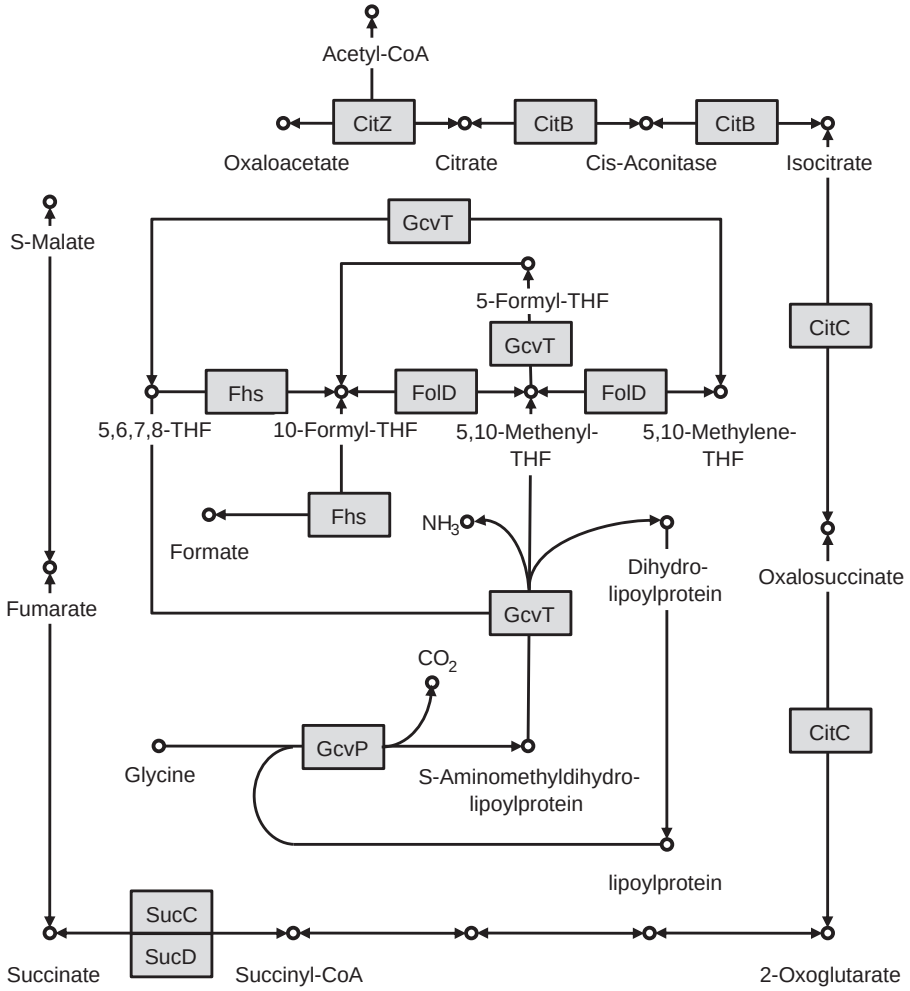


Figure 3. RsaE accumulation downregulates specific metabolic pathways. The metabolic charts are adapted from the Kegg database (90). Enzymes corresponding to downregulated genes by RsaE are indicated by square boxes highlighted in grey [FolD [SA0915] indicated on the figure was not found in this study, but was previously reported (36)]; metabolites are shown as open circles; biochemical reactions that are demonstrated or predicted in N315 by arrows, respectively. The outer pathway is the TCA cycle. The inner pathway is the THF biosynthesis pathway. In certain organisms, these two pathways are connected *via* the glyoxylate shunt, which leads to a modified version of the TCA cycle that bypasses certain steps. However, connecting enzymes in N315 are not yet known.

arginase. (Additional candidates based solely on bio-computing data were also tested: *opp3B* [SA0845], *opp4C* [SA0854], SAS045, SA0308, SA0785 and SA2118.) A synthetic full-length RsaE and truncated mRNAs (~180 nt long; for oligonucleotides used, see Supplementary Data S1) were produced from PCR-generated genes placed under the control of the T7 RNA polymerase using an *in vitro* transcription assay. The putative RsaE targets tested were about 140-nt long with about 80-nt corresponding to the Shine-Dalgarno (SD) upstream sequences. Among these, only one mRNA, corresponding to *opp3A* (SA0849), formed a complex with RsaE *in vitro* (Figure 4A and B). Complex formation between RsaE and SA0849 is specific since a thousand fold excess of tRNA did not disrupt the sRNA–mRNA duplex (Figure 4B).

Predicted pairings between RsaE and the SA0849 mRNA (Figure 6) suggest that RsaE interacts with the opp3A SD sequence, as well as part of the AUG initiation codon and therefore could prevent translation initiation. To test this, toeprint assays were performed on ternary initiation complexes including purified 70S ribosomes, initiator $tRNA^{fMet}$ and the SA0849 mRNA. Two ribosome toeprints were detected on the mRNA, at 14 and 15-nt downstream from the initiation codon respectively (Figure 4C), supporting the location of the SA0849 mRNA start codon as drawn on Figure 6. RsaE reduced ribosome loading onto the SA0849 mRNA in a concentration-dependent manner (Figure 4C). This demonstrates the formation of a translational initiation complex on the SA0849 mRNA. We conclude that RsaE inhibits SA0849 mRNA translation by antisense pairing, thus preventing ribosome binding. The fact that RsaE (i) is predicted to efficiently pair with the *opp3A* mRNA at the SD site, (ii) forms a specific complex *in vitro* with the *opp3A* mRNA transcript and (iii) prevents *in vitro* formation of an *opp3A* mRNA translation initiation complex, together suggest that RsaE targets interacts with the *oppA3* mRNA and prevents its translation.

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

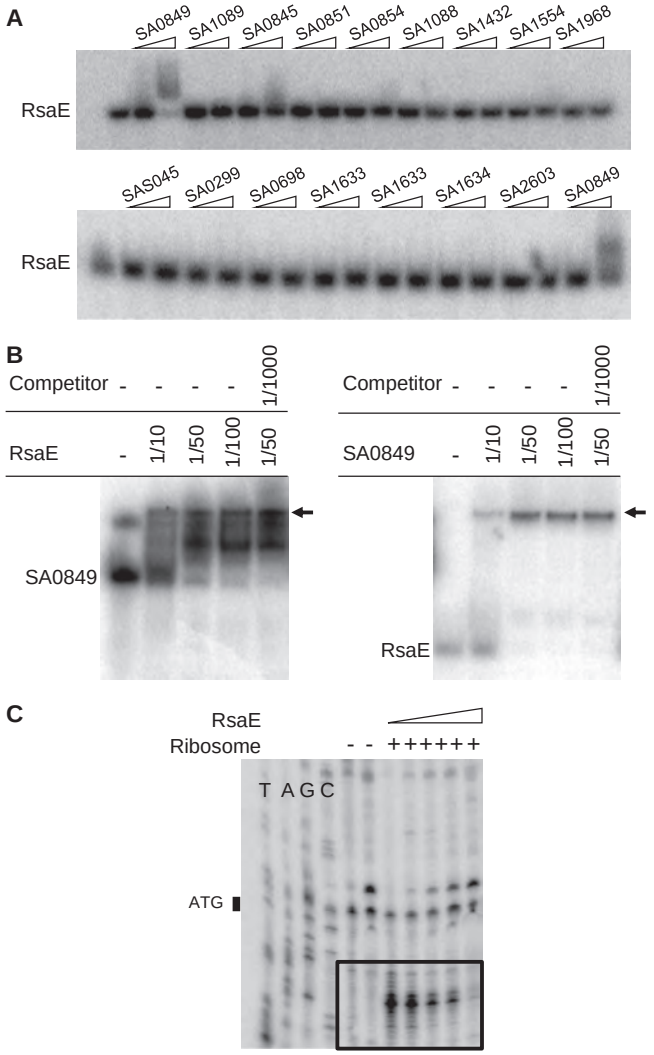


Figure 4. RsaE interacts with the 5′ end region of *opp3A* (SA0849) mRNA. (**A**) Gel retardation assay of labeled RsaE (0.4 pmol) incubated with synthetic RNAs (4 and 40 pmol), corresponding to different 5′ gene ends of the indicated genes (see Supplementary Data S1 for the sequences). The RsaE/SA0849 mRNA interaction is shown by a mobility shift of labeled RsaE in the presence of SA0849 mRNA (RsaE mobility was not affected by the other tested RNAs). (**B**) Left: Gel retardation assay of synthetic labeled SA0849 mRNA 5′ end fragment (0.4 pmol) incubated with increasing concentrations (molar ratio is indicated) of synthetic RsaE. Right: Gel retardation assay of labeled RsaE (0.4 pmol) incubated with increasing concentrations (molar ratio is indicated) of synthetic SA0849 mRNA 5′ end fragment. Addition of a large excess of unlabelled tRNAs did not alter the RsaE/SA0849 mRNA interactions in either assay. (**C**) RsaE prevents ribosome loading and translation initiation on opp3A SA0849 mRNA. Ribosome toeprints on SA0849 mRNA. '–/+' indicates the absence or presence of purified ribosomes. The wedge indicates increasing amounts of RsaE. The experimentally determined toeprints are boxed. T, A, G and C refer to the SA0849 mRNA sequencing ladders. The position of the SA0849 mRNA initiation codon is indicated (ATG).

DISCUSSION

We used an RNomics approach based on deep sequencing of RNA enriched for primary transcripts (73) to search for novel staphylococcal sRNAs. Our analysis of 36 000 reads identified 30 known or putative sRNAs encoded within IGRs (Table 1 and Figure 5). Fourteen sRNAs confirmed by northern blots were categorized as: (i) sRNAs that could target genetically unlinked mRNAs (eight new instances) or (ii) sRNAs that pair with the UTRs of mRNAs expressed from the adjacent gene (two new instances). We also detected small transcripts likely associated with three riboswitch activities**.** Seven sRNAs were expressed from the opposite strand of ORFs; one is a bona fide *cis*-encoded RNA. We recently predicted 50 regions as encoding sRNAs; out of 24 tested, seven were confirmed to express sRNAs (35). Additionally, five of the newly reported sRNAs (RsaE, RsaOM, RsaOP, RsaOQ and RsaOW) are among the untested previously predicted regions. This observation brings further support for this computational screen based on phylogenetic profiling, which is particularly efficient at detecting novel bacterial sRNAs (35).

Most *S. aureus* RNAs are differentially expressed in a growth phase-dependent fashion (34,36,58). The detailed analysis of sRNA expression performed in this study reveals that some sRNAs accumulate in exponential (RsaOH, RsaOR), pre-stationary (RsaH, RsaE, RsaOT, RsaOU) or stationary (RsaOI, RsaA, RsaOP, RsaOV) phases. We observed abrupt changes in the expression levels of several sRNAs between OD_{600} = 6.0 and stationary phase (OD_{600} = 10). It is striking that several sRNAs such as RsaH and RsaE accumulate specifically in pre-stationary phase, and disappear in stationary phase, suggesting a role in adaptation to growth phase changes. Numerous staphylococcal regulatory functions are coordinated with environmental conditions or growth phase. Virulence factor expression is correlated with growth phase: surface adhesins are generally expressed in exponential phase, whereas exoproteins are expressed in stationary phase. The newly identified sRNAs conceivably play role(s) in this regulation, in addition to RNAIII (29). We tested the effect of several regulators on the expression of *rsaE*. RsaE quantities were compared in isogenic 8325-4 strains deficient for regulators SigB (74), SarA (75), Agr (76), PerR (77), Fur (78) or Spa (79), or proficient for RsbU (80); no significant differences were observed in any of those genetic contexts, suggesting that RsaE expression is probably regulated independently of these regulators (data not shown). A σ^A-consensus binding site is located just upstream of the *rsaE* start site, indicating expression driven by the *S. aureus* vegetative σ factor.

The primary RsaE nucleotide sequence is surprisingly well conserved among the sequenced bacillales, and RsaE over-expression is toxic. Altogether, these observations suggest important functional role(s) for RsaE. A variation in the quantity of RsaE induced a significant change in the transcriptional pattern. Among the up or downregulated genes, many express mRNAs that are predicted to pair with RsaE. Structure probing (36) and computational analyses suggest that RsaE folds as two stem loops separated by a 17-nt-long RNA single strand. RsaE has a duplicated, CCCCUUUGUUU accessible sequence with one copy in the loop of the first hairpin and a second within the central single stranded domain. The predicted pairings between RsaE and the targeted mRNAs

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

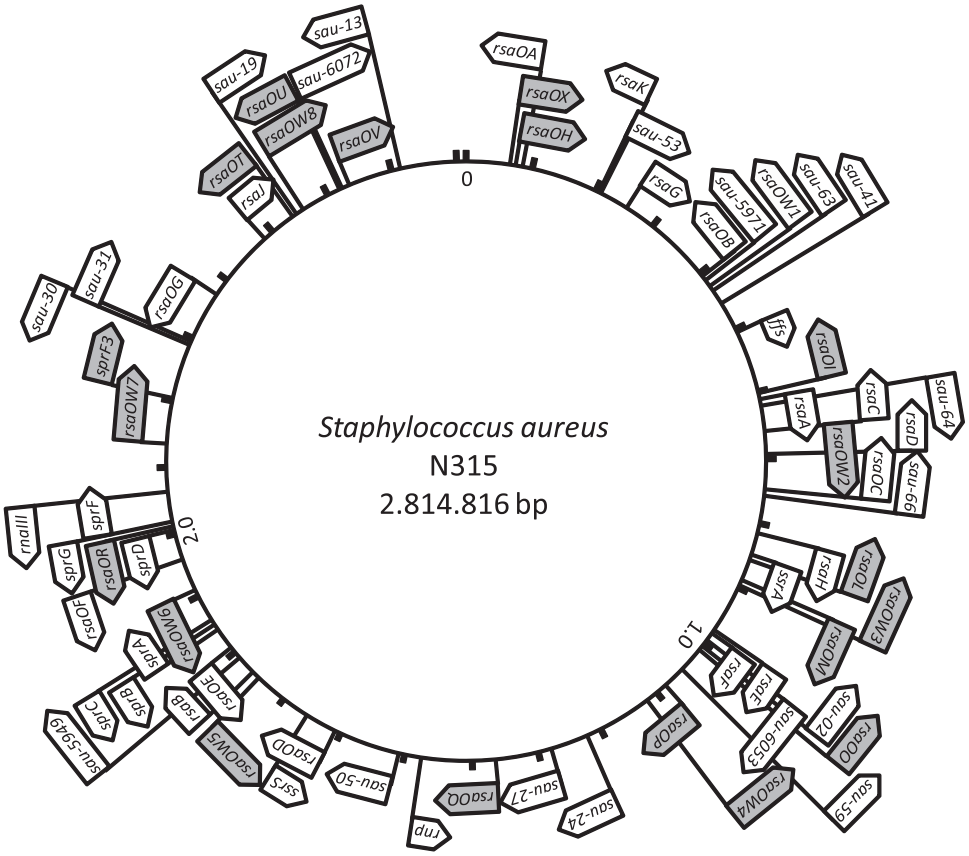


Figure 5. Genomic map showing the distribution of *S. aureus* sRNAs. Flags indicate the names of sRNAs that were predicted and confirmed by northern blot analysis; sRNAs uncovered by the present work are highlighted in grey (34–37,58).

involve the duplicated sequence and the mRNA SD sequences (Supplementary Data S7). Repetition of that interacting sequence on the sRNA suggests several pairing combinations with its mRNA substrates, involving either the RsaE apical loop 1 or its unstructured central domain.

The *opp* operons encode ABC transporters involved in the transport of small peptides. *Staphylococcus aureus* has several putative *opp* operons, three of which (*opp-1*, *opp-2* and *opp-3*) are conserved in all strains (72). The *opp-3* operon is dedicated to nitrogen nutrition by peptide import and is also involved in the regulation of extracellular proteases (SspA and Aur) via an unknown mechanism (81). In several Enterobacteriales, expression of dipeptide, oligopeptide and microcin ABC transport systems are partly controlled by sRNAs such as GcvB and RydC (52,82–84). RsaE targets the *opp* operons and is thus a possible GcvB/RydC functional analog. However, we detected no sequence similarities between RsaE and these sRNAs. In N315, the *opp-3* operon comprises *opp3B*, *opp3C*, *opp3D*, *opp3F* and *opp3A*; it is followed by a second *opp* operon, *opp-4*, comprising *opp4A* and *opp4D* (72). The function of the *opp-4* operon is at present unknown. Toeprint experiments show that RsaE interacts with the *opp-3* mRNA (Figure 4). Surprisingly, RsaE induction had no significant effects on the amounts of *opp-3* mRNAs, but strongly upregulated the downstream operon, *opp-4* (quantified by qRT–PCR, cf. Table 2). Several interpretations can be considered for these results: (i) the observed RsaE mediated inhibition of ribosome loading on *opp3A* mRNA (Figure 4) could in turn, upregulate the *opp-4* operon (i.e. Opp proteins could mediate this regulation) (ii) interactions between RsaE and *opp-3* mRNA generate an anti-termination mechanism that leads to increased *opp-4* operon transcription or (iii) RsaE interacts directly with the *opp-4* mRNA preventing its degradation; however, we did not observe RsaE gel retardation by the *opp4D* 5′ mRNA. Recently, RsaE was reported to target the SD sequence of the *opp3B* mRNA, the first gene of the *opp-3* operon (36). Here, we provide evidence that RsaE targets *opp3A*, the last gene of the *opp3* operon. These observations suggest that RsaE operates simultaneously on two genes of the same mRNA, revealing an unusual mode of sRNA action (Figure 6).

Our results point to a regulatory role of RsaE in primary metabolism. It is astonishing that four enzymes converging directly to the 5, 10 methenyl-THF have their mRNA coordinately downregulated by the same sRNA, RsaE (Figure 3). They correspond to glycine decarboxylase (GcvP) and aminomethyltransferase (GcvT) which belong to the glycine cleavage pathway that provides the one-carbon units to methenyl-THF, and the formate-tetrahydrofolate ligase (Fhs) and methylene-THF

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

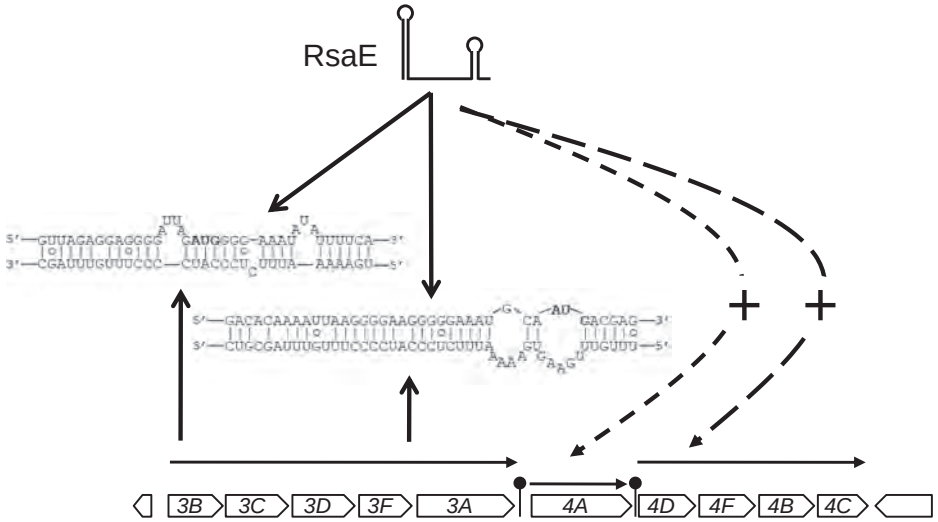


Figure 6. RsaE targets two SD sequences within the same mRNA. The transcriptional organization of the *S. aureus* N315 *opp-3* and *opp-4* operons is presented [based on (72)]. RsaE prevents ribosome loading onto *opp-3* mRNA in front of *opp3B* (36) and *opp3A* ORF (this work). Predicted base pairing between RsaE and the *opp-3* operon is indicated. RsaE positively regulates *opp-3* downstream genes. Dashed lines indicate RsaE-mediated activation, as deduced from microarray and qRT–PCR results.

enzyme (FolD) (36) which convert THF into methenyl-THF. This latter compound is the supplier of one-carbon units for the synthesis of purines and the formyl group of *N*-formylmethionyl-tRNA(f). It is also striking that RsaE downregulates four key enzymes of the tricarboxylic acid (TCA) cycle: the succinyl–CoA synthetase (SucC-SucD) [also in ref (36)], the aconitase (CitB), the isocitrate dehydrogenase (CitC) and the citrate synthase (CitZ) (Table 2, Figure 3). The TCA cycle is involved in the synthesis of usable energy. General downregulation of these different pathways might induce the observed toxicity of *rsaE* over-expression via an ectopic copy. The partial toxicity alleviation by acetate could be due to the utilization of glyoxylate pathway that shunts some RsaE-repressed steps of the TCA cycle. As endogenous RsaE is induced in a pre-stationary phase, an expected role for this sRNA would be to coordinate downregulation of energy metabolism (TCA cycle), and of cofactors, vitamins, translation, and purine biosynthesis (folate-dependent one-carbon metabolism) to facilitate adaptation to the entry into stationary phase.

CcpA is a master regulator for carbon catabolite regulation in numerous Gram-positive bacteria. Many genes that are also modulated by RsaE (e.g. *sucC*, *sucD*, *arg*, *citC*, *citZ* and *ilvB*) show CcpA-dependent regulation (85–87). This suggested that RsaE could affect a CcpA. However, accumulation of RsaE (by means of pAT12-RsaE) did not affect the quantity of the ccpA mRNA (Supplementary Data S5 and S6). In addition, we did not observe any band shift of the *ccpA* UTR mRNA sequence by RsaE (data not shown) suggesting that the RsaE effect is likely not *via* CcpA.

Computer predictions indicate the possibility that RsaE-dependent downregulation is mediated via pairing of RsaE with mRNA 5′ UTRs. However, *in vitro* experiments performed between RsaE and synthetic transcripts of the candidate UTR sequences displayed gel retardation with only one mRNA sequence. It is possible that the sRNA/RNA target binding in *S. aureus* is more labile *in vitro* than with sRNAs from other species such as *E. coli* (possibly because sequences are A/U-rich). For example, the *sucD* mRNA was previously identified and validated as an RsaE target using *in vitro* toeprinting (36); however, the RsaE-*sucD* interaction seemed too labile to be detected in our gel retardation assays (Figure 3A). In *E. coli*, the activity of *trans*-encoded RNA usually requires the RNA chaperone protein Hfq, but it is not the case for known sRNA–mRNA interactions in *S. aureus* (4,30,32,36); thus an unknown factor might be required to promote such regulatory interactions in *S. aureus*. Further evidence for RsaE-dependent gene downregulation via a direct interaction of RsaE with 5′ mRNA UTRs was obtained using a statistical method: Indeed, the probability that the seven of the 22 most downregulated genes (among a total of 2790 *S. aureus* genes) are found by random among the 139 best RsaE/mRNA matching scores (determined using the RNAup software) is 0.000054 (data from Table 2 and Supplementary Table S5). This result rules out the possibility of the random appearance of so many genes matching RsaE among the RsaE-dependent downregulated genes. Strikingly, among these 22 genes that are the most downregulated, 11 are directly involved in Krebs/one carbon by folate metabolism, providing further indication that a pairing with RsaE is required for the observed downregulation of metabolic pathways.

Since RsaE is induced in late exponential phase when the growth medium is depleted, we speculate that RsaE would contribute to downregulating bacterial metabolism (one carbon by folate and TCA cycle), and at the same time activate Opp transporters, as a last-ditch effort to adapt to limited nutrient conditions as cells approach stationary phase.

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

sRNAs are often expressed during specific growth conditions and have numerous targets. As such, they constitute key regulators of an intricate network of metabolic pathways and nutrient uptake systems. So far, most of the evidence for these roles comes from studies on enteric bacteria, while little is known concerning Gram-positive bacteria (36,88,89); results on RsaE make it obvious that the concept of coordinated control of metabolic enzymes by sRNAs should be extended to the Gram-positive pathogen, *S. aureus*.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Pascale Romby and her colleagues for sharing results prior publication, and Armel Guyonvarch, Michelle David and Stella Artois for helpful discussions, technical assistance and warm support. The authors thank Richard Reinhardt and Bernd Timmermann for pyrosequencing and Vertis Biotechnology AG for cDNA construction. The authors thank Sandy Gruss for critical reading of the manuscript and helpful discussions.

FUNDING

Grant 'Microbiologie-Immunologie-Maladies Emergentes' from the 'Agence Nationale pour la Recherche' (ANR-06-MIME-016-01); 'Centre National de la Recherche Scientifique (CNRS)/Conseil Général de l'Essonne' and 'Agence Nationale pour la Recherche' fellowship (to C.R.); 'DIM Malinf' fellowship from 'Région Ile de France' (to P.S.). Funding for open access charge: CNRS.

Conflict of interest statement. None declared.

REFERENCES

1. Waters,L.S. and Storz,G. (2009) Regulatory RNAs in bacteria. *Cell*, **136**, 615–628.
2. Brantl,S. (2007) Regulatory mechanisms employed by cis-encoded antisense RNAs. *Curr. Opin. Microbiol.*, **10**, 102–109.
3. Brennan,R.G. and Link,T.M. (2007) Hfq structure, function and ligand binding. *Curr. Opin. Microbiol.*, **10**, 125–133.
4. Bohn,C., Rigoulay,C. and Bouloc,P. (2007) No detectable effect of RNA-binding protein Hfq absence in *Staphylococcus aureus*. *BMC Microbiol.*, **7**, 10.
5. Babitzke,P. and Romeo,T. (2007) CsrB sRNA family: sequestration of RNA-binding regulatory proteins. *Curr. Opin. Microbiol.*, **10**, 156–163.
6. Wassarman,K.M. (2007) 6S RNA: a small RNA regulator of transcription. *Curr. Opin. Microbiol.*, **10**, 164–168.
7. Pichon,C. and Felden,B. (2007) Proteins that interact with bacterial small RNA regulators. *FEMS Microbiol. Rev.*, **31**, 614–625.
8. Henkin,T.M. (2008) Riboswitch RNAs: using RNA to sense cellular metabolism. *Genes Dev.*, **22**, 3383–3390.
9. Vogel,J., Bartels,V., Tang,T.H., Churakov,G., Slagter-Jager,J.G., Huttenhofer,A. and Wagner,E.G. (2003) RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res.*, **31**, 6435–6443.
10. Loh,E., Dussurget,O., Gripenland,J., Vaitkevicius,K., Tiensuu,T., Mandin,P., Repoila,F., Buchrieser,C., Cossart,P. and Johansson,J. (2009) A trans-acting riboswitch controls expression of the virulence regulator PrfA in *Listeria monocytogenes*. *Cell*, **139**, 770–779.
11. Altuvia,S. (2007) Identification of bacterial small non-coding RNAs: experimental approaches. *Curr. Opin. Microbiol.*, **10**, 257–261.
12. Sharma,C.M. and Vogel,J. (2009) Experimental approaches for the discovery and characterization of regulatory small RNA. *Curr. Opin. Microbiol.*, **12**, 536–546.
13. Gorke,B. and Vogel,J. (2008) Noncoding RNA control of the making and breaking of sugars. *Genes Dev.*, **22**, 2914–2925.
14. Horler,R.S. and Vanderpool,C.K. (2009) Homologs of the small RNA SgrS are broadly distributed in enteric bacteria but have diverged in size and sequence. *Nucleic Acids Res.*, **37**, 5465–5476.
15. Masse,E. and Gottesman,S. (2002) A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **99**, 4620–4625.
16. Masse,E., Vanderpool,C.K. and Gottesman,S. (2005) Effect of RyhB small RNA on global iron use in Escherichia coli. *J. Bacteriol.*, **187**, 6962–6971.
17. Jacques,J.F., Jang,S., Prevost,K., Desnoyers,G., Desmarais,M., Imlay,J. and Masse,E. (2006) RyhB small RNA modulates the free intracellular iron pool and is essential for normal growth during iron limitation in *Escherichia coli*. *Mol. Microbiol.*, **62**, 1181–1190.
18. Guillier,M., Gottesman,S. and Storz,G. (2006) Modulating the outer membrane with small RNAs. *Genes Dev.*, **20**, 2338–2348.
19. Papenfort,K., Pfeiffer,V., Mika,F., Lucchini,S., Hinton,J.C. and Vogel,J. (2006) SigmaE-dependent small RNAs of Salmonella respond to membrane stress by accelerating global omp mRNA decay. *Mol. Microbiol.*, **62**, 1674–1688.
20. Boysen,A., Moller-Jensen,J., Kallipolitis,B.H., Valentin-Hansen,P. and Overgaard,M. (2010) Translational regulation of gene expression by an anaerobically induced small non-coding RNA in Escherichia coli. *J. Biol. Chem.*, **285**, 10690–10702.
21. Durand,S. and Storz,G. (2010) Reprogramming of anaerobic metabolism by the FnrS small RNA. *Mol. Microbiol.*, **75**, 1215–1231.
22. Lenz,D.H., Mok,K.C., Lilley,B.N., Kulkarni,R.V., Wingreen,N.S. and Bassler,B.L. (2004) The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in *Vibrio harveyi* and *Vibrio cholerae*. *Cell*, **118**, 69–82.
23. Christiansen,J.K., Nielsen,J.S., Ebersbach,T., Valentin-Hansen,P., Sogaard-Andersen,L. and Kallipolitis,B.H. (2006) Identification of small Hfq-binding RNAs in *Listeria monocytogenes*. *RNA*, **12**, 1383–1396.
24. Sousa,S.A., Ramos,C.G., Moreira,L.M. and Leitao,J.H. (2010) The *hfq* gene is required for stress resistance and full virulence of *Burkholderia cepacia* to the nematode *Caenorhabditis elegans*. *Microbiology*, **156**, 896–908.
25. Chao,Y. and Vogel,J. (2010) The role of Hfq in bacterial pathogens. *Curr. Opin. Microbiol.*, **13**, 24–33.
26. Toledo-Arana,A., Repoila,F. and Cossart,P. (2007) Small noncoding RNAs controlling pathogenesis. *Curr. Opin. Microbiol.*, **10**, 182–188.
27. Kuroda,M., Ohta,T., Uchiyama,I., Baba,T., Yuzawa,H., Kobayashi,I., Cui,L., Oguchi,A., Aoki,K. and Nagai,Y.F. (2001) Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. *Lancet*, **357**, 1225–1240.
28. Klevens,R.M., Morrison,M.A., Nadle,J., Petit,S., Gershman,K., Ray,S., Harrison,L.H., Lynfield,R., Dumyati,G., Townes,J.M. *et al.* (2007) Invasive methicillin-resistant *Staphylococcus aureus* infections in the United States. *JAMA*, **298**, 1763–1771.
29. Novick,R.P. (2003) Autoinduction and signal transduction in the regulation of staphylococcal virulence. *Mol. Microbiol.*, **48**, 1429–1449.
30. Boisset,S., Geissmann,T., Huntzinger,E., Fechter,P., Bendridi,N., Possedko,M., Chevalier,C., Helfer,A.C., Benito,Y., Jacquier,A. *et al.* (2007) *Staphylococcus aureus* RNAIII coordinately represses

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

the synthesis of virulence factors and the transcription regulator Rot by an antisense mechanism. *Genes Dev.*, **21**, 1353–1366.
31. Huntzinger,E., Boisset,S., Saveanu,C., Benito,Y., Geissmann,T., Namane,A., Lina,G., Etienne,J., Ehresmann,B., Ehresmann,C. *et al.* (2005) *Staphylococcus aureus* RNAIII and the endoribonuclease III coordinately regulate *spa* gene expression. *EMBO J.*, **24**, 824–835.
32. Geisinger,E., Adhikari,R.P., Jin,R., Ross,H.F. and Novick,R.P. (2006) Inhibition of rot translation by RNAIII, a key feature of *agr* function. *Mol. Microbiol.*, **61**, 1038–1048.
33. Anderson,K.L. and Dunman,P.M. (2009) Messenger RNA turnover processes in Escherichia coli, Bacillus subtilis, and emerging studies in Staphylococcus aureus. *Int. J. Microbiol.*, **2009**, 525491.
34. Pichon,C. and Felden,B. (2005) Small RNA genes expressed from *Staphylococcus aureus* genomic and pathogenicity islands with specific expression among pathogenic strains. *Proc. Natl Acad. Sci. USA*, **102**, 14249–14254.
35. Marchais,A., Naville,M., Bohn,C., Bouloc,P. and Gautheret,D. (2009) Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles. *Genome Res.*, **19**, 1084–1092.
36. Geissmann,T., Chevalier,C., Cros,M.J., Boisset,S., Fechter,P., Noirot,C., Schrenzel,J., Francois,P., Vandenesch,F., Gaspin,C. *et al.* (2009) A search for small noncoding RNAs in *Staphylococcus aureus* reveals a conserved sequence motif for regulation. *Nucleic Acids Res.*, **37**, 7239–7257.
37. Abu-Qatouseh,L.F., Chinni,S.V., Seggewiss,J., Proctor,R.A., Brosius,J., Rozhdestvensky,T.S., Peters,G., von Eiff,C. and Becker,K. (2010) Identification of differentially expressed small non-protein-coding RNAs in Staphylococcus aureus displaying both the normal and the small-colony variant phenotype. *J. Mol. Med.*, **88**, 565–75.
38. Anderson,K.L., Roberts,C., Disz,T., Vonstein,V., Hwang,K., Overbeek,R., Olson,P.D., Projan,S.J. and Dunman,P.M. (2006) Characterization of the *Staphylococcus aureus* heat shock, cold shock, stringent, and SOS responses and their effects on log-phase mRNA turnover. *J. Bacteriol.*, **188**, 6739–6756.
39. Roberts,C., Anderson,K.L., Murphy,E., Projan,S.J., Mounts,W., Hurlburt,B., Smeltzer,M., Overbeek,R., Disz,T. and Dunman,P.M. (2006) Characterizing the effect of the *Staphylococcus aureus* virulence factor regulator, SarA, on log-phase mRNA half-lives. *J. Bacteriol.*, **188**, 2593–2603.
40. Peng,H.L., Novick,R.P., Kreiswirth,B., Kornblum,J. and Schlievert,P. (1988) Cloning, characterization, and sequencing of an accessory gene regulator (*agr*) in *Staphylococcus aureus*. *J. Bacteriol.*, **170**, 4365–4372.
41. Iordanescu,S. and Surdeanu,M. (1976) Two restriction and modification systems in *Staphylococcus aureus* NCTC8325. *J. Gen. Microbiol.*, **96**, 277–281.
42. Cheung,A.L., Eberhardt,K.J., Chung,E., Yeaman,M.R., Sullam,P.M., Ramos,M. and Bayer,A.S. (1994) Diminished virulence of a sar-/agr- mutant of Staphylococcus aureus in the rabbit model of endocarditis. *J. Clin. Invest.*, **94**, 1815–1822.
43. Booth,M.C., Cheung,A.L., Hatter,K.L., Jett,B.D., Callegan,M.C. and Gilmore,M.S. (1997) Staphylococcal accessory regulator (sar) in conjunction with agr contributes to *Staphylococcus aureus* virulence in endophthalmitis. *Infect. Immun.*, **65**, 1550–1556.
44. Kreiswirth,B.N., Lofdahl,S., Betley,M.J., O'Reilly,M., Schlievert,P.M., Bergdoll,M.S. and Novick,R.P. (1983) The toxic shock syndrome exotoxin structural gene is not detectably transmitted by a prophage. *Nature*, **305**, 709–712.
45. Bateman,B.T., Donegan,N.P., Jarry,T.M., Palma,M. and Cheung,A.L. (2001) Evaluation of a tetracycline-inducible promoter in *Staphylococcus aureus* in vitro and in vivo and its application in demonstrating the role of *sigB* in microcolony formation. *Infect. Immun.*, **69**, 7851–7857.
46. Sambrook,J. and Russell,D. (2001) *Molecular Cloning: A Laboratory Manual*, 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
47. Rigoulay,C., Entenza,J.M., Halpern,D., Widmer,E., Moreillon,P., Poquet,I. and Gruss,A. (2005) Comparative analysis of the roles of HtrA-like surface proteases in two virulent *Staphylococcus aureus* strains. *Infect. Immun.*, **73**, 563–572.
48. Oh,E.T. and So,J.S. (2003) A rapid method for RNA preparation from Gram-positive bacteria. *J. Microbiol. Methods*, **52**, 395–398.
49. Douchin,V., Bohn,C. and Bouloc,P. (2006) Down-regulation of porins by a small RNA bypasses the essentiality of the regulated intramembrane proteolysis protease RseP in *Escherichia coli*. *J. Biol. Chem.*, **281**, 12253–12259.
50. Argaman,L., Hershberg,R., Vogel,J., Bejerano,G., Wagner,E.G., Margalit,H. and Altuvia,S. (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.*, **11**, 941–950.
51. Bensing,B.A., Meyer,B.J. and Dunny,G.M. (1996) Sensitive detection of bacterial transcription initiation sites and differentiation from RNA processing sites in the pheromone-induced plasmid transfer system of *Enterococcus faecalis*. *Proc. Natl Acad. Sci. USA*, **93**, 7794–7799.
52. Antal,M., Bordeau,V., Douchin,V. and Felden,B. (2005) A small bacterial RNA regulates a putative ABC transporter. *J. Biol. Chem.*, **280**, 7901–7908.
53. Benito,Y., Kolb,F.A., Romby,P., Lina,G., Etienne,J. and Vandenesch,F. (2000) Probing the structure of RNAIII, the *Staphylococcus aureus agr* regulatory RNA, and identification of the RNA domain involved in repression of protein A expression. *RNA*, **6**, 668–679.
54. Hartz,D., McPheeters,D.S., Traut,R. and Gold,L. (1988) Extension inhibition analysis of translation initiation complexes. *Methods Enzymol.*, **164**, 419–425.
55. Sittka,A., Lucchini,S., Papenfort,K., Sharma,C.M., Rolle,K., Binnewies,T.T., Hinton,J.C. and Vogel,J. (2008) Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet.*, **4**, e1000163.
56. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
57. Bury-Mone,S., Nomane,Y., Reymond,N., Barbet,R., Jacquet,E., Imbeaud,S., Jacq,A. and Bouloc,P. (2009) Global analysis of extracytoplasmic stress signaling in Escherichia coli. *PLoS Genet.*, **5**, e1000651.
58. Novick,R.P., Ross,H.F., Projan,S.J., Kornblum,J., Kreiswirth,B. and Moghazeh,S. (1993) Synthesis of staphylococcal virulence factors is controlled by a regulatory RNA molecule. *EMBO J.*, **12**, 3967–3975.
59. Berghoff,B.A., Glaeser,J., Sharma,C.M., Vogel,J. and Klug,G. (2009) Photooxidative stress-induced and abundant small RNAs in *Rhodobacter sphaeroides*. *Mol. Microbiol.*, **74**, 1497–1512.
60. Case,C.C., Simons,E.L. and Simons,R.W. (1990) The IS10 transposase mRNA is destabilized during antisense RNA control. *EMBO J.*, **9**, 1259–1266.
61. Ma,C. and Simons,R.W. (1990) The IS10 antisense RNA blocks ribosome binding at the transposase translation initiation site. *EMBO J.*, **9**, 1267–1274.
62. Zhu,L., Inoue,K., Yoshizumi,S., Kobayashi,H., Zhang,Y., Ouyang,M., Kato,F., Sugai,M. and Inouye,M. (2009) *Staphylococcus aureus* MazF specifically cleaves a pentad sequence, UACAU, which is unusually abundant in the mRNA for pathogenic adhesive factor SraP. *J. Bacteriol.*, **191**, 3248–3255.
63. Gardner,P.P., Daub,J., Tate,J.G., Nawrocki,E.P., Kolbe,D.L., Lindgreen,S., Wilkinson,A.C., Finn,R.D., Griffiths-Jones,S., Eddy,S.R. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, D136–D140.
64. Grundy,F.J. and Henkin,T.M. (1998) The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in gram-positive bacteria. *Mol. Microbiol.*, **30**, 737–749.
65. Epshtein,V., Mironov,A.S. and Nudler,E. (2003) The riboswitch-mediated control of sulfur metabolism in bacteria. *Proc. Natl Acad. Sci. USA*, **100**, 5052–5056.
66. Winkler,W.C., Nahvi,A., Sudarsan,N., Barrick,J.E. and Breaker,R.R. (2003) An mRNA structure that controls gene expression by binding S-adenosylmethionine. *Nat. Struct. Biol.*, **10**, 701–707.
67. Naville,M. and Gautheret,D. (2009) Transcription attenuation in bacteria: theme and variations. *Brief. Funct. Genomic Proteomic*, **8**, 482–492.

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

68. Grundy,F.J., Rollins,S.M. and Henkin,T.M. (1994) Interaction between the acceptor end of tRNA and the T box stimulates antitermination in the *Bacillus subtilis tyrS* gene: a new role for the discriminator base. *J. Bacteriol.*, **176**, 4518–4526.
69. Langbein,I., Bachem,S. and Stulke,J. (1999) Specific interaction of the RNA-binding domain of the bacillus subtilis transcriptional antiterminator GlcT with its RNA target, RAT. *J. Mol. Biol.*, **293**, 795–805.
70. Knezevic,I., Bachem,S., Sickmann,A., Meyer,H.E., Stulke,J. and Hengstenberg,W. (2000) Regulation of the glucose-specific phosphotransferase system (PTS) of *Staphylococcus carnosus* by the antiterminator protein GlcT. *Microbiology*, **146(Pt 9)**, 2333–2342.
71. Muckstein,U., Tafer,H., Hackermuller,J., Bernhart,S.H., Stadler,P.F. and Hofacker,I.L. (2006) Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**, 1177–1182.
72. Hiron,A., Borezee-Durant,E., Piard,J.C. and Juillard,V. (2007) Only one of four oligopeptide transport systems mediates nitrogen nutrition in *Staphylococcus aureus*. *J. Bacteriol.*, **189**, 5119–5129.
73. Sharma,C.M., Hoffmann,S., Darfeuille,F., Reignier,J., Findeiss,S., Sittka,A., Chabas,S., Reiche,K., Hackermuller,J., Reinhardt,R. *et al.* (2010) The primary transcriptome of the major human pathogen. *Helicobacter pylori*. *Nature*, **464**, 250–255.
74. Chan,P.F., Foster,S.J., Ingham,E. and Clements,M.O. (1998) The *Staphylococcus aureus* alternative sigma factor sigmaB controls the environmental stress response but not starvation survival or pathogenicity in a mouse abscess model. *J. Bacteriol.*, **180**, 6082–6089.
75. Chan,P.F. and Foster,S.J. (1998) Role of SarA in virulence determinant production and environmental signal transduction in *Staphylococcus aureus*. *J. Bacteriol.*, **180**, 6232–6241.
76. Chan,P.F. and Foster,S.J. (1998) The role of environmental factors in the regulation of virulence-determinant expression in *Staphylococcus aureus* 8325-4. *Microbiology*, **144(Pt 9)**, 2469–2479.
77. Horsburgh,M.J., Clements,M.O., Crossley,H., Ingham,E. and Foster,S.J. (2001) PerR controls oxidative stress resistance and iron storage proteins and is required for virulence in *Staphylococcus aureus*. *Infect. Immun.*, **69**, 3744–3754.
78. Horsburgh,M.J., Ingham,E. and Foster,S.J. (2001) In *Staphylococcus aureus*, Fur is an interactive regulator with PerR, contributes to virulence, and Is necessary for oxidative stress resistance through positive regulation of catalase and iron homeostasis. *J. Bacteriol.*, **183**, 468–475.
79. McDevitt,D., Francois,P., Vaudaux,P. and Foster,T.J. (1995) Identification of the ligand-binding domain of the surface-located fibrinogen receptor (clumping factor) of *Staphylococcus aureus*. *Mol. Microbiol.*, **16**, 895–907.
80. Horsburgh,M.J., Aish,J.L., White,I.J., Shaw,L., Lithgow,J.K. and Foster,S.J. (2002) sigmaB modulates virulence determinant expression and stress resistance: characterization of a functional *rsbU* strain derived from *Staphylococcus aureus* 8325-4. *J. Bacteriol.*, **184**, 5457–5467.
81. Borezee-Durant,E., Hiron,A., Piard,J.C. and Juillard,V. (2009) Dual role of the oligopeptide permease Opp3 during growth of *Staphylococcus aureus* in milk. *Appl. Environ. Microbiol.*, **75**, 3355–3357.
82. Urbanowski,M.L., Stauffer,L.T. and Stauffer,G.V. (2000) The *gcvB* gene encodes a small untranslated RNA involved in expression of the dipeptide and oligopeptide transport systems in *Escherichia coli*. *Mol. Microbiol.*, **37**, 856–868.
83. Sharma,C.M., Darfeuille,F., Plantinga,T.H. and Vogel,J. (2007) A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites. *Genes Dev.*, **21**, 2804–2817.
84. McArthur,S.D., Pulvermacher,S.C. and Stauffer,G.V. (2006) The *Yersinia pestis gcvB* gene encodes two small regulatory RNA molecules. *BMC Microbiol.*, **6**, 52.
85. Moreno,M.S., Schneider,B.L., Maile,R.R., Weyler,W. and Saier,M.H. Jr (2001) Catabolite repression mediated by the CcpA protein in *Bacillus subtilis*: novel modes of regulation revealed by whole-genome analyses. *Mol. Microbiol.*, **39**, 1366–1381.
86. Seidl,K., Goerke,C., Wolz,C., Mack,D., Berger-Bachi,B. and Bischoff,M. (2008) *Staphylococcus aureus* CcpA affects biofilm formation. *Infect. Immun.*, **76**, 2044–2050.
87. Seidl,K., Muller,S., Francois,P., Kriebitzsch,C., Schrenzel,J., Engelmann,S., Bischoff,M. and Berger-Bachi,B. (2009) Effect of a glucose impulse on the CcpA regulon in Staphylococcus aureus. *BMC Microbiol.*, **9**, 95.
88. Heidrich,N., Moll,I. and Brantl,S. (2007) In vitro analysis of the interaction between the small RNA SR1 and its primary target ahrC mRNA. *Nucleic Acids Res.*, **35**, 4331–4346.
89. Preis,H., Eckart,R.A., Gudipati,R.K., Heidrich,N. and Brantl,S. (2009) CodY activates transcription of a small RNA in Bacillus subtilis. *J. Bacteriol.*, **191**, 5446–5457.
90. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.

Downloaded from http://nar.oxfordjournals.org by on September 8, 2010

ARNold: a web tool for the prediction of Rho-independent transcription terminators

Magali Naville[1], Adrien Ghuillot-Gaudeffroy[1] , Antonin Marchais[1] and Daniel Gautheret[1§]

[1]Univ. Paris-Sud 11, CNRS, UMR8621, Institut de Génétique et Microbiologie, Bâtiment 400, F-91405 Orsay Cedex, France

[§]Corresponding author

Email addresses:

MN: magali.naville@igmors.u-psud.fr

DG: daniel.gautheret@igmors.u-psud.fr

Keywords: transcription, Rho-independent termination, Bacteria

Abstract

Rho-independent termination is a major mechanism of transcriptional arrest in bacteria that controls both normal 3' termination and a wide array of regulatory attenuation events. Detecting Rho-independent terminators is an obliged step in the annotation of bacterial operons. Yet, while several efficient algorithms are available for this purpose, there is no freely available web site enabling a rapid scanning of raw genomic sequence for the presence of terminators. Here we implemented such a web server, which combines two published prediction algorithms, Erpin and RNAmotif, and performs nearly as well as more complex procedures while being accessible to the non specialist. The ARNold Web server is available at : http://rna.igmors.u-psud.fr/toolbox/arnold/

Introduction

Bacteria use two alternative mechanisms to terminate transcription: Rho-dependent termination involves the protein factor Rho and occurs at sites that cannot be easily characterized by their structure or sequence[1], while Rho-independent termination involves sequence and structure signals that can be computationally identified in genomic sequences[2,3]. These signals, also known as intrinsic terminators, consist of an 8-10 bp and GC-rich hairpin followed by a stretch of Us in the RNA transcript, which correspond in DNA to an inverted palindrome followed by a poly-T. These two features are sufficient to induce transcriptional arrest. Even though Rho-independent termination is not the dominant mode of termination in all bacterial species[4,5], identifying these signals remains an obligatory step for characterizing transcript 3' ends as well as many "premature" terminators occuring in the 5' part of genes and involved in multiple regulatory mechanisms[6].

Several efficient algorithms have been developed for intrinsic terminator detection. One class uses scoring metrics developed from trusted example sets to distinguish true terminators from other stem-loop[4,5,7]. Another class uses generic motif detection algorithms,

such as Erpin[8], that takes as input a sequence alignment, or RNAmotif[3], that uses an expert-made descriptor to represent sequence, structure and thermodynamic constraints. The later was for instance applied for the characterization of *Escherichia coli* terminators[9]. Yet, although efficient algorithms have been published for several years, no freely accessible online tool is available for finding intrinsic terminators in a raw DNA/RNA sequence. One of the state of the art tools, TranstermHP[5], requires an annotated, Genbank-formatted sequence and restricts search to intergenic regions, which makes it unfit for certain applications, including detection of transcriptional attenuators that occur after gene starts. Other available tools include commercial programs such as Softberry's Findterm (http://www.softberry.com), or require computational skills for operation and are thus out of the reach of many biology labs.

To satisfy the need for a simple, freely accessible terminator detection tool, we developed ARNold, a web interface allowing biologists to rapidly seek Rho-independent terminators in any bacterial sequence, with detection performances on a par with the most recent published algorithms.

Results and Discussion

ARNold finds terminators in DNA or RNA sequences up to 2 Mb in size. The search procedure uses two complementary programs, Erpin and RNAmotif. Erpin[8] is given a structure-annotated alignment of 1200 terminator sequences from *Bacillus subtilis* and *Escherichia coli* as a training set. It builds a lod-score profile from this alignment and seeks high scoring instances of the profile in input sequences. The descriptor for RNAmotif was developed by Lesnik et al.[9] to recognize *E. coli* terminators; however it can in principle be applied to find terminators from any species. Essentially, the descriptor consists of a 4-18 bp helix, a 0-2 nt spacer and a 12 nt T-rich region. RNAmotif matches are scored using the sequence contents of the T-rich region and stability of the stem loop region and an empirical score cutoff is defined to accept or reject matches.

As the two algorithms used by ARNold implement different scoring schemes, we opted to score again all predictions to obtain a homogeneous evaluation of candidates. To this aim, we compute the free energy of predicted structures using RNAfold[10]. This uniform ΔG value is provided as a confidence value for predicted terminators. Overlapping hits located at 5 nt or less from one another on the same strand are clustered and represented by the best-scoring hit.

Table 1. Predictive performances of programs used in ARNold and of TransTermHP[5].

Program	Sensitivity (%)	Specificity (%)*
Erpin	79.1	99.6
RNAmotif	69.2	95.6
ARNold (Erpin U RNAmotif)	87.8	95.3
TransTermHP[5]	90.4	97.5

* Computed as (true negative / true negative * false positives), assuming the 1Mb negative training set is composed of 8700 shuffled 115nt intergenic regions (the average size of intergenic regions in *B. subtilis*).

To evaluate Arnold's prediction performance, we used the same dataset of 425 experimentally verified operons of *B. subtilis* as in other recent studies[4,5]. For sensitivity measures we scanned the whole *B. subtilis* genome and recorded a positive hit only when the prediction overlapped one of the 425 trusted terminators. For specificity measures, we built a 1Mb negative dataset of same dinucleotide composition as the concatenated 425 trusted terminators (100nt around each terminator). Specificity and sensitivity values computed from this data are shown in Table 1.

The specificity and sensitivity of Arnold are comparable, albeit slightly inferior, to those of the reference program TransTermHP[5]. Arnold has a 87.8% sensitivity compared to 90.4% for TransTermHP. Combination of the Erpin and RNAmotif algorithms allows a significant gain (Erpin and RNAmotif used alone have 79.1% and 69.2% sensitivity, respectively) which indicates the two programs tend to detect different classes of terminators. As the Erpin training set was made partly of *B. subtilis* terminators, we also measured Erpin sensitivity

after removing all *B. subtilis* sequences from the training set. Although this amputed the training set by 57%, sensitivity was only mildly reduced from 79.1% to 76.2%.

Arnold and TransTermHP both achieve high specificities at 95.3% and 97.5%, respectively, which amounts to 47 and 25 false positives hits per 1000 intergenic regions of size 115nt (the average size in *B. subtilis*. We noted a sharp drop in the number of TransTermHP detections when input sequences were given in 100 nt chunks instead of large fragments. This effect was not as important with Arnold, indicating that TranstermHP is subject to stronger edge effects. Note that TransTermHP cannot analyze input sequences shorter than 74 nt while Arnold has no lower size limit.

Figure 1. Terminator prediction obtained in a fragment from the *Buchnera aphidicola* genome, showing 33 predicted terminators. Each line contains: starting position; name of program that emitted the prediction or "BOTH" when applicable; strand direction; color-coded terminator sequence (blue: base-pairs, red: apical loop); predicted free energy of terminator hairpin (Kcal/mol).

The ARNold interface is intuitive and user friendly. Input sequences in Fasta or free format are entered into a text area or uploaded as a file. Options enable search strand selection

(plus, minus or both) as well as some control on output format. Runtimes (about 10 sec/Mb using a single 64-bit CPU) are short enough to allow complete scans of bacterial genomes in under a minute. Results are presented in a highly legible format (Fig. 1), indicating, for each prediction, the starting position, strand and stem-loop free energy, as well as the name of the program that generated the prediction and the terminator sequence with colored stem-loop region. Hits predicted by both Erpin and RNAmotif are reported as such. Predictions are listed according to their position in the input sequence, which facilitates the identification of bidirectional terminators, such as the one at position 105864-7 in Figure 1.

References

1. Ciampi MS. Rho-dependent terminators and transcription termination. Microbiology 2006; 152: 2515–28.

2. d'Aubenton Carafa Y, Brody E, Thermes C. Prediction of Rho-independent Escherichia coli transcription terminators. A statistical analysis of their RNA stem-loop structures. J Mol Biol 1990; 216(4):835-58.

3. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res 2001; 29:4724-35.

4. de Hoon MJ, Makita Y, Nakai K, Miyano S. Prediction of transcriptional terminators in Bacillus subtilis and related species. PLoS Comp Biol 2005; 1(3):e25.

5. Kingsford CL, Ayanbule K, Salzberg SL. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biol 2007; 8(2):R22.

6. Naville M, Gautheret D. Transcription attenuation in Bacteria: theme and variations. Brief Funct Genomic Proteomic 2009. 8:482-92.

7. Unniraman S, Prakash R, Nagaraja V. Conserved economics of transcription termination in eubacteria. Nucleic Acids Res 2002; 30:675-84.

8. Gautheret D, Lambert A. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. J Mol Biol 2001; 313:1003-11.

9. Lesnik EA, Sampath R, Levene HB, Henderson TJ, McNeil JA, Ecker DJ. Prediction of rho-independent transcriptional terminators in Escherichia coli. Nucleic Acids Res 2001; 29: 3583-94.

10. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P. Fast Folding and Comparison of RNA Secondary Structures. Monatshefte f. Chemie 1994; 125:167-88.