

Université de Montréal

***Riboswitches* : Le cas des atténuateurs de la  
transcription du type terminateur/antitermineur  
chez les bactéries**

par

Maria de los Angeles Abella

Département de biochimie

Faculté de médecine

Mémoire présenté à la Faculté des études supérieures et postdoctorales  
en vue de l'obtention du grade de maîtrise  
en bio-informatique

Décembre 2011

© Maria de los Angeles Abella, 2011

Université de Montréal  
Faculté des études supérieures et postdoctorales

Ce mémoire intitulé :

*Riboswitches* : Le cas des atténuateurs de la transcription du type  
terminateur/antitermineur chez les bactéries

Présenté par :

Maria de los Angeles Abella

a été évalué par un jury composé des personnes suivantes :

Serguei Chteinberg, président-rapporteur

François Major, directeur de recherche

Daniel Lafontaine, membre du jury

## Résumé

Il est essentiel pour chaque organisme d'avoir la possibilité de réguler ses fonctions afin de permettre sa survie et d'améliorer sa capacité de se reproduire en divers habitats. Avec l'information disponible, il semble que les organismes consacrent une partie assez importante de leur matériel génétique à des fonctions de régulation. On peut envisager que certains mécanismes de régulation ont persisté dans le temps parce qu'ils remplissent bien leurs rôles.

Les premières études sur les procaryotes ont indiqué qu'il y avait peu de mécanismes de régulation exerçant le contrôle des gènes, mais il a été démontré par la suite qu'une variété de ces mécanismes est utilisée pour la régulation de gènes et d'opérons. En particulier, les opérons bactériens impliqués dans la biosynthèse des acides aminés, l'ARNt synthétase, la dégradation des acides aminés, les protéines ribosomales et l'ARN ribosomal font l'objet d'un contrôle par l'atténuation de la transcription. Ce mécanisme d'atténuation de la transcription diffère d'autres mécanismes pour la génération de deux structures différentes de l'ARNm, où l'une de ces structures réprime le gène en aval, et l'autre permet de continuer la transcription/traduction.

Dans le cadre de cette recherche, nous nous sommes intéressé au mécanisme d'atténuation de la transcription chez les procaryotes où aucune molécule ne semble intervenir comme facteur de régulation, en me concentrant sur la régulation des opérons bactériens. Le but principal de ce travail est de présenter une nouvelle méthode de recherche des riborégulateurs qui combine la recherche traditionnelle des riborégulateurs avec la recherche structurale. En incorporant l'étude du repliement de l'ARNm, nous pouvons mieux identifier les atténuateurs répondant à ce type de mécanisme d'atténuation.

Ce mémoire est divisé en quatre chapitres. Le premier chapitre présente une revue de la littérature sur l'ARN et un survol sur les mécanismes de régulation de l'expression génétique chez les procaryotes. Les chapitres 2 et 3 sont consacrés à la méthodologie utilisée dans cette recherche et à l'implémentation du logiciel TA-Search. Enfin, le chapitre 4 expose les conclusions et les applications potentielles de la méthode.

**Mots clefs :** riborégulateur, atténuation de la transcription, ARN, structure secondaire, repliement de l'ARN

## Abstract

It is essential for each organism to have the possibility to regulate its functions to allow its survival and improve its capacity to reproduce in different environments. With the information available, it is apparent that most organisms dedicate an important piece of their genetic material to regulating functions. We could think that certain regulating mechanisms have most likely persisted over time because they fulfilled their roles.

The first prokaryotes studies indicated that there are few regulating mechanisms that take control over genes, but it has been proven that a variety of these mechanisms are used in the regulation of genes and operons. In particular, the bacterial operons involved in the biosynthesis of amino acids, tRNA synthetase, the degradation of amino acids, the ribosomal proteins and RNA ribosomal could be controlled by transcription attenuation. This mechanism of regulation differs from others for the creation of two different structures of the mRNA where one of these structures represses the gene in 3' and the other one allows the transcription/translation to continue.

In this work, I'm interested in the mechanism of transcription attenuation in prokaryotes where no molecule appears to act as a regulatory factor. In particular, I'm interested in the regulation of bacterial operons. The principal goal of this work is to present a new method for detecting riboswitches that combines the traditional research of these elements with the structural research by incorporating the study of mRNA folding.

This thesis is divided into four chapters. Chapter 1 is a review of the literature on RNA and an overview of the regulatory mechanism of gene expression in prokaryotes. Chapter 2 and 3 present the method developed for this work and its implementation in new software, TA-Search. Finally, Chapter 4 is dedicated to providing a discussion and conclusion for this work.

**Keywords:** riboswitch, transcription attenuation, RNA, secondary structure, RNA folding

<b>RÉSUMÉ</b>	<b>iii</b>
<b>ABSTRACT</b>	<b>v</b>
<b>TABLE DE MATIÈRES</b>	<b>vii</b>
<b>LISTE DE FIGURES</b>	<b>x</b>
<b>LISTE DE TABLEAUX</b>	<b>xv</b>
<b>ABBREVIATIONS</b>	<b>xvii</b>
<b>DÉDICACE</b>	<b>xviii</b>
<b>REMERCIEMENTS</b>	<b>xix</b>
<b>Chapitre 1 : Introduction</b>	<b>1</b>
<b>1.1 L'ARN</b>	<b>1</b>
1.1.1 Les acides ribonucléiques	2
1.1.2 La structuration de l'ARN	5
La représentation de la structure secondaire	8
1.1.3 Le modèle de repliement	9
1.1.4 L'analyse et la prédiction des structures secondaires	10
Les méthodes informatiques	10
<b>1.2 Les riborégulateurs et la régulation de l'expression chez les procaryotes</b>	<b>13</b>
1.2.1 Les opérons bactériens	16
1.2.2 La transcription chez les procaryotes	17
1.2.3 La traduction chez les procaryotes	19
1.2.4 L'atténuation de la transcription chez les bactéries	22

1.2.5 Le mécanisme d'atténuation intrinsèque, ou Rho-indépendant	23
1.2.6 Les mécanismes d'atténuation de la transcription de l'opéron <i>trp</i> chez <i>Escherichia coli</i>	26
1.2.7 Les méthodes de prédiction des atténuateurs de la transcription	30
1.2.8 Le but de ce mémoire	31
<b>Chapitre 2 : Étude du repliement des atténuateurs de la transcription</b>	<b>33</b>
2.1 Étude de l'impact de cofacteurs dans le repliement de l'ARN	33
2.2 Les cas du mécanisme d'atténuation de l'opéron <i>trp</i>	35
2.3 La structure secondaire comme élément d'analyse du modèle d'atténuation	44
<b>Chapitre 3 : TA-Search</b>	<b>50</b>
3.1 TA-Search : Recherche des atténuateurs de la transcription	50
3.1.1 Analyse du génome	52
3.1.2 Identification des changements dans les repliements	54
3.2 Recherche de candidats : validation de TA-Search sur des atténuateurs de la transcription déjà identifiés	54
3.3 Recherche des nouveaux candidats sur des génomes complets	59
<b>CHAPITRE 4 : Discussion et conclusion</b>	<b>62</b>
4.1 L'utilisation de la prédiction de structures secondaires	62
4.2 TA-Search	67
4.3 TA-Search et la prédiction chez les opérons <i>leu</i> et <i>his</i>	68
4.4 Conclusion	69
<b>BIBLIOGRAPHIE</b>	<b>70</b>
<b>ANNEXE A</b>	<b>I</b>
<b>ANNEXE B</b>	<b>III</b>

**ANNEXE C**

**XV**

**ANNEXE C**

**XX**

**LICENCE DE REPRODUCTION**

## Liste de Figures

Figure 1.1 : Le dogme central de la biologie permet d'expliquer le processus de l'expression génétique, la conservation et l'utilisation de l'information. Il définit le processus où le contenu de l'ADN est transcrit en ARN et, par la suite, synthétisé en protéine.

Figure 1.2 : Le brin d'ARN est formé à partir de l'alternance de groupes phosphate et des sucres où les bases (la cytosine, la guanine, l'adénine et l'uracile) sont attachées. (Figure adaptée de <http://commons.wikimedia.org/wiki/File:RNA-Nucleobases.svg> )

Figure 1.3 : a) Les bases possèdent trois faces leur permettant de former des ponts hydrogène entre eux. Les appariements sont en général observés sur la face Watson-Crick. (Figure adaptée de <http://fr.wikipedia.org/wiki/Fichier:NucleotideFaces.svg> ). b) Les appariements de type canonique sont les plus stables et les plus fréquemment observés, comme les appariements G:C et A:U. L'autre appariement est G:U, ou Wobble. (Figure adaptée de Varani, 2000)

Figure 1.4 : La structure secondaire de l'ARN<sup>Phe</sup>. a) La séquence de nucléotides, ou structure primaire. b) La séquence de nucléotides et la structure secondaire de ARN<sup>Phe</sup>. c) La structure tertiaire de ARN<sup>Phe</sup>.

Figure 1.5 : Les éléments de structures secondaires de l'ARN. La tige-boucle peut être considérée soit comme un élément de structure, soit comme un ensemble formé d'une tige et d'une boucle terminale.

(Figure adaptée de <http://commons.wikimedia.org/wiki/File:RNAsecondary.svg> et <http://commons.wikimedia.org/wiki/File:PseudoknotFold.svg>)

Figure 1.6 : La structure secondaire peut être représentée : a) en utilisant la notation planaire ou b) la notation parenthésée.

Figure 1.7 : Les riborégulateurs sont des éléments d'ARN dont, en présence d'un effecteur, le repliement change, affectant ainsi le gène auquel ils sont associés.

Figure 1.8 : Les riborégulateurs sont des éléments de régulation de la transcription ou de la traduction chez les procaryotes. Différents types ont été signalés : les riborégulateurs dépendants de la présence de métabolites, ceux qui sont sensibles à la présence d'ARNt chargés et ceux qui interagissent avec des protéines pour modifier le repliement de l'ARN. (Figure adaptée de Naville, 2010)

Figure 1.9 : La structure consensus des opérons bactériens; a) la forme consensus à partir du travail sur les opérons lac et trp; b) la forme consensus de l'opéron trp et les points de régulation possibles.

Figure 1.10 : L'ADN étant déroulé, un des deux brins d'ADN exposés agit comme matrice pour permettre un appariement de bases complémentaires avec les ribonucléotides entrants. De cette façon, la nouvelle chaîne s'allonge d'un nucléotide à la fois, dans la direction 5' → 3'.

(Figure adaptée de [http://en.wikipedia.org/wiki/File:Simple\\_transcription\\_elongation1.svg](http://en.wikipedia.org/wiki/File:Simple_transcription_elongation1.svg) )

Figure 1.11 : La traduction chez les procaryotes commence avec l'assemblage du ribosome à partir des sous-unités 30S et 50S. Une fois le ribosome en place, la synthèse de la protéine commence lorsqu'un aminoacyl ARNt pénètre dans le site A du ribosome assemblé.

Figure 1.12 : La terminaison de la synthèse d'une protéine se fait quand le ribosome trouve un des trois codons d'arrêt (UAA, UAG, ou UGA). En même temps, la présence de facteurs de libération de la chaîne permet de libérer la chaîne polypeptidique, puis l'ARNm.

Figure 1.13 : Un segment d'ADN répété inversé pourra se replier en forme de tige-boucle. a) Quatre éléments définissent la tige-boucle décrite par de d'Aubenton Carafa. b) Des années plus tard, Lesnik revisite la définition de d'Aubenton Carafa et redéfinit le segment postérieur à la tige; le segment de poly-T.

Figure 1.14 : La séquence du gène *trpL* et le segment d'atténuation de l'opéron *trp*. Quatre segments sont identifiés; ils peuvent former divers appariements entre eux et diverses conformations en dépendant du passage et du rythme de synthèse de l'ARN polymérase et l'action du ribosome.

Figure 1.15 : Le mécanisme d'atténuation de la transcription de l'opéron *trp* chez *E. coli* peut suivre deux scénarios possibles selon la concentration de *trp*: soit la formation du terminateur à haute concentration de *trp*, soit la formation d'une tige-boucle antiterminatrice à basse concentration de *trp*. (Figure adaptée de Elf et Ehrenberg, 2005)

Figure 1.16 : Modèle consensus des structures proposées par le modèle d'atténuation de la transcription. La première structure correspond à la structure d'arrêt de la transcription, ou structure terminatrice. La deuxième correspond à l'antiterminateur, ou structure qui permet la poursuite de la transcription/traduction.

Figure 2.1 : Pour représenter l'interaction d'un cofacteur, on utilise le masque pour indiquer les positions de fixation sur l'ARNm qui ne doivent pas former de paires de bases. Trois types de masques sont considérés: le type 1 bloque 15 nucléotides à partir de la position 5' en avançant sur la séquence vers l'extrémité 3'; le type 2 bloque à partir d'un premier segment de 15 nucléotides, il ajoute trois nucléotides pour bloquer de façon progressive la séquence et le type 3 bloque la séquence de trois nucléotides à la fois.

Figure 2.2 : La méthode appliquée en utilisant le masque de type 1; chaque ligne représente le repliement MFE pour chaque position de fixation du cofacteur. La boîte rouge indique où le repliement a changé et la boîte bleue, la position du cofacteur qui provoque le changement. Dans ce cas, les positions 24 à 35 codent deux codons trp.

Figure 2.3 : La méthode appliquée en utilisant le masque de type 2; chaque ligne représente le repliement MFE pour chaque position de fixation du ligand. La boîte rouge indique où le repliement a changé et la boîte bleue, la position du ligand qui provoque le changement. Dans ce cas, les positions 24 à 35 codent deux codons trp.

Figure 2.4 : La méthode appliquée en utilisant le masque de type 3; chaque ligne représente le repliement MFE pour chaque position de fixation du ligand. En ligne 5, en rouge, on observe un changement de repliement par rapport aux premiers repliements; en ligne 9, en bleu, le repliement change encore une fois et on observe la formation de la tige-boucle terminatrice.

Figure 2.5 : Le repliement de l'ARNm de l'atténuateur de l'opéron trp dans *E.coli* avec les quatre logiciels. Quand le masque de type 2 se trouve sur les deux codons régulateurs trp, le repliement présente un changement tel qu'observé à l'option b et prédit par MC-Fold. Les options a, c, et d prédites avec les autres logiciels ne montrent pas le résultat attendu, ni le changement du repliement. Dans chaque figure, la tige-boucle terminatrice est indiquée en rouge et la tige-boucle antiterminatrice en bleu, telle que présentée dans le modèle à deux structures.

Figure 2.6 : Le repliement par MC-Fold de la séquence d'*E.coli* (NC\_000913, segment 4118381 :418500c) ne montre pas les changements de repliement du modèle, bien que les éléments qui caractérisent l'atténuateur soient dans la séquence (les codons trp régulateurs, la tige-boucle terminatrice et les distances entre les segments qui composent l'atténuateur).

Figure 3.1 : Le modèle de recherche des atténuateurs commence à partir d'un fichier descripteur qu'alimente l'analyse du génome de TA-Search. TA-Search est divisé principalement en deux analyses: celle du génome et celle du repliement.

Figure 3.2 : Les éléments qui sont recherchés par TA-Search sont les codons d'initiation et de stop sur le même cadre de lecture, le segment poly-T, les codons régulateurs dans l'opéron leader, la distance entre initiation-stop (rouge), la distance entre le codon d'initiation et le dernier élément de la poly-T (bleu), la distance entre le dernier élément de la poly-T et le gène le plus proche en 3' (vert).

Figure 3.3 : Les candidats de TA-Search peuvent être décrits selon sa position par rapport aux gènes, une partie du candidat est localisé sur un gène codant et l'autre partie sur une région non codante de l'ARNm juste avant un gène. D'autre part, le candidat est localisé dans une région non codante de l'ARNm avant un gène.

Figure 3.4 : À partir des résultats de Search-MFE, a) dans le candidat, final le changement du repliement se fait quand le masque est sur les deux codons trp et qu'après le passage du masque sur ces codons, le repliement change pour présenter une très forte tige-boucle similaire à la tige-boucle terminatrice du modèle. b) En prenant un autre candidat résultant de l'analyse du génome, on remarque qu'aucun changement n'apparaît dans le repliement avec le masque. Dans ce cas, le candidat est rejeté.

## Liste des tableaux

Tableau 2.1 : L'analyse du repliement de l'opéron *trp*, en quatre génomes avec quatre logiciels de prédiction utilisant les trois masques, permet d'observer si les logiciels sont capables de prédire le changement du repliement du modèle consensus de l'atténuation. Si « + », le changement a été aperçu dans le repliement MFE ; si « - », le modèle n'a pas été aperçu dans le repliement; si « ~ », le changement a été aperçu dans l'ensemble de sous-optimaux.

Tableau 2.2 : Liste de génomes bactériens où le mécanisme a été étudié ou prédit auparavant. (Vitreschak et Lyubetsky, 2004).

Tableau 2.3 : Liste des opérons dont les séquences se trouvent annotées dans Genbank. Les opérons choisis sont *trp*, *phe*, *thr*, *ilv*, *leu* et *his*.

Tableau 2.4 : Pour chaque opéron dans chaque génome, selon le tableau 2.3, MC-Fold a été lancé en considérant les trois masques. Après avoir récolté les résultats du repliement en utilisant Search-MFE, annexe A, ils reflètent si le changement de repliement a été observé selon le modèle. Si « + », le modèle a été aperçu dans le repliement MFE ; si « - », le modèle n'a pas été aperçu dans le repliement ; si « ~ », le modèle a été aperçu dans les repliements sous-optimaux.

Tableau 3.1 : TA-Search a été lancé avec les paramètres énumérés dans le fichier descripteur sur chaque génome sur les deux brins,  $5' \rightarrow 3'$  et  $3' \rightarrow 5'$ .

Tableau 3.2 : Une fois changée la définition du motif régulateur de l'opéron *thr*, l'analyse de génome de TA-Search a donné les candidats pour l'opéron *thr* dans tous les génomes.

Tableau 3.3 : Le nombre de candidats proposés par TA-Search pour les quatre génomes analysés.

Tableau 3.4 : TA-Search a été lancé sur quatre génomes, où l'annotation est partielle ou hypothétique. En utilisant les mêmes paramètres que dans la section 2.2, après l'analyse du génome, le logiciel est arrivé à proposer des candidats.

Tableau 3.5 : TA-Search a été lancé sur quatre génomes, où l'annotation est partielle ou hypothétique. En utilisant les mêmes paramètres que dans la section 2.2, après l'analyse du repliement, le logiciel est arrivé à proposer des candidats.

Tableau 3.6 : Dans le cas de thr et ilv, les paramètres ont été changés pour améliorer la prédiction. Le plus grand changement est lié au motif qui identifie les codons régulateurs dans chaque cas.

## Abréviations

A	Adénine
C	Cytosine
G	Guanine
U	Uridine
T	Thymine
Trp	Tryptophane
Leu	Leucine
Phe	Phénylalanine
His	Histidine
ILV	Isoleucine-Leucine-Valine
THR	Thréonine
TrpL	Trp Leader
PheL	Phe Leader
TEC	<i>Transcription Elongation Complex</i> (Complexe d'élongation de la transcription)
ADN	Acide désoxyribonucléique
ARN	Acide ribonucléique
ARNm	ARN messenger
ARNt	ARN de transfert
ARNp	ARN polymérase
MFE	Minimum Free Energie

*À mon père, Rogelio*

*À mon mari, Sergio*

*À mon fils, Martin*

*Qui ont encouragé mes rêves le plus fous.*

*A mi papá, Rogelio*

*A mi marido, Sergio*

*A mi hijo, Martin*

*Quienes siempre apoyaron mis sueños locos*

## Remerciements

Je veux remercier mon directeur de recherche, D<sup>r</sup> François Major, pour ses conseils et son support tout au long de ce travail. Je ne veux pas oublier les membres de son laboratoire : Paul Dallaire, Nicolas Scott, Karine St-Onge, Véronique Lisi, Cédric Saule et Marc-Frédéric Blanchet, qui m'ont aidée à comprendre pourquoi la bio-informatique et la recherche me passionnent.

Je tiens à remercier également M<sup>me</sup> Éline Meunier, dont le soutien constant et les judicieux conseils m'ont été d'un grand secours.

Finalement, je remercie le D<sup>r</sup> Serguei Chteinberg et le D<sup>r</sup> Daniel Lafontaine d'avoir accepté d'être membres du jury d'évaluation de ce mémoire.

# Chapitre 1 : Introduction

## 1.1 L'ARN

Pendant des années, l'acide ribonucléique (ARN) a été considéré par les chercheurs comme un simple intermédiaire de l'information génétique, alors qu'il est un élément fondamental de la vie.

En biologie moléculaire, le modèle schématique de la conservation et de l'utilisation de l'information génétique est appelé le « dogme central de la biologie ». Ce modèle établit la liaison qu'il existe entre le matériel génétique contenu dans la cellule, l'acide désoxyribonucléique (ADN) et les protéines que cette cellule synthétise (Figure 1.1) (Crick, 1970; Lewin, 1997).

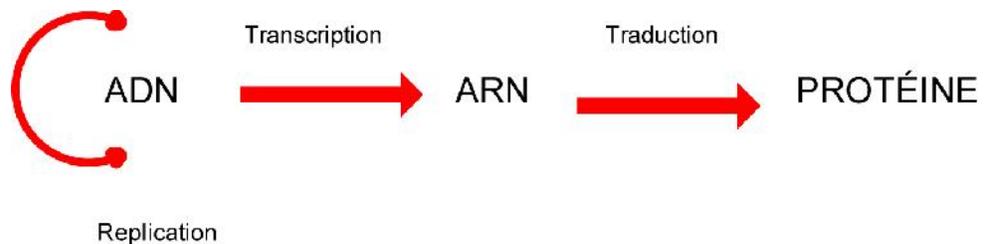


Figure 1.1 : Le dogme central de la biologie permet d'expliquer le processus de l'expression génétique, de la conservation et de l'utilisation de l'information. Il définit le processus par lequel le contenu de l'ADN est transcrit en ARN et, par la suite, synthétisé en protéine.

Le passage du gène à la protéine se fait en deux étapes : tout d'abord, le segment d'ADN, correspondant au gène est transcrit sur un brin d'acide ribonucléique que l'on appelle l'« ARN messager » (ARNm); il s'agit de la transcription. Par la suite, le brin d'ARNm est traité pour donner la séquence correspondant à une protéine; il s'agit de la traduction (Voet et Voet, 2005).

Il y a plus d'ARN non codants qui codent pour des protéines. Bien que non démontrés clairement, ils doivent assurer une grande variété de fonctions au niveau cellulaire. Les ARN sont capables de catalyser diverses réactions chimiques, de corriger les problèmes de traduction ou de réguler l'expression de gènes. Plusieurs découvertes ont laissé supposer que l'ARN aurait été à l'origine même de la vie sur Terre (Atkins, Gesteland et al., 2011).

### 1.1.1 Les acides ribonucléiques

L'ARN est un polymère linéaire composé de nucléotides. Le nucléotide est composé d'un pentose, d'un groupe phosphate (un acide phosphorique) et d'une base azotée parmi l'adénine, l'uracile, la guanine ou la cytosine. L'alternance des phosphates et des sucres produit le squelette des acides nucléiques sur lequel s'attachent les bases azotées. Le polymère ainsi formé possède des extrémités différentes relatives à la géométrie des sucres et nommées 5' et 3' (Figure 1.2) (Alberts, Johnson et al., 2004; Bolsover, Hyams et al., 2006).

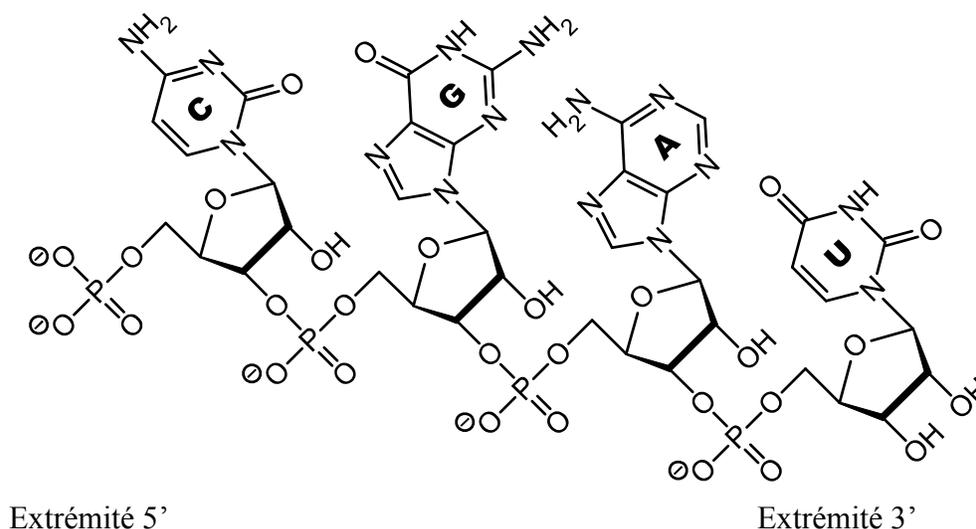


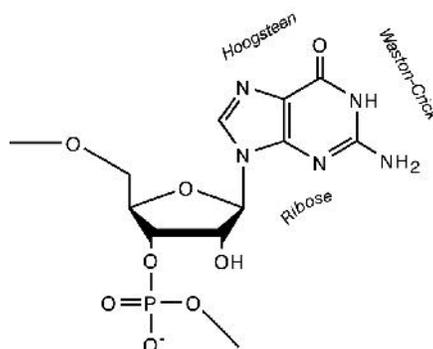
Figure 1.2 : Le brin d'ARN est formé à partir de l'alternance de groupes phosphate et de sucres, où les bases (la cytosine, la guanine, l'adénine et l'uracile) sont attachées. (Figure adaptée de <http://commons.wikimedia.org/wiki/File:RNA-Nucleobases.svg>)

La présence d'un groupe hydroxyle sur la position 2' du ribose influence la conformation du cycle furanose du ribose. Ce cycle à cinq atomes n'est pas plan, ce qui conduit à deux conformères principaux du sucre, appelés C2'-endo et C3'-endo.

La structure chimique des bases azotées permet de distinguer deux groupes : les purines, constituées de l'adénine et de la guanine; et les pyrimidines, constituées de la cytosine et de l'uracile. Une caractéristique des bases azotées est qu'elles possèdent trois faces: celle de Watson-Crick (WC), la face Hoogsteen (H) et la face sucre (S), qui permettent la formation de liaisons de type pont hydrogène (Figure 1.3 a). Les deux appariements se forment entre l'adénine et l'uracile et entre la cytosine et la guanine. Cet appariement, impliquant la face WC de deux bases azotées, se fait par des ponts hydrogène, et son nombre dépend du nombre d'atomes accepteurs/donneurs de chaque base (Figure 1.3 b). Un autre type d'appariement, dit « Wobble », vu dans les structures d'ARN se forme entre la guanine et l'uridine (Figure 1.3 c). Les appariements de ce type sont appelés canoniques (Leontis et Westhof, 2001; Lee et Gutell, 2004; Major et Thibault, 2007).

D'après l'analyse des structures dans les structures tridimensionnelles d'ARN résolues par cristallographie ou par résonance magnétique nucléaire (RMN), certains appariements ne correspondent pas aux appariements canoniques. Ces appariements non canoniques impliquent par exemple les faces Hoogsteen-Hoogsteen ou Watson-Hoogsteen. De la même façon, il a été observé également des interactions base-ribose, base-phosphate ou base-empilement (Fritsch et Westhof, 2000; Lemieux et Major, 2002; Leontis et Westhof, 2003; Abreu-Goodger et Merino, 2005; Mukherjee, Bansal et al., 2006; Olson, Esquerra et al., 2009; Zirbel, Sponer et al., 2009).

a)



b)

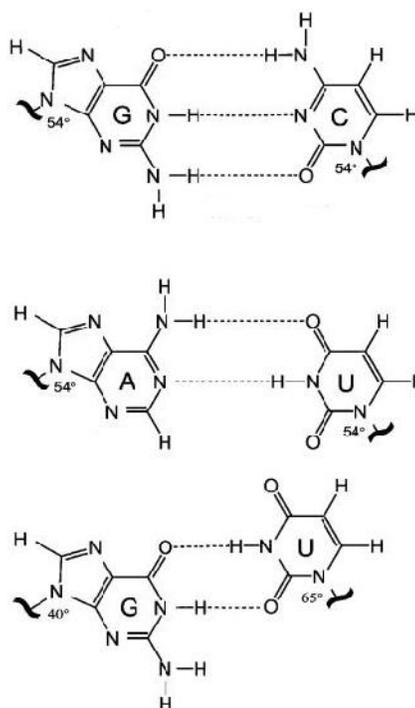


Figure 1.3 : a) Les bases possèdent trois faces leur permettant de former des ponts hydrogène entre eux. Les appariements sont en général observés sur la face Watson-Crick. (Figure adaptée de <http://fr.wikipedia.org/wiki/Fichier:NucleotideFaces.svg>). b) Les appariements de type canonique sont les plus stables et les plus fréquemment observés, comme les appariements G:C et A:U. Un autre appariement est l'appariement G:U, ou Wobble. (Figure adaptée de Varani, 2000) <sup>1</sup>

<sup>1</sup> Reproduit avec la permission de Macmillan Publishers ltd : EMBO reports, Varani, G. et McClain, W. « The G-U wobble base pair », (2000) July 17; 1(1): 18–23. doi: 10.1093/embo-reports/kvd001 , copyright 2000.

### 1.1.2 La structuration de l'ARN

L'ARN comporte trois niveaux de structure; la structure primaire, la structure secondaire et la structure tertiaire (Mount, 2004; Voet et Voet, 2005).

La structure primaire désigne la séquence de nucléotides (Figure 1.4 a).

La structure secondaire est la représentation planaire des interactions d'appariement et l'empilement de bases formées entre les nucléotides de la structure primaire (Figure 1.4 b). L'ARN se replie pour former différents éléments qui se trouvent de façon récurrente dans les repliements d'ARN. L'ensemble des appariements permet de définir des régions appariées, les tiges, et des régions non appariées, les boucles. De plus longues séquences d'ARN permettent d'obtenir divers éléments de structure en combinant les tiges et les boucles, la tige-boucle, les divers types de boucles (interne, terminale, multiple), de renflements et le pseudoknot qui résulte de l'interaction à longue distance entre la boucle et une région non appariée à l'extérieur de la tige à laquelle elle appartient (Figure 1.5) (Brion et Westhof, 1997).

La structure tertiaire est représentée par les coordonnées atomiques de la séquence dans l'espace. Ce type de structure provient d'interactions des éléments de structures secondaires pour former des interactions tertiaires (Figure 1.4 c).

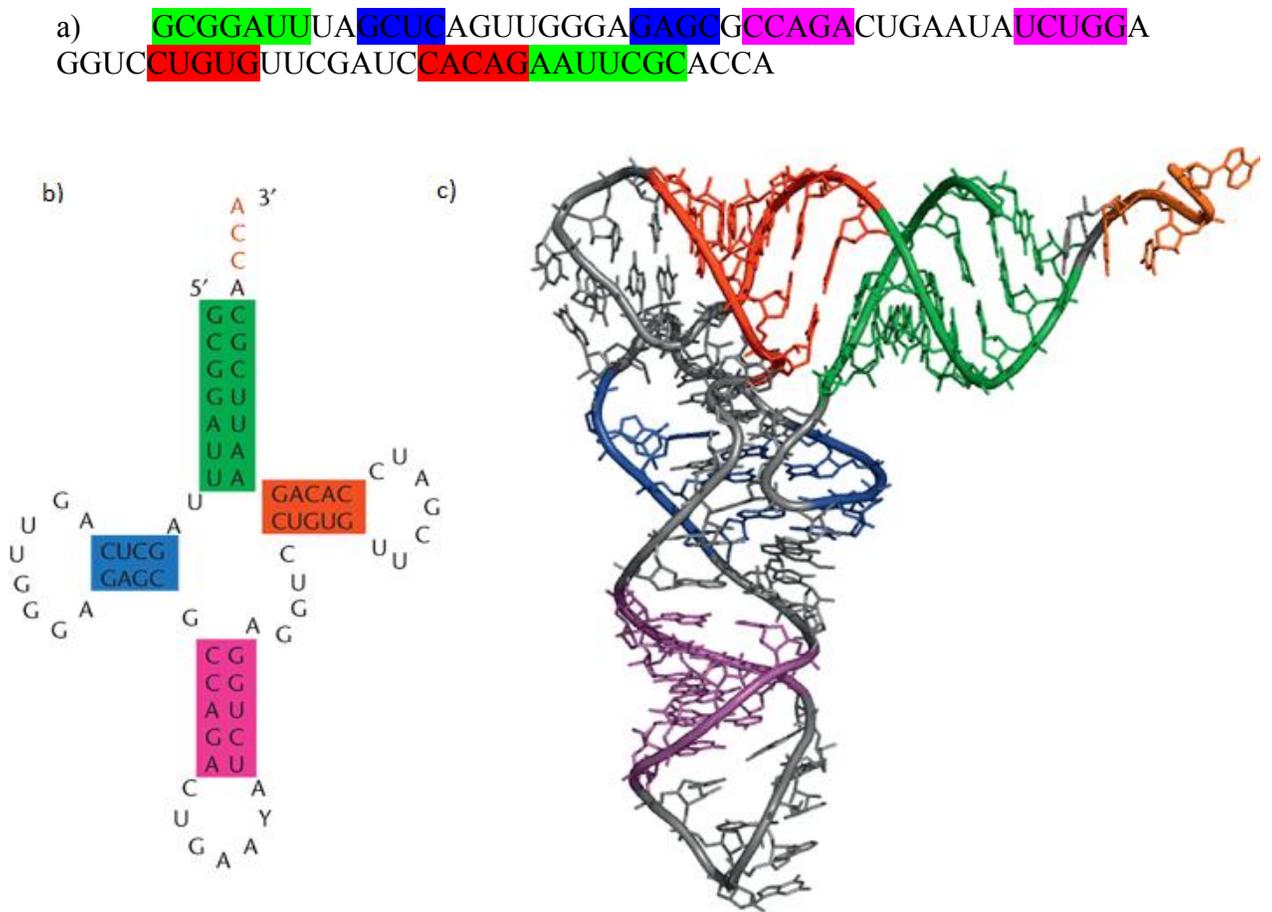


Figure 1.4 : La structure secondaire de l'ARNt<sup>Phe</sup>. a) La séquence de nucléotides, ou structure primaire.  
 b) La séquence de nucléotides et la structure secondaire de ARNt<sup>Phe</sup>, c) la structure tertiaire de ARNt<sup>Phe</sup>.

Dans ce travail, nous nous pencherons sur l'étude de la structure secondaire.

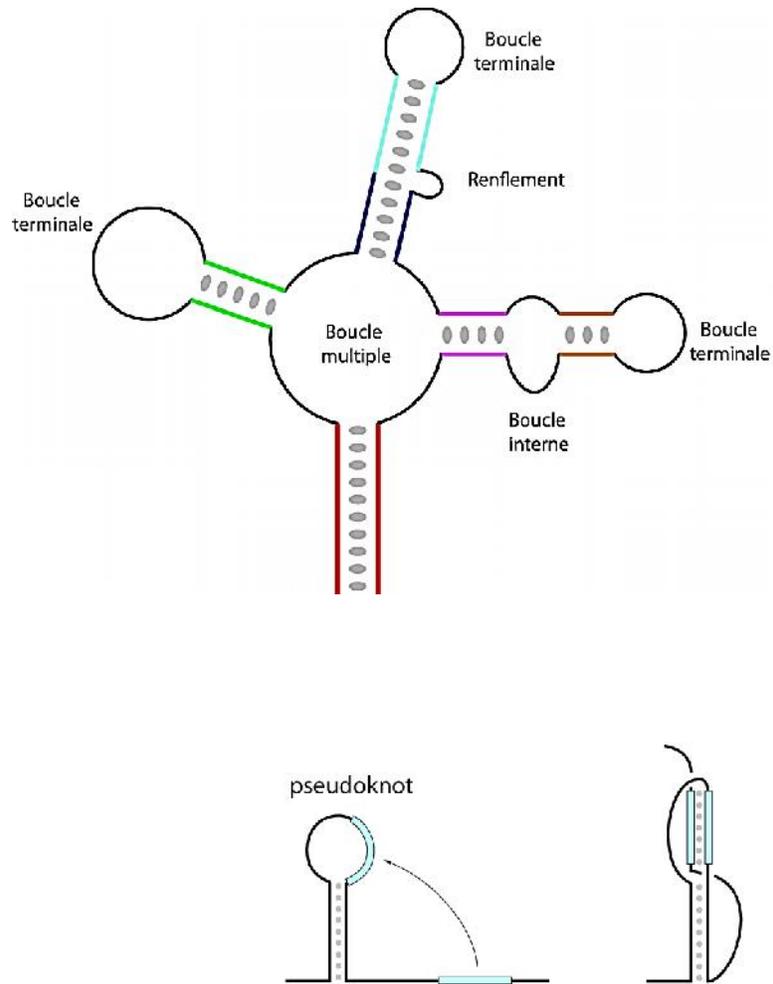


Figure 1.5 : Les éléments de structures secondaires de l'ARN. La tige-boucle peut être considérée soit comme un élément de structure, soit comme un ensemble formé par une tige et une boucle terminale. (Figure adaptée de <http://commons.wikimedia.org/wiki/File:RNAsecondary.svg> et <http://commons.wikimedia.org/wiki/File:PseudoknotFold.svg>)

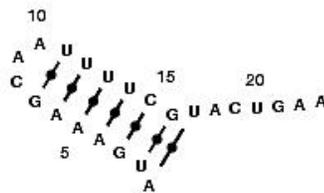
## La représentation de la structure secondaire

Les structures secondaires d'ARN peuvent être représentées de diverses façons. Les notations planaire, en matrice de points, en matrice circulaire ou parenthésée sont les représentations les plus utilisées. Dans le cadre de ce mémoire, nous nous pencherons plus particulièrement l'utilisation des notations planaire et parenthésée (Mount, 2004).

La notation planaire est en deux dimensions. Chaque résidu y est représenté par un point et le squelette (*backbone*), et les ponts hydrogène, par des lignes entre les points (Figure 1.6 a).

Dans la notation parenthésée, les appariements sont représentés à l'aide de parenthèses balancées placées sous la séquence. À chaque nucléotide de la séquence correspond un caractère : une parenthèse ouvrante, une parenthèse fermante ou un point. Ainsi, un point représente une position non appariée, alors que les parenthèses correspondent à des bases appariées. Ce type de notation est la plus utilisée pour représenter la structure secondaire (Figure 1.6 b).

a)



b)

```
AUGAAAGCAAUUUUCGUACUGAA
((( ((( (. . . ) ) ) ) ) ) . . . . .
```

Figure 1.6 : La structure secondaire peut être représentée a) en utilisant la notation planaire ou b) la notation parenthésée.

### 1.1.3 Le modèle de repliement

Le repliement de l'ARN a été perçu ces dernières années comme un processus hiérarchique. Les études de la transcription chez les bactéries au cours des années 1990 ont démontré que le processus de repliement est influencé par divers éléments de l'environnement, par exemple des pauses transcriptionnelles ou des métabolites.

Le repliement hiérarchique considère que l'ARN commence à se replier à partir de la structure primaire et qu'il adopte par la suite une structure secondaire avant finir le repliement dans sa structure tertiaire. Dans sa forme primaire, l'ARN a le plus haut niveau d'énergie libre, mais quand la structure secondaire se forme, le processus libère de l'énergie en stabilisant la structure résultante. Le brin d'ARN se replie sur lui-même, en rapprochant les bases qui forment des appariements de type pont hydrogène et en fermant des boucles terminales pour donner forme aux doubles hélices d'ARN. Une fois que les éléments de structures secondaires sont assemblés, les interactions tertiaires se forment pour produire la structure tertiaire. La structure tertiaire ajoute des interactions entre les éléments de la structure secondaire et contribue de façon moins significative à l'énergie libre de la structure. Dans ce modèle, l'idée de base est que l'ARN dispose de toute l'information nécessaire pour se replier par lui-même (Brion et Westhof, 1997; Tinocco Jr et Bustamante, 1999; Xayaphoummine et Isambert, 2007).

En considérant l'environnement, le repliement de l'ARN est influencé par les facteurs extérieurs à la séquence d'ARN qui modifient son repliement. Lorsque l'ADN est transcrit, le brin d'ARN naissant peut déjà commencer à former sa structure, mais des éléments de structure comme une protéine, des métabolites, de l'ADN ou des pauses provoquées par l'ARN polymérase peuvent influencer la structure qu'adoptera le reste de la chaîne une fois transcrit. Ainsi, si une partie de la structure est déjà produite, les nucléotides qu'elle contient ne sont plus disponibles pour s'apparier avec la séquence qui est en train d'être synthétisée. Dans ce cas, le repliement doit considérer les étapes de production de la chaîne d'ARN en incluant

tous les intervenants pour en comprendre la structure (Keller et Calvo, 1979; Gardner, 1982; Henkin et Yanofsky, 2002; Sosnick et Pan, 2003; Pan et Sosnick, 2006).

#### **1.1.4 L'analyse et la prédiction des structures secondaires**

Un ARN de longueur suffisante peut, en principe, adopter un grand nombre de conformations différentes. Le repliement le plus stable est celui qui correspond au repliement de plus basse énergie libre identifié, ou énergie libre minimale (MFE).

Afin de déterminer ou de prédire la structure secondaire des ARN, deux méthodes sont proposées actuellement : les méthodes expérimentales et les méthodes informatiques. Les méthodes expérimentales permettent de déterminer la structure secondaire en utilisant les méthodes chimiques et enzymatiques. D'autres méthodes permettent également, à partir de la structure tertiaire, d'obtenir la structure secondaire telle que la cristallographie, la résonance magnétique nucléaire ou, récemment, le transfert d'énergie par résonance de type Förster (FRET) (Egli et Pallan, 2007; Hall, 2008; Baird, Kulshina et al., 2010; Sigel et Gallo, 2010; Tinocco, Chen *et al.*, 2010; Klostermeier, 2011).

Les méthodes informatiques sont utilisées essentiellement pour la prédiction des structures secondaires. Dans ce cas, deux approches sont employées: soit l'approche comparative et l'approche ab-initio. Dans le cadre de ce mémoire, nous nous attarderons à ces deux méthodes.

##### **Les méthodes informatiques**

L'approche comparative infère des structures en comparant l'évolution de plusieurs séquences d'ARN homologues à la séquence d'intérêt. Par contre, l'approche ab-initio est basée sur le calcul de l'énergie minimale de la structure la plus stable dérivée directement de la séquence d'ARN (Xiong, 2006).

### *L'approche comparative*

L'approche comparative se base systématiquement sur l'alignement des séquences à analyser. L'alignement de séquences consiste à établir une correspondance maximale entre les éléments des séquences, et à relever les ressemblances, soit la similarité, l'identité ou l'homologie. Quand on compare les ARN de plusieurs espèces qui remplissent la même fonction, on observe une conservation importante de la structure secondaire. Cette approche est de plus en plus employée à des fins prédictives, par exemple pour déterminer si les séquences partagent une structure commune ou pour déterminer si elles sont des séquences homologues.

Les algorithmes utilisant la méthode comparative peuvent être divisés en deux catégories, selon le type de données utilisé: la première catégorie se base sur des séquences alignées, et la seconde, sur des séquences non alignées.

Dans la prédiction, en utilisant des données préalablement alignées, des programmes comme R-Coffee (Hofacker, Fekete et al., 2002), MARNA (Siebert et Backofen, 2005) ou Clustal (Thompson, Higgins et al., 1994) sont utilisés pour générer l'alignement des séquences d'intérêt avec un ensemble de séquences homologues. Par la suite, cet alignement pourra être analysé, par exemple, par RNAalifold de la suite Vienna (Hofacker, Fontana et al., 1994), pour déduire la structure de minimum énergie de l'ensemble de séquences.

Dans le cas des données non alignées, ce type d'algorithme réalise l'alignement des séquences et infère la structure consensus. Des exemples de cette implémentation sont les applications FoldalignM (Torarinsson, Havgaard et al., 2007), Dynalign (Mathews et Turner, 2002) ou Infernal (Nawrocki, Kolbe et al., 2009).

### *L'approche ab-initio*

Cette approche fait des prédictions structurales basées sur la séquence d'ARN elle-même. Ainsi, ces algorithmes peuvent être conçus pour rechercher les structures stables de l'ARN avec le minimum d'énergie libre.

La stabilité de la structure secondaire de l'ARN est mesurée par la quantité d'énergie libre libérée ou nécessaire pour former les paires de bases. Les énergies libres sont additives; de cette façon, l'énergie libre totale d'une structure secondaire peut s'obtenir en ajoutant toutes les énergies du système. Avec une énergie plus négative, la formation de la structure est plus probable parce que plus d'énergie stockée est libérée. Cette méthode fonctionne par la recherche de toutes les paires de bases possibles une première fois, puis en calculant l'énergie libre de structures secondaires potentielles. S'il existe plusieurs structures secondaires, cette méthode trouve la conformation avec l'énergie minimale, c'est-à-dire la conformation énergétiquement la plus favorable.

En recherchant la conformation de plus basse énergie, toutes les paires de bases doivent être examinées. Il existe plusieurs méthodes pour accomplir cette tâche à partir d'une séquence d'acides nucléiques; par exemple la méthode de la matrice de points, la programmation dynamique, l'utilisation de fonctions de partition ou une combinaison de ces méthodes (Mathews, 2006).

Divers logiciels ont implémenté une ou plusieurs de ces approches. Les plus intéressants d'entre eux pour ce travail sont MFold (Zucker et Stiegler, 1981), Sfold (Ding et Lawrence, 2003), RNAsubopt de la suite Vienna (Hofacker et Standler, 2006) et MC-Fold (Parisien et Major, 2008). Tous ces logiciels permettent d'obtenir un repliement de plus basse énergie (MFE) et, de plus, un ensemble de structures secondaires qui forme une collection de conformations possibles de la molécule.

## **1.2 Les riborégulateurs et la régulation de l'expression chez les procaryotes**

Chez les procaryotes, il existe divers mécanismes de régulation de l'expression des gènes extrêmement efficaces pour activer ou inhiber l'expression de certains gènes ou des opérons selon les besoins de la cellule en produits de ces gènes (Browning et Busby, 2004; Ishihama, 2010; Turnbough, 2011).

Un de ces mécanismes montre que l'ARN est capable de réguler par activation ou inhibition l'expression d'un gène. Dans ce cas, un élément d'ARN, le riborégulateur, change son repliement de façon très marquée en présence d'un effecteur, soit un métabolite ou un élément extérieur qui, en interagissant avec le riborégulateur, lui permet de réorganiser sa structure en permettant ainsi la régulation du gène auquel cet élément est associé (Figure 1.7).

En général, les riborégulateurs sont décrits comme des séquences d'ARN situées dans des régions non codantes de certains ARNm pouvant se lier à un effecteur qui provoque la réorganisation de sa structure. Ils sont composés de deux domaines fonctionnels: un aptamère et la plateforme d'expression. L'aptamère est une structure formée en général par des tiges ou tiges-boucles assemblées autour d'une boucle multiple et il semble être le domaine de reconnaissance de l'effecteur. Immédiatement après l'aptamère, on trouve la plateforme d'expression qui est composée d'une séquence sans repliement défini. A priori, la fonction de cette plateforme est de moduler l'expression du gène en 3' du site de liaison entre l'effecteur et l'aptamère (Vitreschak, Rodionov et al., 2004; Winkler et Breaker, 2005; Lemay et Lafontaine, 2006; Blouin et Lafontaine, 2009; Sergarov, 2009; Baird, Kulshina et al., 2010; Bastet, Dube et al., 2011).

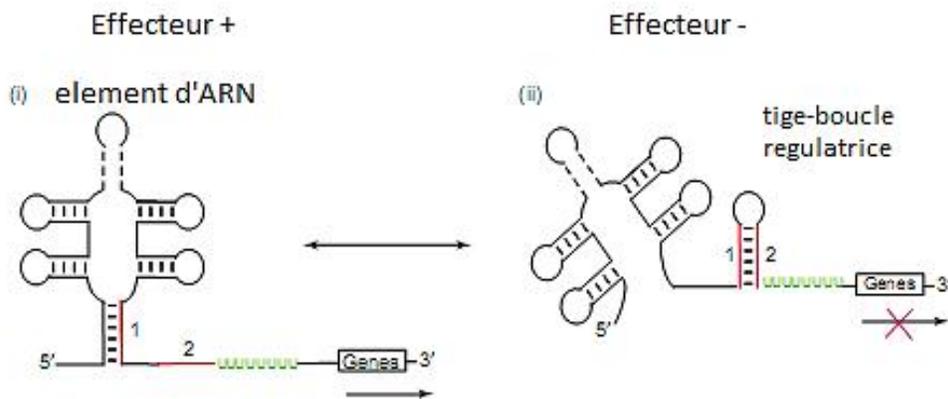


Figure 1.7 : Les riborégulateurs sont des éléments d'ARN dont, en présence d'un effecteur, le repliement change, affectant ainsi le gène auquel ils sont associés.

Pendant la dernière décennie, plusieurs travaux ont cherché à différencier les riborégulateurs des autres mécanismes similaires de régulation comme le T-Box, l'atténuation de la transcription par une protéine ou l'atténuation par interaction avec la tige-boucle terminatrice. Mais dans tous ces cas, ils partagent la même fonction, soit le contrôle de l'expression génique par un changement dans le repliement de la séquence d'ARNm en aval d'un gène (Figure 1.8) (Du et Babitzke, 2000; Means, Wolf et al., 2007; McGraw et Babitzke, 2008; Montagne et Batey, 2008; Breaker, 2010; Green, Grundy et al., 2010; Naville et Gautheret, 2010; Wachter, 2010; Haller, Souliere et al., 2011).

Dans le contexte de ce travail, nous nous intéressons au modèle du mécanisme d'atténuation de la transcription chez les bactéries où l'effecteur, connu comme atténuateur, est la concentration d'un acide aminé. Ce mécanisme a été étudié sur l'opéron *trp* chez *E.coli* par le groupe de Charles Yanofsky dans les années 1970 et 1980, qui a proposé en 1977 un modèle considéré comme le modèle consensus à utiliser dans l'étude de ce mécanisme chez les opérons bactériens (Lee et Yanofsky, 1977; Oxender et Yanofsky, 1979; Yanofsky, 1981; Kolter et Yanofsky, 1982; Stroynowski et Yanofsky, 1983; Kuroda et Yanofsky, 1984; Landick et Yanofsky, 1984; Roesser et Yanofsky, 1991; Yanofsky, 2004).

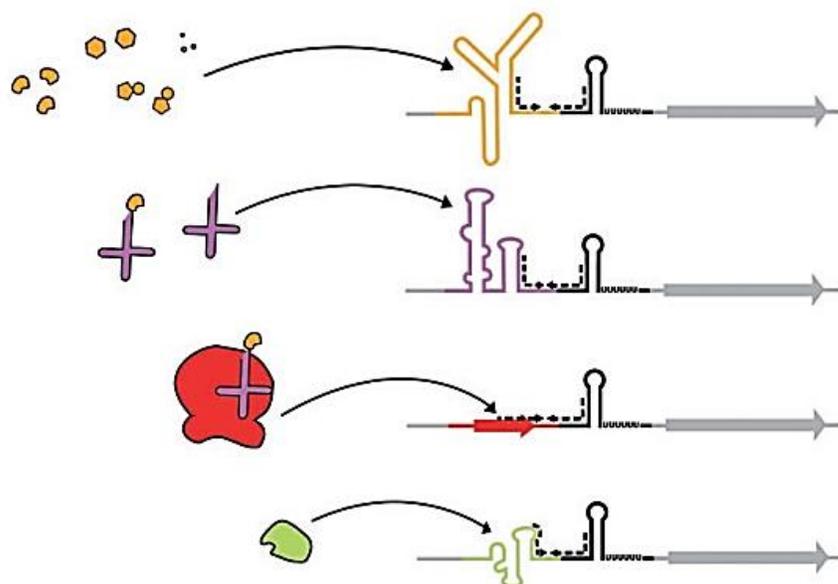


Figure 1.8 : Les riborégulateurs sont des éléments de régulation de la transcription ou de la traduction chez les procaryotes. Différents types ont été signalés : les riborégulateurs dépendants de la présence de métabolites, ceux qui sont sensibles à la présence d'ARNt chargés et ceux qui interagissent avec des protéines pour modifier le repliement de l'ARN (Figure adaptée de Naville 2010).<sup>2</sup>

Dans la suite de ce mémoire, nous proposons de traiter les atténuateurs comme des riborégulateurs, où le contexte principal n'est pas la présence d'une molécule, d'un métabolite ou la concentration de ces éléments, sinon la présence d'un changement important dans le repliement de la séquence d'ARNm qui mène à la régulation du gène.

Afin de faciliter la compréhension de ce mécanisme d'atténuation, ce chapitre propose une révision des concepts de transcription/traduction chez les procaryotes, puis du mécanisme d'atténuation de la transcription chez les bactéries.

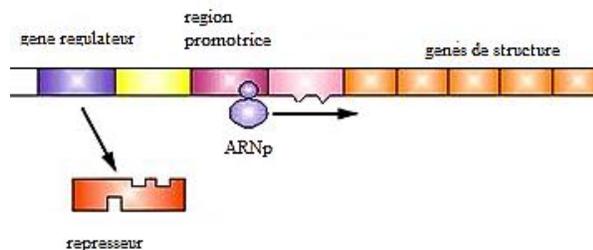
<sup>2</sup> Reproduit avec la permission d'Oxford University Press : *Briefings in functional Genomics*, Naville M. et Gautheret D. « Transcription attenuation in bacteria: theme and variations » (2009) 8 (6): 482-492. doi: 10.1093/bfpg/elp025, copyright 2010.

### 1.2.1 Les opérons bactériens

Chez les procaryotes, les gènes codants des enzymes dont les fonctions sont liées, par exemple les gènes impliqués dans le métabolisme du lactose, tendent à être regroupés physiquement, et se trouvent souvent sous le contrôle d'éléments régulateurs uniques. Ces types de gènes regroupés sont appelés « opérons ». Les travaux faits sur l'opéron lac, ainsi que sur d'autres opérons bactériens comme l'opéron trp, ont permis de proposer un modèle consensus pour ce type de structure (Reznikoff, 1972; Lewin, 1997; Bolsover, Hyams et al., 2006; Osbourn et Field, 2009).

Ce modèle est composé d'un gène répresseur, d'une région régulatrice, d'une région leader ou promotrice, d'une région d'atténuation et de la région de gènes de structure, qui se compose elle-même de plusieurs gènes regroupés en un segment codant (Figure 1.9 a). Dans ce mémoire, nous nous intéressons à la structure des opérons régulateurs du métabolisme des acides aminés qui correspondent en général au modèle de l'opéron trp (Figure 1.9 b), où les régions promotrice et d'atténuation sont identifiées comme des éléments importants dans le mécanisme d'atténuation de la transcription du gène (Roesser et Yanofsky, 1991; Mackey, 2001; Merino et Yanofsky, 2008).

a)



b)

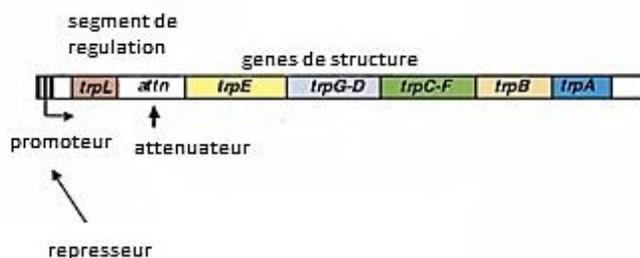


Figure 1.9 : La structure consensus des opérons bactériens, a) la forme consensus à partir du travail sur les opérons lac et trp; b) la forme consensus de l'opéron trp et les points de régulation possibles.

## 1.2.2 La transcription chez les procaryotes

Chez les procaryotes, la transcription commence par l'ouverture et le déroulement d'une petite portion de la double hélice d'ADN. Un des deux brins de cette hélice, le brin codant, sert de matrice pour la synthèse. Les enzymes qui effectuent la transcription sont appelées ARN polymérase (ARNp), et elles catalysent la formation des liaisons phosphodiester qui relient les nucléotides et forment une chaîne linéaire.

Les molécules d'ARNp, composées de plusieurs sous-unités, n'adhèrent que faiblement à l'ADN bactérien et glissent rapidement le long de l'ADN jusqu'à ce qu'elles se dissocient à nouveau. La polymérase glisse jusqu'une région spécifique de la double hélice d'ADN, appelée région promotrice, qui contient une séquence spécifique indiquant le point de départ de la synthèse d'ARN. Après s'être solidement

fixé sur le promoteur d'ADN, grâce au facteur  $\sigma$ , l'ARNp ouvre la double hélice pour exposer sur chaque brin une courte région de nucléotides. L'ADN étant déroulé, un des deux brins d'ADN exposés agit comme matrice pour permettre un appariement de bases complémentaires avec les ribonucléotides entrants, qui sont reliés par la polymérase débutant la chaîne d'ARN de façon que la nouvelle chaîne s'allonge d'un nucléotide à la fois dans la direction  $5' \rightarrow 3'$  (Figure 1.10) (Ishihama, 2000; Voet et Voet, 2005; Becker, Hardin et al., 2008).

Une fois que les dix premiers nucléotides de l'ARN ont été synthétisés, la polymérase subit des modifications structurales qui lui permettent de se déplacer rapidement en aval. L'élongation de la chaîne se poursuit, à une vitesse d'environ 50 nucléotides/minute pour les ARN polymérases bactériens, jusqu'à ce que l'enzyme rencontre un second signal sur l'ADN, le signal de terminaison ou le signal de stop, où elle s'arrête et libère à la fois la matrice d'ADN et la chaîne d'ARN synthétisé (Berg, Tymoczko et al., 2002; Becker, Hardin et al., 2008).

La libération presque immédiate du brin d'ARN ou d'ARNm signifie que de nombreuses copies d'ARN peuvent être faites à partir du même ADN et que la synthèse d'une deuxième molécule d'ARN commence avant la fin de la synthèse de la première chaîne (Alberts, Johnson et al., 2004; Voet et Voet, 2005; Bolsover, Hyams et al., 2006; Muller-Esterl, 2007).

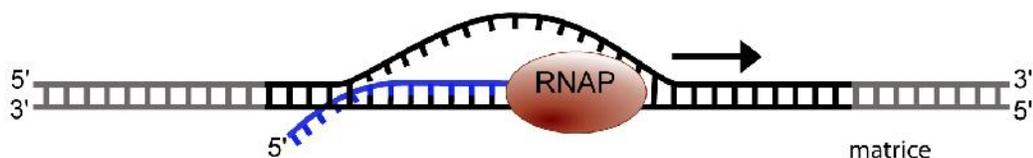


Figure 1.10 : L'ADN étant déroulé, un des deux brins d'ADN exposés agit comme matrice pour permettre un appariement de bases complémentaires avec les ribonucléotides entrants. De cette façon, la nouvelle chaîne s'allonge d'un nucléotide à la fois dans la direction 5'→3' (Figure adaptée de [http://en.wikipedia.org/wiki/File:Simple\\_transcription\\_elongation1.svg](http://en.wikipedia.org/wiki/File:Simple_transcription_elongation1.svg)).

### 1.2.3 La traduction chez les procaryotes

L'étape suivante, la traduction, permet la synthèse de protéines. Pour commencer la synthèse, un ribosome doit s'attacher à la matrice d'ARNm. Chaque ribosome est formé de deux sous-unités, une grande et une petite, toutes deux constituées d'ARN auquel s'ajoute un grand nombre de protéines. Les ribosomes de procaryotes sont 70S quand ils sont assemblés; ils sont constitués d'une grande sous-unité 50S et d'une plus petite sous-unité 30S (Figure 1.11). Les sous-unités sont caractérisées par la valeur S, qui indique la vitesse de sédimentation, c'est-à-dire la mesure de la rapidité avec laquelle une molécule se déplace dans le champ gravitationnel (Tortora, Funke et al., 2004; Malys et McCarthy, 2011).

Une fois assemblé, le ribosome protège un segment de 20 à 30 nucléotides sur l'ARNm pour éviter l'action des ribonucléases qui risqueraient de dégrader le messenger (Kozak, 1999). Cependant, c'est un segment plus court d'environ 8 à 15 nucléotides qui va se trouver exposé à l'interface de la sous-unité 30S durant la traduction (Yusupova, Yusupov et al., 2001).

La séquence ARNm est lue consécutivement par groupes de trois nucléotides ou codons. En considérant que l'ARN est un polymère linéaire composé de quatre nucléotides différents, 64 combinaisons de lecture de trois nucléotides sont possibles. Chaque codon spécifie soit un acide aminé, soit l'arrêt ou l'initiation du processus de traduction. L'ensemble des règles qui permettent d'établir une correspondance entre un codon et un acide aminé est résumé dans le code génétique.

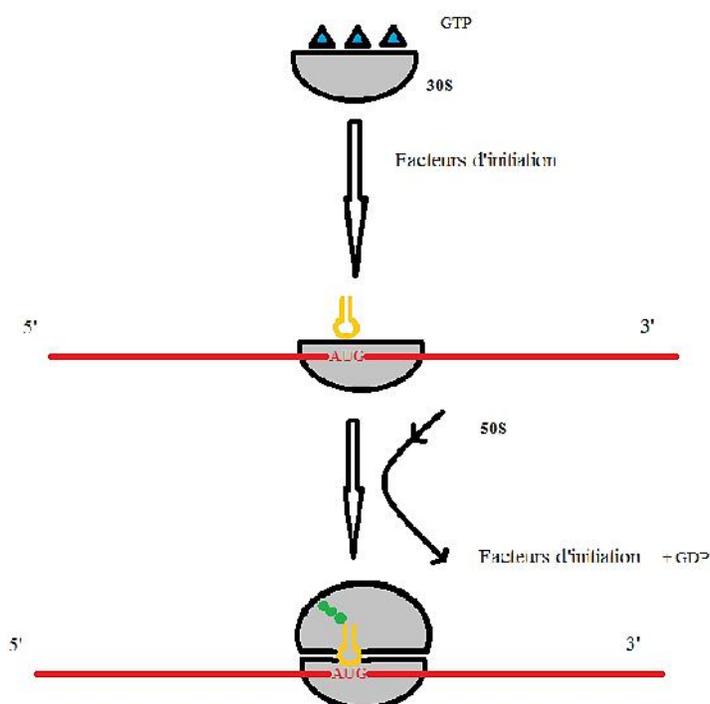


Figure 1.11 : La traduction chez les procaryotes commence avec l'assemblage du ribosome à partir des sous-unités 30S et 50S. Une fois le ribosome en place, la synthèse de la protéine commence lorsqu'un aminoacyl ARNt pénètre dans le site A du ribosome assemblé.

Il existe trois façons (cadres de lecture) de lire la séquence. La lecture se fait par groupe de trois nucléotides, où chaque groupe est décalé d'un nucléotide par rapport aux autres (+1 +2 ou +3 dans le cas de brin 5'→3'). Cependant, un seul de ces trois cadres de lecture possibles code dans l'ARNm pour la protéine requise. Le

cadre de lecture utilisé est défini par le premier AUG que le ribosome rencontre en aval du site de fixation du ribosome. Le premier acide aminé à être incorporé dans un nouveau polypeptide est toujours une méthionine modifiée, le formyle méthionine. La synthèse de la protéine commence lorsqu'un aminoacyl ARNt pénètre dans le site A du ribosome assemblé. L'identité de l' aminoacyl ARNt entrant est déterminée par le codon de l'ARNm à cette position. L'incorporation des acides aminés à la chaîne peptidique en train de se synthétiser, c'est-à-dire l'étape d'élongation, se fait par la catalyse d'une liaison peptidique entre deux acides aminés par l'enzyme peptidyl transférase. L'élongation de cette chaîne nécessite la participation de protéines appelées « facteurs d'élongation », qui ne seront pas étudiés dans le contexte de ce travail (Yusupova, Yusupov et al., 2001; Voet et Voet, 2005; Bolsover, Hyams et al., 2006; Steitz, 2008).

La terminaison de la synthèse d'une protéine se fait quand le ribosome trouve un des trois codons d'arrêt (UAA, UAG ou UGA). En même temps, la présence de facteurs de libération de la chaîne permet de libérer la chaîne polypeptidique, puis l'ARNm (Figure 1.12) (Lodish, Berk et al., 2000).

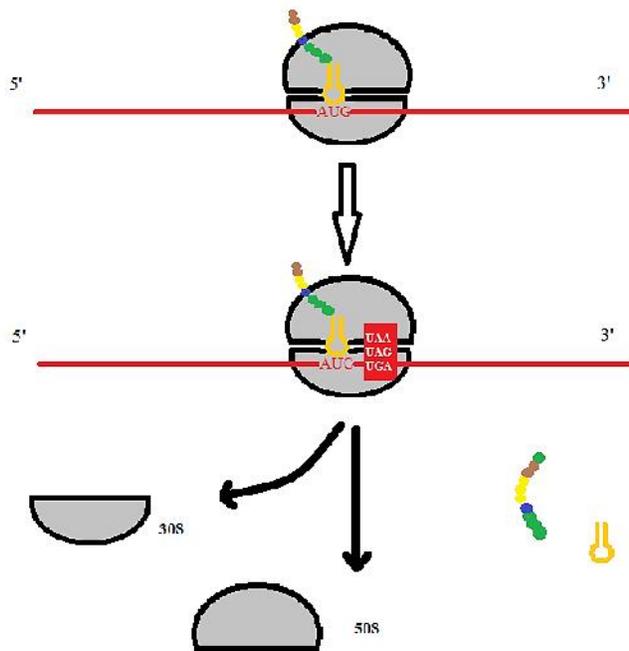


Figure 1.12 : La terminaison de la synthèse d'une protéine se fait quand le ribosome trouve un des trois codons d'arrêt (UAA, UAG ou UGA). En même temps, la présence de facteurs de libération de la chaîne permet de libérer la chaîne polypeptidique, puis l'ARNm.

#### 1.2.4 L'atténuation de la transcription chez les bactéries

Les bactéries ont développé divers mécanismes complexes pour contrôler la transcription ou la traduction de gènes ou d'opérons en réponse à des changements de leur milieu. Dans certains cas, la transcription est régulée au niveau de l'initiation par des protéines qui ciblent l'ADN et inhibent (répresseurs) ou stimulent (activateurs) l'initiation de la transcription. Une fois que la transcription commence, le complexe d'élongation de la transcription et le transcrit naissant sont des cibles potentielles de régulation.

Il existe plusieurs mécanismes d'atténuation de la transcription qui dépendent de la participation ou non d'une ou plusieurs molécules. Un de ces mécanismes, l'atténuation de la transcription du peptide leader, désigne un processus de régulation négative qui a lieu principalement pendant la phase d'élongation de la transcription (Elf, Berg et al., 2001). Dans ce cas, le transcrit naissant pourra se replier en une structure secondaire qui pourra déclencher un ralentissement de l'ARN-polymérase en l'obligeant à faire une pause ou à finir la transcription avant d'arriver aux gènes. Ce mécanisme diffère d'autres mécanismes de régulation par la génération de deux structures alternatives de l'ARNm, dont l'une réprime le gène en aval et l'autre permet de continuer la transcription/traduction (Fisher et Yanofsky, 1983; Landick et Phung, 1990; Bolsover, Hyams et al., 2006; Lyubetsky et Pirogov, 2006).

En général, ils sont classés en deux groupes, soit les mécanismes intrinsèques ou Rho-indépendant, soit les mécanismes de facteur dépendant (Yanofsky, 2000; Gollnick et Babitzke, 2002; Merino et Yanofsky, 2008; Naville et Gautheret, 2009).

### **1.2.5 Le mécanisme d'atténuation intrinsèque, ou Rho-indépendant**

Les terminateurs intrinsèques ou terminateurs rho-indépendants se composent de deux éléments de séquence: une répétition inversée courte d'environ 20 nucléotides suivie d'une séquence d'uridines. Lorsque la polymérase transcrit la séquence répétée inversée, l'ARNm peut adopter une structure en tige-boucle. Cette tige-boucle est censée provoquer la terminaison de la transcription en perturbant le complexe d'élongation formé par l'ADN, l'ARNm et l'ARN polymérase (Platt et Farhham, 1980; Wilson et Hippel, 1995; Uptain et Chamberlin, 1997; Gusarov et Nudler, 1999; Yamell et Roberts, 1999; Korzheva et Mustaev, 2001; Komissarova et Kashlev, 2002).

Plusieurs travaux ont essayé de décrire ce modèle. Ils suggèrent en général que le complexe d'élongation de la transcription est activement perturbé par la tige-

boucle, bien que plusieurs aspects de ce mécanisme demeurent inconnus (Ramesh, 1993; Wilson et Hippel, 1995; Uptain et Chamberlin, 1997; Mitra et Nagaraja, 2009; Naville et Gautheret, 2010; Peters, Vangeloff et al., 2011).

Deux études, soit le travail de d'Aubenton-Carafa (d'Aubenton-Carafa et Thermes, 1990) et, a posteriori, celle de Lesnik (Lesnik et Ecker, 2001), définissent les éléments qui permettent l'identification de ce type de structure en tige-boucle. Le travail de d'Aubenton-Carafa propose les éléments d'ADN qui permettent de définir les terminateurs intrinsèques, soit la tige-boucle et un segment de polythymines (poly-T) à la suite de la tige-boucle. Les observations faites sur 148 terminateurs ont permis à d'Aubenton-Carafa de définir quatre éléments de la tige-boucle : le nombre de paires GC de la tige ( $n_{GC}$ ); le nombre de nucléotides de la boucle ( $n_L$ ); le nombre de paires de bases de la base de la tige en 3' jusqu'au premier nucléotide de la poly-T ( $n_B$ ) et le nombre de nucléotides de la base de la tige en 5' jusqu'au premier nucléotide de la poly-T ( $L_h$ ) (Figure 1.13 a). Dans le cas du segment poly-T, d'Aubenton-Carafa a proposé une formule qui prend en considération la concentration en thymine d'un segment de 15 nucléotides suivant la tige-boucle ( $n_T$ ).

Des années plus tard, Lesnik a redéfini la description de la tige-boucle en redéfinissant les composants du segment de thymine et ceux de la tige-boucle terminatrice. À partir de ses observations des terminateurs chez *E.coli*, il a proposé de diviser le segment de poly-T en trois régions appelées « proximal », « distal » et « extra » (Figure 1.13 b) et d'incorporer un score qui prend en considération la stabilité de l'hybride ARN:ADN du segment poly-T.



### 1.2.6 Les mécanismes d'atténuation de la transcription de l'opéron *trp* dans *Escherichia coli*

Comme nous l'avons mentionné, le phénomène fut découvert au début des années 1970 par l'équipe de Charles Yanofsky, dans le cadre de l'étude de la régulation de l'opéron gouvernant la biosynthèse du tryptophane chez *E.coli* (opéron *trp* EDCBA). Cependant, le terme « atténuation » fut introduit par Kasai (1974), qui a observé un phénomène semblable en analysant la régulation de l'opéron histidine chez *Salmonella typhimurium*. Un point important dans ces études a été de confirmer que l'atténuation chez les procaryotes est fondée sur le couplage entre la transcription et la traduction; c'est-à-dire que la traduction démarre sur l'ARNm alors que ce dernier est encore en cours de synthèse (Richardson, 1993; Henkin, 1996).

L'équipe de Yanofsky s'est aperçue que même chez les mutants dépourvus du répresseur actif *trpR*, un autre élément de régulation, l'opéron *trp*, était toujours régulé. Les chercheurs ont constaté que la régulation était préservée chez les mutants à condition qu'une courte région située juste avant de l'opéron *trp* soit présente (Lee et Yanofsky, 1977; Oxender et Yanofsky, 1979; Yanofsky, 1981).

Dans le cas de l'opéron *trp*, en moyenne 140 nucléotides séparent le point de départ de la transcription et le codon d'initiation du premier gène structural, *trpE*. L'extrémité 5' de cette séquence, appelée leader ou *trpL*, code un court peptide de 14 acides aminés contenant deux résidus de tryptophane. Le transcrit du segment leader et le segment atténuateur jusqu'à *trpE* peuvent former deux types de structures secondaires en fonction de la disponibilité des segments notés 1, 2, 3 et 4 dans la Figure 1.14. La structure formée par les segments 1 et 2 provoque une pause de l'ARNp; les segments 3 et 4 forment une tige-boucle terminatrice ou Rho-indépendante qui sert de terminateur de transcription, tandis que les segments 2 et 3 forment une tige-boucle qui empêche la formation de la tige-boucle terminatrice et qu'on l'appelle tige-boucle antiterminatrice (Stroynowski et Yanofsky, 1983; Kuroda

et Yanofsky, 1984; Landick et Phung, 1990; Gusarov et Nudler, 1999; Elf et Ehrenberg, 2005).

Une fois transcrits, les segments 1 et 2 forment donc sur l'ARNm une structure qui provoque une pause de l'ARNp sans que celle-ci décroche de sa matrice. La durée de cette pause est suffisante pour permettre au ribosome de commencer la traduction du peptide leader (Landick et Yanofsky, 1984; Yanofsky et Fisher, 1984; Landick et Yanofsky, 1987). L'avancement du ribosome pousse l'ARNp à reprendre la transcription, et l'ARNp ainsi que le ribosome avancent de façon synchrone sur leurs matrices (Keller et Calvo, 1979; Landick et Yanofsky, 1985).

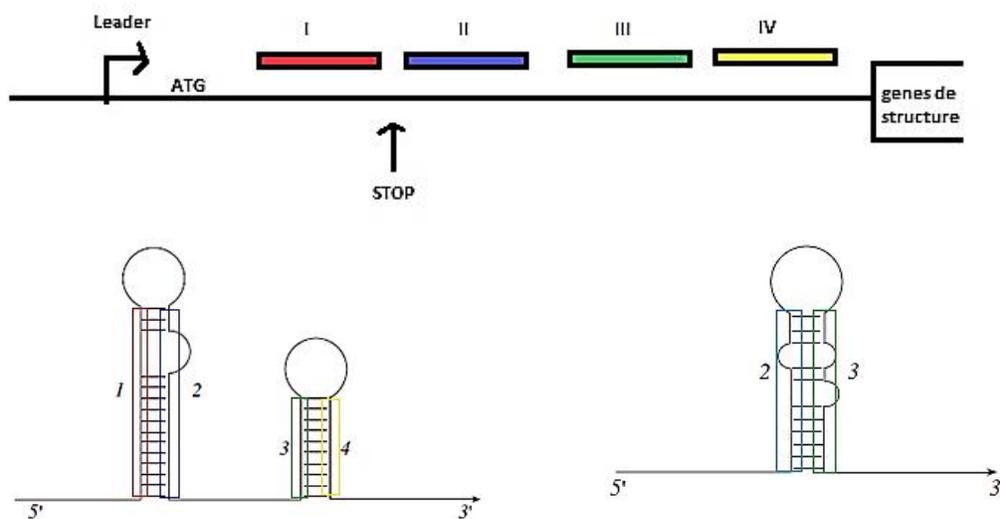


Figure 1.14 : La séquence du gène *trpL* et le segment d'atténuation de l'opéron *trp*. Quatre segments sont identifiés; ils peuvent former divers appariements entre eux et diverses conformations selon le passage et le rythme de la synthèse de l'ARN polymérase et l'action du ribosome.

La suite peut se dérouler suivant deux scénarios. Si la concentration d'ARNt chargé de tryptophane (ARNt<sup>trp</sup>) est faible, le ribosome s'arrête sur un ou les deux codons tryptophane du peptide leader pendant que l'ARNp poursuit la transcription. Ces conditions favorisent la formation d'une structure secondaire antiterminatrice (segments 2 et 3) permettant la poursuite de la traduction de l'ensemble des gènes de l'opéron trp. Dans le cas contraire, en présence d'une forte concentration d'ARNt<sup>trp</sup>, la transcription du peptide leader se déroule sans problème jusqu'à son codon stop. L'avancement du ribosome synchronisé avec celui de l'ARNp prévient la formation de la structure antiterminatrice et favorise la formation de la structure des segments 3 et 4, qui agit comme un terminateur de la transcription. Lorsque cette structure est formée, l'ARNp cesse son activité prématurément et le complexe d'élongation se détache sans même atteindre le premier gène de l'opéron trp (Figure 1.15) (Kuroda et Yanofsky, 1984; Gollnick et Babitzke, 2002; Henkin et Yanofsky, 2002; Nudler et Gottesman, 2002; Grundy et Henkin, 2006).

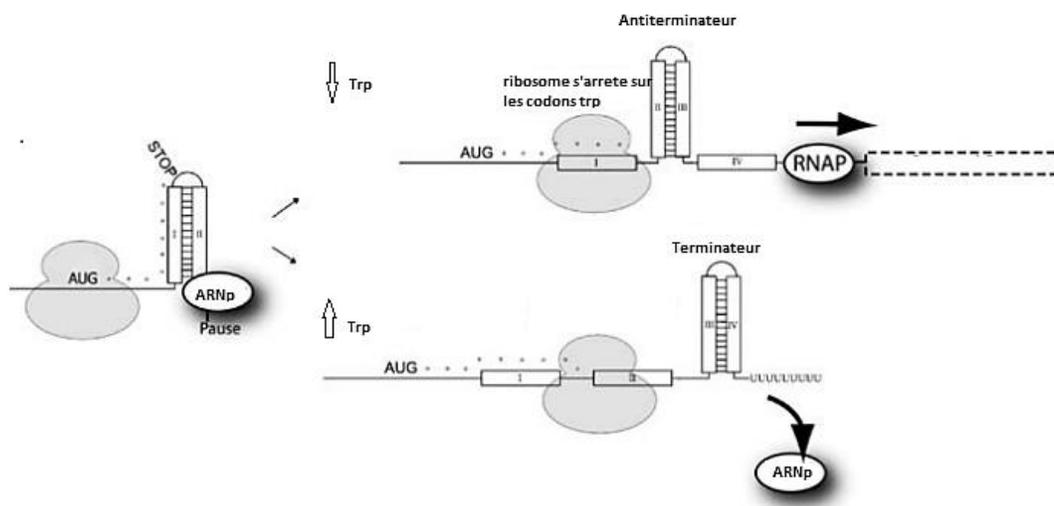


Figure 1.15 : Le mécanisme d'atténuation de la transcription de l'opéron trp chez *E. coli* peut suivre deux scénarios selon la concentration de trp: soit la formation du termineur à haute concentration de trp, soit la formation d'une tige-boucle antiterminatrice à basse concentration de trp. (Figure adaptée de Elf et Ehrenberg, 2005).

Des travaux similaires sur d'autres opérons (*leu*, *his*, *phe*, *ilv*, *thr*) chez *E. coli* ont démontré que le mécanisme d'atténuation joue un rôle important pour l'atténuation de la transcription. Dans chaque cas, le peptide leader contient plusieurs acides aminés correspondant à la voie de biosynthèse régulée et ils semblent suivre la même méthode de régulation que l'opéron *trp* (Keller et Calvo, 1979; Johnston et Roth, 1980; Wessler et Calvo, 1981; Gardner, 1982; Carter et Calvo, 1986; Lyn et Gardner, 1987; Chan et Landick, 1989; Bartkus et Calvo, 1991; Chen et Umbarger, 1991; Panina et Gelfand, 2001).

En 1977, Lee et Yanofsky (Lee et Yanofsky, 1977) ont suggéré un modèle consensus à même d'expliquer le mécanisme d'atténuation. En 1979, Oxender a proposé un premier modèle qui est devenu le modèle consensus et celui à utiliser pour la recherche et l'identification de candidats des atténuateurs de la transcription (Figure 1.16) (Oxender et Yanofsky, 1979).

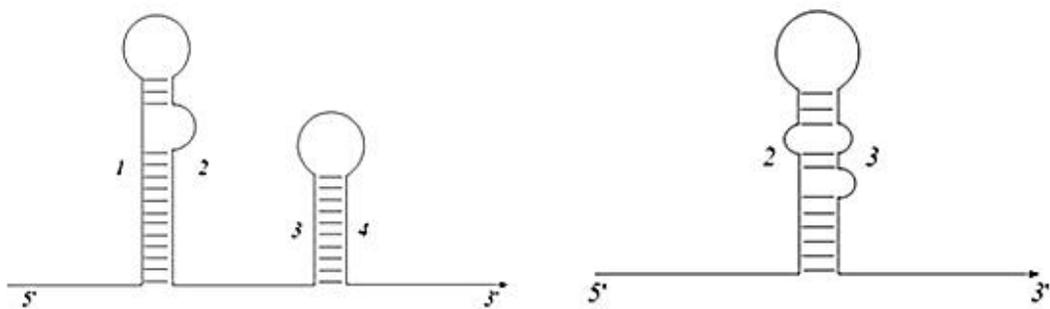


Figure 1.16 : Modèle consensus des structures proposées par le modèle d'atténuation de la transcription. La première structure correspond à la structure d'arrêt de la transcription, ou structure terminatrice. La deuxième correspond à l'antitermineur, ou structure qui permet la poursuite de la transcription/traduction.

### 1.2.7 Les méthodes de prédiction des atténuateurs de la transcription

Dans les années qui ont suivi à la découverte du mécanisme d'atténuation chez *E.coli*, plusieurs autres travaux sur les opérons leucine et histidine chez *Salmonella typhimurium* ont démontré que ce mécanisme existe chez plusieurs organismes bactériens (Kasai, 1974; Johnston et Roth, 1980; Chan et Landick, 1989; Bartkus et Calvo, 1991; Delorme et Renault, 1999).

D'autres travaux indiquent que ce même mécanisme s'applique aux opérons *thr*, *phe* et *ilv* en utilisant des méthodes expérimentales ou en se basant principalement sur les analyses de similarité de la séquence. À la suite du travail de d'Aubenton-Carafa en 1990, divers chercheurs ont commencé à développer des méthodes pour identifier l'élément terminateur ou les gènes qui pourront être contrôlés en utilisant le modèle d'atténuation (Ermolaeva et Salzberg, 2000; Elf, Berg et al., 2001; Lyubetskaya et Lyubetsky, 2003; Vitreschak et Lyubetsky, 2004; Kingsford et Salzberg, 2007; Keller et Calvo, 1979; Yanofsky, 1981; Gardner, 1982; Kolter et Yanofsky, 1982; Lyn et Gardner, 1987).

En 2000 et 2003, deux travaux intéressants de Vitreschak (Vitreschak et Gelfand, 2000) et Lyubetskaya (Lyubetskaya et Lyubetsky, 2003) présentent des approches qui utilisent principalement le modèle de comparaison au modèle de consensus de l'opéron *trp*, la reconnaissance de la tige-boucle terminatrice, la présence d'une région leader, les éléments de régulation sur cette région et un segment d'atténuation après le gène leader.

Après 2001, l'apparition des logiciels tels que RibEx (Abreu-Goodger et Merino, 2005), Erpin (Lambert et Gautheret, 2004) ont permis l'identification de riborégulateurs en utilisant des méthodes comparatives. Plus récemment, ARNold (Naville et Gautheret, 2010) permet l'identification des tiges-boucles terminatrice à partir de l'implémentation du modèle de Lesnik sur des séquences de longueur limitée. Mais dans tout cas, ce type de logiciel ne permet pas d'identifier les

séquences d'ARN susceptibles de présenter les changements de repliement vus dans le modèle d'atténuation.

### **1.2.8 Le but de ce mémoire**

Comme il a été discuté dans ce chapitre, le repliement de l'ARNm peut être influencé par divers facteurs. Dans le mécanisme d'atténuation de la transcription chez les bactéries, il est évident que l'existence de deux différentes structures secondaires est due à la présence d'un ou plusieurs facteurs. L'existence d'une structure secondaire définit la fin ou la poursuite du processus de transcription/traduction. Si l'on regarde de près ce mécanisme, on pourra conclure que le repliement de la séquence est influencé essentiellement par les pauses transcriptionnelles, le déplacement du ribosome sur l'ARNm et la concentration cellulaire d'un ou plusieurs acides aminés.

La recherche d'ARNm répondant au modèle d'atténuation de la transcription dans un génome complet est principalement basée sur les méthodes comparatives de séquences; la reconnaissance des éléments qui définissent ce type d'atténuateurs, comme l'identification des segments qui incluent des gènes leaders; la distance entre les éléments; le type codon que contient la séquence et l'existence d'une tige-boucle en 3' de la séquence. Mais dans ce type de recherche, le processus de repliement et la reconnaissance de structures alternatives de la séquence ne semblent pas être pris en considération à cause de la présence des facteurs et de ses effets sur le repliement. Les principaux problèmes d'une méthode qui considère le repliement de l'ARNm est de savoir quels sont les facteurs d'influence principaux du repliement, et comment les prendre en compte dans la prédiction de ce dernier.

Le but de ce mémoire est de proposer une méthode d'analyse de repliement de l'ARNm qui considère la présence de ces facteurs d'influence et qui permettra d'identifier les séquences dont le repliement se verra affecté par ces derniers. Par la suite, nous démontrerons que la méthode a été testée sur des atténuateurs de

transcription dans des génomes bactériens. Enfin, nous présenterons comment la méthode a été incorporée au logiciel TA-Search (*Transcription Attenuation Search*), qui permet d'identifier des riborégulateurs potentiels qui répondent au changement de repliement en combinant la méthode proposée avec la recherche d'éléments génomiques.

## **Chapitre 2 : Étude du repliement des atténuateurs de la transcription**

Dans ce chapitre, nous démontrerons qu'on peut prédire efficacement l'effet de la fixation d'un ligand/cofacteur sur la structure d'un ARN en contraignant localement la prédiction de la structure secondaire. Nous démontrerons que cette méthode peut être utilisée comme un élément supplémentaire pour identifier les riborégulateurs de la transcription.

Comme nous l'avons mentionné dans le chapitre 1, le modèle d'atténuation de la transcription chez les procaryotes est fondé sur l'existence de deux structures alternatives dans l'ARNm situées avant un gène ou un opéron. Un des éléments le plus importants de ce modèle est le changement du repliement provoqué par la fixation d'un ligand/cofacteur qui permettra ou non la transcription et la traduction du gène ou de l'opéron (Elf et Ehrenberg, 2005).

### **2.1 Étude de l'impact de cofacteurs dans le repliement de l'ARN**

Le repliement de l'ARNm peut être affecté par la présence d'autres intervenants moléculaires. Comme nous l'avons mentionné dans la section 1.1.3, ces éléments ou cofacteurs peuvent être des métabolites, des protéines, des enzymes, de l'ADN ou tout autre ARN interagissant avec la molécule en contraignant le repliement à adopter certaines conformations (Keller et Calvo, 1979; Henkin et Yanofsky, 2002; Sosnick et Pan, 2003; Pan et Sosnick, 2006).

Pour étudier l'effet du ligand/cofacteur sur la molécule en cours de repliement, ce travail propose une méthode fondée dans l'analyse des repliements en considérant la possibilité de contraindre le repliement en indiquant des zones qui

seront non appariées. Par la suite, la méthode compare les résultats avec un modèle de repliement de la molécule sans ligand ou avec un repliement suggéré.

Pour une séquence d'ARNm, la méthode commence par générer un ensemble d'instructions qui seront soumises à un logiciel de prédiction de la structure secondaire. Chaque ensemble correspond à la position de fixation du cofacteur sur différents nucléotides de l'ARNm. Les positions sont choisies en considérant les sites possibles de fixation; par exemple, sur un nucléotide ou un ensemble de nucléotides qui ne seront pas appariés durant le repliement. Par la suite, la méthode récupère les repliements pour le comparer à une structure proposée avant l'analyse. Deux possibilités sont examinées au cours de l'analyse : soit le changement se manifeste dans le repliement MFE, soit il se trouve dans l'ensemble des solutions ou sous-optimaux. Dans le dernier cas, pour considérer que le changement est significatif par rapport à l'ensemble des solutions, il doit se présenter dans un grand nombre des solutions. La récupération de repliements pour l'analyse se fait en utilisant l'application Search-MFE décrite en Annexe A. Une fois que l'analyse est terminée, la méthode produit un fichier contenant les positions où l'interaction du cofacteur provoque de changements sur le repliement. Ainsi, la méthode permet non seulement d'identifier les changements, mais aussi d'identifier les positions où le cofacteur induit ces derniers.

Un grand nombre de logiciels de prédiction de structure secondaire de l'ARN permettent d'implémenter la méthode en incorporant des contraintes simulant les sites de fixation du ligand/cofacteur et, en même temps, produisent un ensemble de solutions pour chaque repliement. Des logiciels comme RNAsubopt (Hofacker et Standler, 2006), MFold (Zucker et Stiegler, 1981), SFold (Ding et Lawrence, 2003) ou MC-Fold (Parisien et Major, 2008) permettent, en utilisant des options « Structural mask » ou « Constrains information », d'incorporer la fixation de cofacteurs sur l'ARN.

## 2.2 Les cas du mécanisme d'atténuation de l'opéron trp

Pour évaluer la pertinence de la méthode, nous l'avons testée en analysant le repliement de l'ARNm de l'atténuateur de l'opéron trp chez les bactéries. Comme nous l'avons mentionné dans la section 1.2.6, le mécanisme d'atténuation de la transcription a la particularité de présenter deux structures différentes selon la présence de cofacteurs. La revue de littérature nous a permis de constater la présence de trois cofacteurs à même d'influencer le repliement de l'ARNm. Le premier cofacteur est lié au passage de l'ARNp sur le brin d'ADN qui génère des pauses transcriptionnelles pendant la transcription; le deuxième, à la disponibilité de ARNt<sup>Trp</sup> et le dernier, au passage du ribosome sur l'ARNm durant la traduction (Keller et Calvo, 1979; Landick et Yanofsky, 1984; Landick et Yanofsky, 1985; Landick et Yanofsky, 1987). En prenant en considération la littérature, nous avons décidé de tester la méthode en considérant seulement le passage du ribosome, le cofacteur, sur l'ARNm. Dans ce cas, nous avons utilisé trois types de masques qui représentent les différentes positions où le cofacteur va se fixer sur l'ARNm. Chaque masque est défini par :

Type 1 - 15 nucléotides sont bloqués à la fois à partir de l'extrémité 5' et en parcourant l'ARNm vers l'extrémité 3';

Type 2 - à partir de 15 nucléotides en 5', on ajoute 3 nucléotides à la fois jusqu'à la fin de l'ARNm à l'extrémité 3';

Type 3 - 3 nucléotides sont bloqués à la fois à partir de 5' et en parcourant l'ARNm vers 3'.

Le choix de ces trois masques était dû, dans le premier et le troisième cas, à la recherche de l'intervention minimale du cofacteur susceptible d'affecter le repliement de l'ARNm. Dans le premier cas, le segment d'interface de la sous-unité 30S tel que décrit dans la section 1.2.3 est le facteur pouvant déclencher le changement et, dans le troisième cas, seule l'interface du site A du ribosome provoque l'effet recherché.

Dans le deuxième masque, le choix a été fait en considérant le déplacement du cofacteur sur le brin d'ARNm, comme il a été décrit pour le masque du type 1, mais en considérant qu'après le passage du ribosome, la séquence ne sera pas capable de se replier et que le brin restera disponible à un autre ribosome pour démarrer un processus de traduction sur le même brin (Figure 2.1). En appliquant les trois masques sur la séquence d'atténuateur de l'opéron *trp* chez *E.coli*, nous avons observé que, dans les trois masques, le changement de repliement se présente au MFE lorsque le cofacteur se trouve sur les nucléotides 29 à 34, qui représentent deux codons *trp* et qui correspondent au modèle d'atténuation de la transcription dans l'opéron *trp* (Figures 2.2 à 2.4).

```

> Escherichia-coli
ATGAAAGCAATTTTCGTACTGAAAGGTTGGTGGCGCACTTCC TGAAACGGGCAGTGTATTACCATGCGTAAAGCAATCAGATACCCAGCCCGCCTAATGAGCGGGC

..... ***** type 1
***..... *****
*****..... *****
*****..... *****
*****..... *****

..... ***** type 2
..... *****
..... *****
..... *****
..... *****

... ***** type 3
*... *****
**... *****
***... *****
***... *****

```

Figure 2.1 : Pour représenter l'interaction d'un cofacteur, on utilise le masque pour indiquer les positions de fixation sur l'ARNm qui ne doivent pas former de paires de bases. Trois types de masques sont considérés: le type 1 bloque 15 nucléotides à partir de la position 5' en avançant sur la séquence vers l'extrémité 3'; le type 2 bloque à partir d'un premier segment de 15 nucléotides, il ajoute trois nucléotides pour bloquer de façon progressive la séquence; le type 3 bloque la séquence de trois nucléotides à la fois.



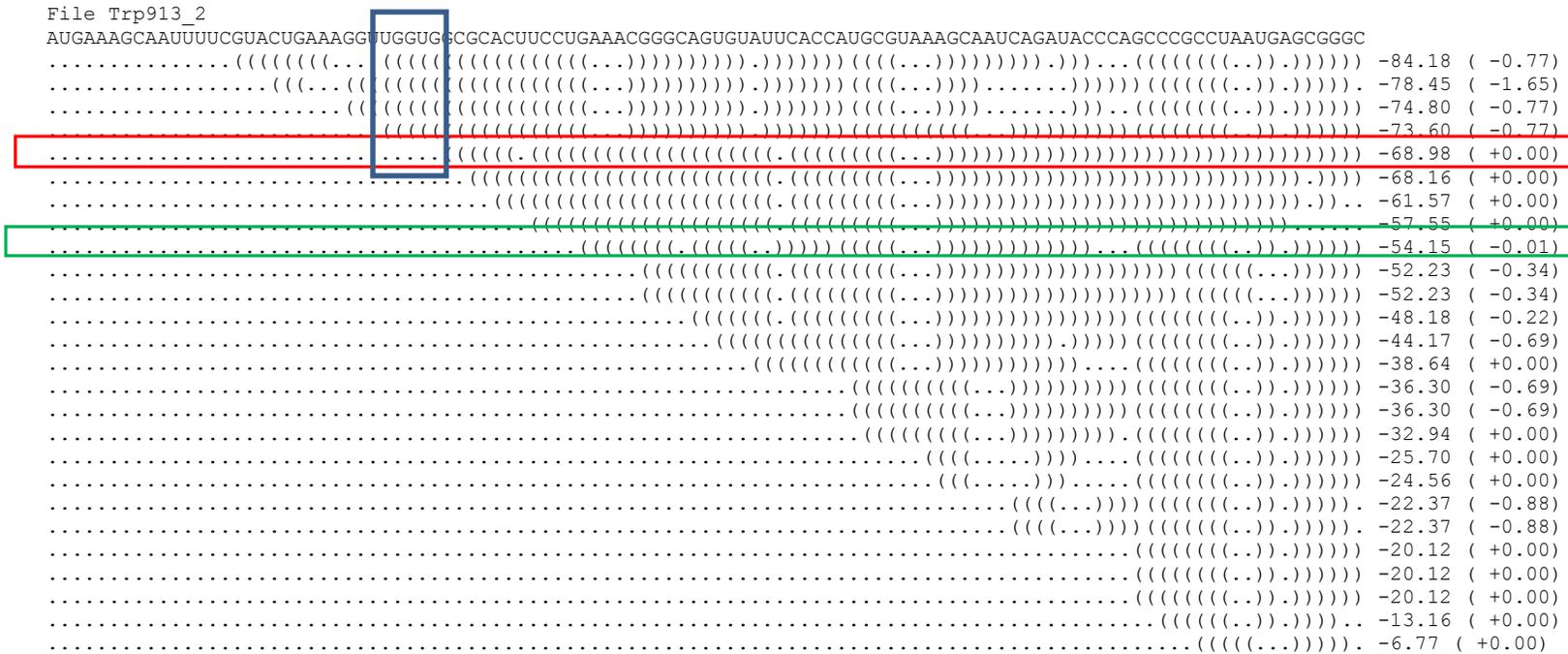


Figure 2.3 : La méthode appliquée en utilisant le masque de type 2; chaque ligne représente le repliement MFE pour chaque position de fixation du ligand. La boîte rouge indique où le repliement a changé, et la boîte bleue, la position du ligand qui provoque le changement. Dans ce cas, les positions 24 à 25 codent deux codons trp.



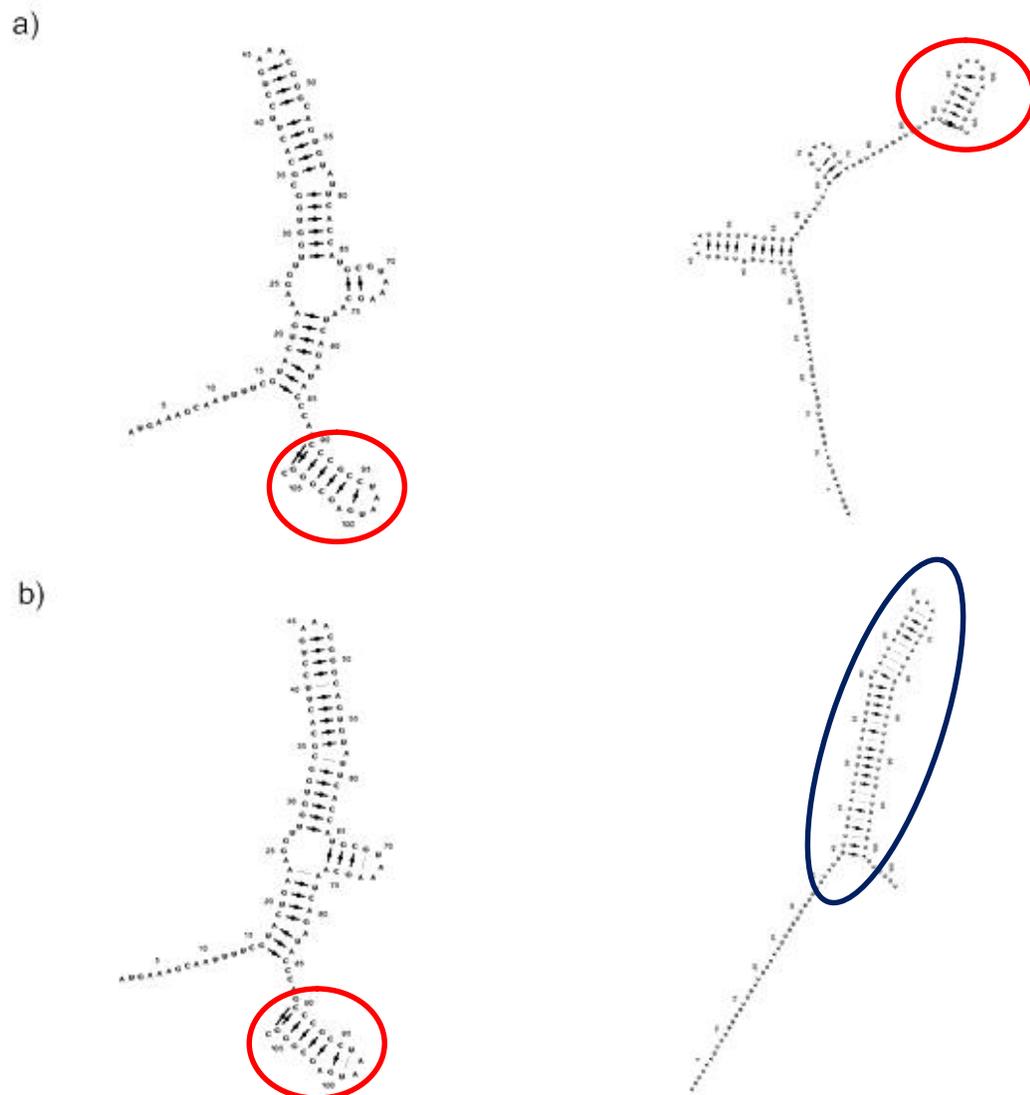
Afin d'étudier le repliement de l'ARNm de l'atténuateur de l'opéron *trp*, nous avons décidé de faire une première analyse avec quatre séquences de l'atténuateur de l'opéron *trp* en utilisant les logiciels présentés dans la section 2.1. Les séquences prévenantes d'*Escherichia coli* (NC\_000913), *Salmonella enterica Tryphimurium* (NC\_003197), *Citrobacter rodentium* (NC\_013716) et *Yersenia pestis* (NC\_004088), comprennent le gène leader (*trpL*) et le segment d'atténuation entre le codon stop de *trpL* et le premier résidu de la poly-T avant le premier gène structural de l'opéron.

Dans l'analyse des résultats obtenus avec les quatre logiciels (Tableau 2.1), on distingue trois types de résultats, qui dépendent de la présence d'un changement de conformation dans le repliement conforme à ceux présentés dans le modèle consensus à deux structures. Le premier résultat montre la présence d'un changement du repliement entre les deux structures du modèle d'atténuation qui se produit dans le repliement MFE de chaque séquence quand le masque se trouve sur les codons *trp* (Figure 2.5 b); ce cas est indiqué dans le tableau par le signe « + ». Le deuxième résultat est l'absence du changement du repliement dans le MFE ou dans l'ensemble de solutions sous-optimales de chaque séquence, et il est indiqué par « - ». Le dernier résultat correspond à l'identification du changement qui se présente dans le premier 5 % de l'ensemble des solutions avec une prévalence de plus de 45 % dans tous les résultats; ce résultat est indiqué par « ~ » dans le tableau 2.1.

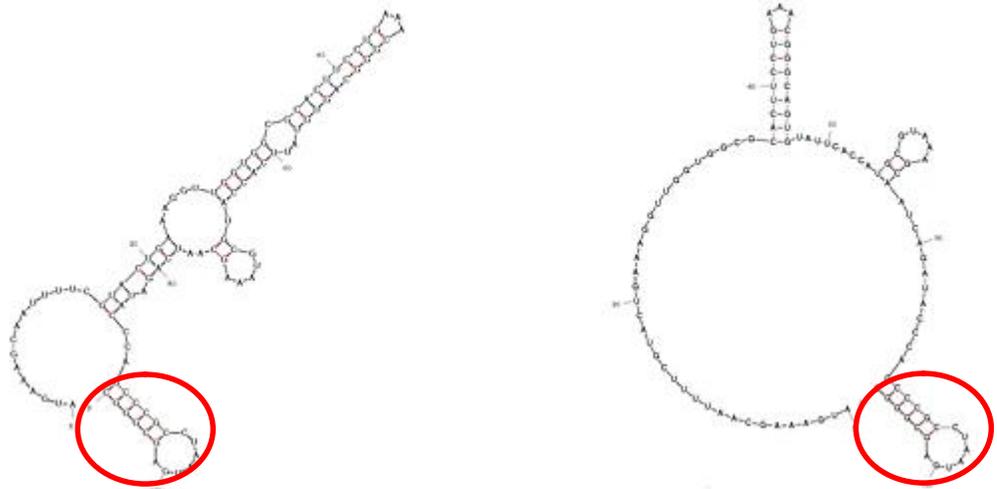
	RNAsubopt			Mfold			MC-Fold			Sfold		
	1	2	3	1	2	3	1	2	3	1	2	3
NC_000913	-	-	-	-	-	-	+	+	+	-	-	-
NC_013716	-	-	-	-	-	-	+	+	~	-	-	-
NC_003197	-	-	-	-	-	-	+	~	+	-	-	-
NC_004088	-	-	-	-	-	-	+	+	~	-	-	-

Tableau 2.1 : L'analyse du repliement de l'opéron *trp*, en quatre génomes avec quatre logiciels de prédiction en utilisant les trois masques, permet d'observer si les logiciels sont capables de prédire le changement du repliement du modèle consensus de l'atténuation. Si « + », le changement a été aperçu en le repliement MFE; si « - », le modèle n'a pas aperçu dans le repliement; si « ~ », le changement a été aperçu dans l'ensemble de sous-optimaux.

En examinant les résultats obtenus dans le tableau 2.1, nous avons remarqué que les changements de repliement de l'ARN sont reflétés seulement dans les résultats avec MC-Fold. Dans ce cas, pour l'implémentation de la méthode, nous avons décidé d'utiliser MC-Fold, parce qu'il semble reproduire l'effet d'un ligand/cofacteur sur le repliement de l'ARNm.



c)



d)

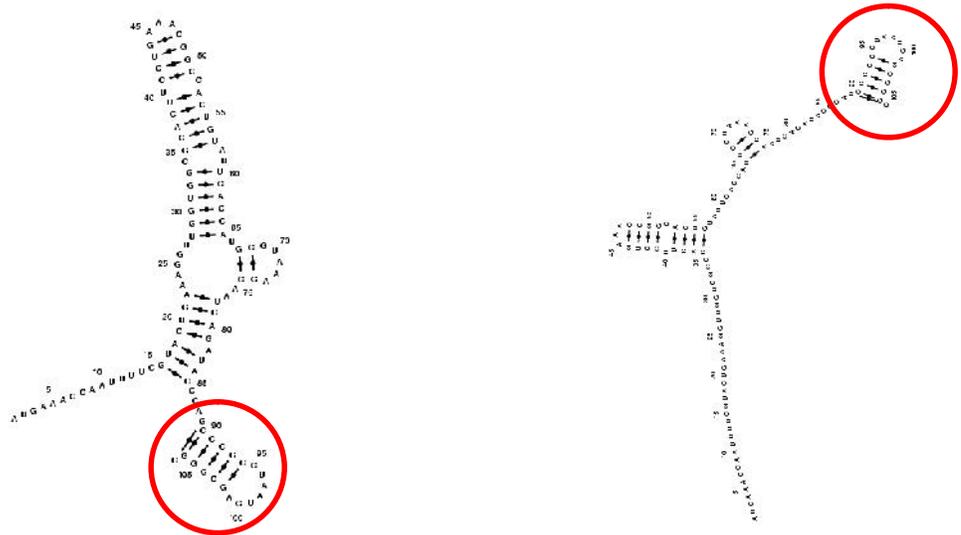


Figure 2.5 : Le repliement de l'ARNm de l'atténuateur de l'opéron *trp* chez *E.coli* avec les quatre logiciels. Quand le masque de type 2 se trouve sur les deux codons régulateurs *trp*, le repliement présente un changement tel qu'observé dans l'option b et prédit par MC-Fold. Les options a, c et d, prédites à l'aide des autres logiciels, ne montrent pas le résultat attendu, ni le changement du repliement. Dans chaque figure, la tige-boucle terminatrice est indiquée en rouge et la tige-boucle antiterminatrice, en bleu, telle que présentée au modèle à deux structures.

### 2.3 La structure secondaire comme élément d'analyse du modèle d'atténuation

Par la suite, nous avons appliqué la méthode à un plus grand ensemble de séquences des différents opérons provenant d'un sous-ensemble de données de Vitreschak (Tableau 2.2) (Vitreschak et Lyubetsky, 2004).

NC_000913	<i>Escherichia coli</i> str. K-12 substr. MG1655
NC_002655	<i>Escherichia coli</i> O157:H7 str. EDL933 chromosome, complete genome
NC_002695	<i>Escherichia coli</i> O157:H7 str. Sakai
NC_003197	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Typhimurium str. LT2
NC_003198	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Typhi str. CT18
NC_003910	<i>Colwellia psychrerythraea</i> 34H
NC_004088	<i>Yersinia pestis</i> KIM 10
NC_004337	<i>Shigella flexneri</i> 2a str. 301
NC_004741	<i>Shigella flexneri</i> 2a str. 2457T
NC_006840	<i>Vibrio fischeri</i> ES114
NC_006905	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Choleraesuis str. SC-B67
NC_007384	<i>Shigella sonnei</i> Ss046
NC_007606	<i>Shigella dysenteriae</i> Sd197
NC_007613	<i>Shigella boydii</i> Sb227
NC_009648	<i>Klebsiella pneumoniae</i> subsp. <i>pneumoniae</i> MGH 78578
NC_009708	<i>Yersinia pseudotuberculosis</i> IP 31758
NC_010554	<i>Proteus mirabilis</i> HI4320
NC_010658	<i>Shigella boydii</i> CDC 3083-94
NC_011283	<i>Klebsiella pneumoniae</i> 342
NC_011740	<i>Escherichia fergusonii</i> ATCC 35469
NC_012125	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Paratyphi C strain RKS4594
NC_013364	<i>Escherichia coli</i> O111:H- str. 11128
NC_013716	<i>Citrobacter rodentium</i> ICC168

Tableau 2.2 : Liste de génomes bactériens chez lesquels le mécanisme a été étudié ou prédit auparavant (Vitreschak et Lyubetsky, 2004)

Les séquences contenant les opérons leaders et les segments d'atténuation des organismes du tableau 2.2, ont été obtenus de GenBank.

(<http://www.ncbi.nlm.nih.gov/genbank/>, 31 avril 2011).

Les opérons dont les séquences d'atténuateurs ont été analysées à l'aide de cette méthode correspondent à trp, phe, thr, ilv, leu et his (Tableau 2.3). Une liste complète de ces séquences se trouve en Annexe B. Par la suite, nous avons généré l'ensemble d'instructions de MC-Fold pour chaque séquence avec les trois masques signalés dans la section 2.2.

TRP	ILV	LEU	THR
NC_000913	NC_000913	NC_000913	NC_000913
NC_002655	NC_002695	NC_002655	NC_002695
NC_002695	NC_003197	NC_003198	NC_003198
NC_003197	NC_003910	NC_003197	NC_004088
NC_004088	NC_004088	NC_003910	NC_004741
NC_004337	NC_004741	NC_004741	NC_006840
NC_004741	NC_007384	NC_006840	NC_009648
NC_006840	NC_007606	NC_006905	NC_009708
NC_007606	NC_007613	NC_007384	NC_010658
NC_007613	NC_010554	NC_007606	NC_011283
NC_009708	NC_011283	NC_007613	NC_012125
NC_011740	NC_011740	NC_011283	NC_013716
NC_012125	NC_012125	NC_011740	
NC_013364	NC_013716	NC_013716	
NC_013716			
HIS	PHE		
NC_000913	NC_000913		
NC_002695	NC_003198		
NC_003198	NC_004088		
NC_003910	NC_004337		
NC_004337	NC_004741		
NC_007384	NC_007613		
NC_007613	NC_009708		
NC_011283	NC_011740		
NC_011740	NC_013364		
NC_012125			
NC_013716			

Tableau 2.3 : Liste des opérons dont les séquences se trouvent annotées dans Genbank. Les opérons choisis sont trp, phe, thr, ilv, leu et his.

Les résultats complets de l'analyse sont présentés dans le tableau 2.4. De ces résultats, la méthode détecte que quatre des six opérons semblent suivre le modèle à deux structures du mécanisme d'atténuation. Dans les cas de *his* et *leu*, le changement a été observé dans deux cas isolés en *leu* sur les séquences de NC\_003197 et NC\_003198. Dans le cas de ces deux opérons, une analyse différente sera proposée dans la discussion.

Nous avons effectué une dernière analyse avec une séquence d'ARNm qui contient les mêmes éléments génomiques que ceux qui composent l'atténuateur de l'opéron *trp* chez *E.coli* (voir les sections 1.2.6 et 1.2.7); la séquence correspond à *E. coli* NC\_000913, segment 4118381:418500c. Les résultats ont démontré que le repliement de l'ARNm ne change pas en considérant le cofacteur qui a été utilisé dans le tableau 2.4. Dans ce cas, on pourra conclure que le cofacteur choisi a un impact important sur les repliements des opérons bactériens, impact qui ne se reflète pas en séquences de composition similaire (Figure 2.6).

	TRP			PHE			ILV			THR			LEU			HIS		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
NC_000913	+	+	+	-	+	-	-	+	-	+	+	+	-	-	-	-	-	-
NC_002655	+	+	+										-	-	-			
NC_002695	+	+	+				-	+	-	+	+	+				-	-	-
NC_003197	+	+	+				-	+	-				-	+	-			
NC_003198				+	~	+				+	+	+	-	+	-	-	-	-
NC_003910							-	+	-				-	-	-	-	-	-
NC_004088	+	+	~	-	~	-	-	+	-	+	+	+						
NC_004337	+	+	+	-	+	-										-	-	-
NC_004741	+	+	+	-	+	-	-	+	-	+	+	+	-	-	-			
NC_006840	+	+	-							+	+	+	-	-	-			
NC_006905								+					-	-	-			
NC_007384							-	+	-				-	-	-	-	-	-
NC_007606	+	+	+				-	+	-				-	-	-			
NC_007613	~	+	~	+	+	-	-	+	-				-	-	-	-	-	-
NC_009648										+	+	+						
NC_009708	~	+	~	+	+	-				+	+	+						

	TRP			PHE			ILV			THR			LEU			HIS		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
NC_010554							-	+	-									
NC_010658										+	+	+						
NC_011283							-	+	-	+	+	+	-	-	-	-	-	-
NC_011740	~	+	+	-	+	-	-	+	-				-	-	-	-	-	-
NC_012125	+	+	+				-	+	-	+	+	+				-	-	-
NC_013364	+	+	+	+	+	-												
NC_013716	+	+	+				-	+	-	+	+	+	-	-	-	-	-	-
MFE	12/16	16/16	12/16	4/9	7/9	1/9	0/14	14/14	0/14	12/12	12/12	12/12	0/12	2/12	0/12	0/11	0/11	0/11

Tableau 2.4 : Pour chaque opéron dans chaque génome, selon le tableau 2.3, MC-Fold a été lancé en considérant les trois masques. Après avoir récolté les résultats du repliement en utilisant Search-MFE ,Annexe A, ils reflètent si le changement de repliement a été observé selon le modèle. Si « + », le modèle a été aperçu dans le repliement MFE ; si « - », le modèle n'a pas aperçu dans le repliement; si « ~ », le modèle a été aperçu dans les repliements sous-optimaux.



## **Chapitre 3 : TA-Search**

Dans le chapitre précédent, nous avons démontré qu'une méthode d'analyse des repliements de structures secondaires permet d'identifier si l'interaction d'un ligand/cofacteur peut altérer le repliement de l'ARNm. En même temps, nous avons vérifié que la méthode peut être utilisée pour étudier les atténuateurs du mécanisme d'atténuation de la transcription chez les procaryotes. À partir de ces résultats, nous voulons implémenter la méthode pour la recherche de riborégulateurs sur des génomes bactériens complets en particulier, dans la recherche des atténuateurs de transcription.

L'objectif de ce chapitre est d'automatiser la recherche de ces riborégulateurs sur des génomes bactériens complets, en combinant la recherche classique des atténuateurs vus dans la section 1.2.7 avec la méthode décrite dans le chapitre 2.

### **3.1 TA-Search : Recherche des atténuateurs de la transcription**

La recherche des atténuateurs de la transcription sur les génomes complets a été limitée à l'observation des éléments qui caractérisent ce type d'atténuateurs. Un des inconvénients de ce type de recherche est le grand volume de candidats qui ne sont pas impliqués dans les mécanismes de régulation, mais qui contiennent des éléments recherchés. L'inclusion de la méthode d'analyse du repliement permet de limiter les nombres de candidats en sachant que les atténuateurs de transcription sont capables de changer leur repliement en interagissant avec des cofacteurs. L'étape suivante a consisté à implémenter un logiciel, TA-Search, développé en Java, qui marie les deux approches (Figure 3.1).

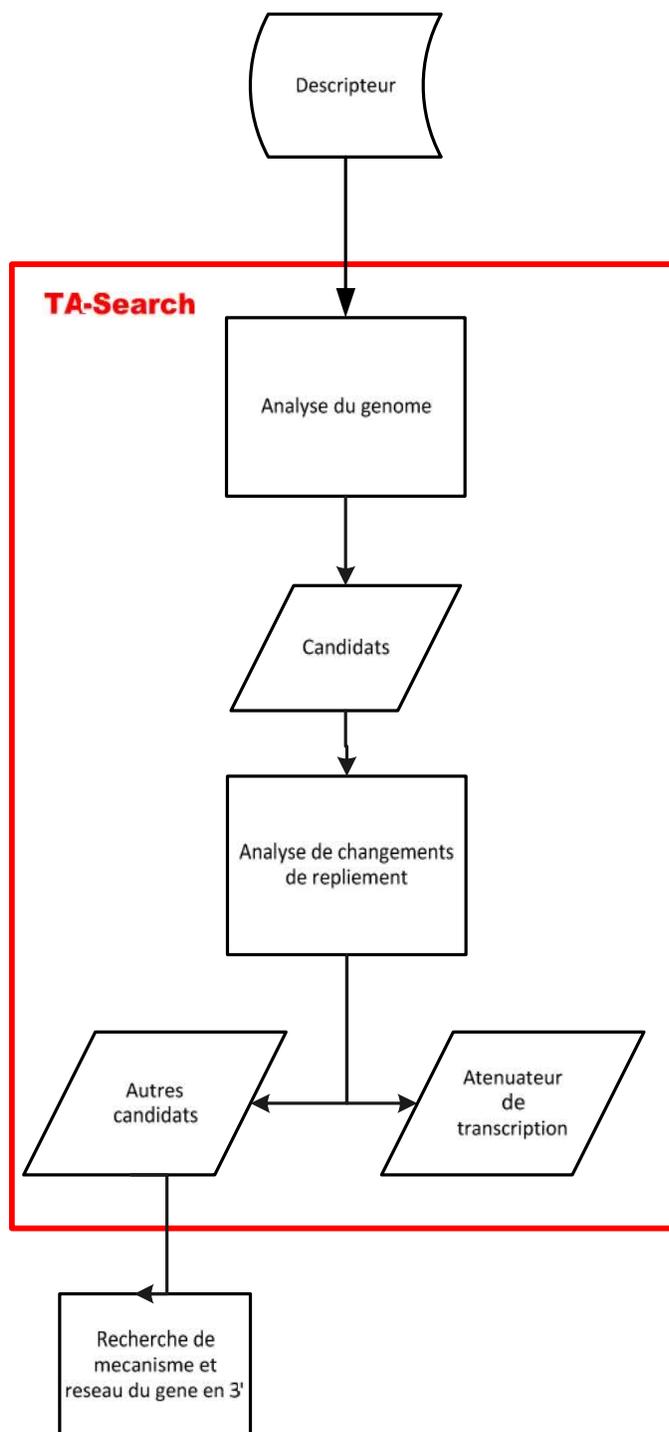


Figure 3.1 : Le modèle de recherche des atténuateurs commence à partir d’un fichier descripteur qui alimente l’analyse du génome de TA-Search. TA-Search est divisé principalement en deux analyses : celle du génome et celle du repliement.

### 3.1.1 Analyse du génome

TA-Search est divisé en deux analyses. La première cherche à identifier des candidats en se basant sur les éléments qui caractérisent la séquence d'un opéron bactérien et les atténuateurs de cet opéron.

Le logiciel commence l'analyse en prenant un fichier descripteur qui contient l'identification des génomes à analyser, l'identification d'une base de données contenant les notations génomiques des divers génomes et un descripteur qui contient les éléments et des paramètres dont le logiciel a besoin pour identifier chaque candidat à atténuateur d'un opéron bactérien.

Le descripteur contient, pour chaque opéron bactérien, un vecteur qui définit ses éléments caractéristiques, comme il a été vu dans la section 1.2.6. Dans le contexte de ce travail, le terme « vecteur » définit une liste de valeurs qui caractérisent les éléments à rechercher. Ces éléments sont les distances entre le codon d'initiation et le codon stop du segment leader; la distance maximale entre la fin du segment atténuateur et le codon d'initiation du leader; le type de codons régulateurs du segment leader; le nombre de codons et la distance maximale du candidat au codon d'initiation du gène structural en 3' (Figure 3.2).

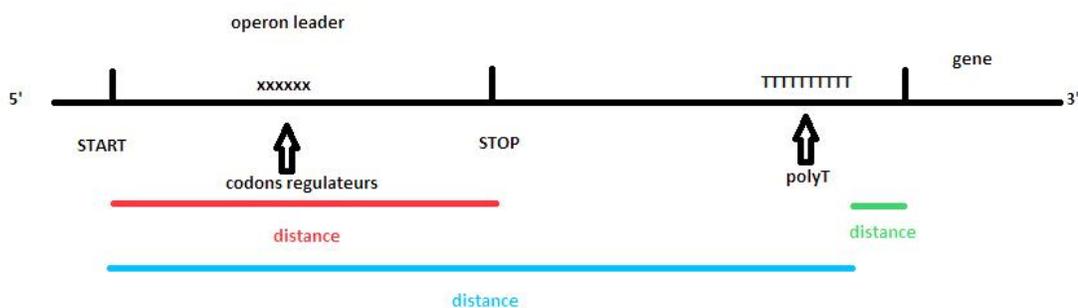


Figure 3.2 : Les éléments qui sont recherchés par TA-Search sont les codons d'initiation et de stop sur le même cadre de lecture, le segment poly-T, les codons régulateurs dans l'opéron leader, la distance entre initiation-stop (rouge), la distance entre le codon d'initiation et le dernier élément de la poly-T (bleu), la distance entre le dernier élément de la poly-T et le gène le plus proche en 3' (vert).

TA-Search commence en identifiant les codons d'initiation et les codons stop pour lesquels le logiciel garde sa position et son cadre de lecture. À cette étape, le logiciel cherche également le segment de haute concentration en Thymine ou poly-T. Par la suite, le logiciel combine en respectant les distances proposées pour chaque opéron (Figure 3.2): les codons d'initiation, les codons de stop et les segments poly-T, afin de générer un premier ensemble de candidats. Par la suite, une fois récupérées les séquences de chaque candidat, le logiciel vérifie si les codons régulateurs indiqués pour l'opéron dans le fichier descripteur se trouvent dans le segment considéré comme leader. Les codons régulateurs considérés par le logiciel peuvent être un motif simple, la répétition d'un même acide aminé ou encore un motif complexe, c'est-à-dire un ensemble d'acides aminés dans un ordre spécifique, avec ou sans répétition e la séquence. Par la suite, le logiciel calcule le score de Lesnik pour le segment de poly-T (Lesnik et Ecker, 2001). Les détails des éléments et les valeurs choisies pour la recherche des atténuateurs sont décrits en Annexe C.1.

Pour réduire l'ensemble, le logiciel retient les candidates qui se trouvent dans des segments indiqués dans la figure 3.3

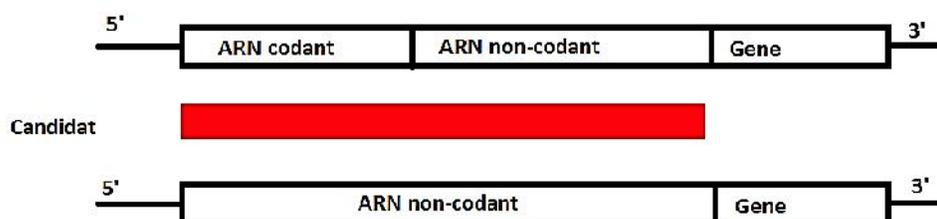


Figure 3.3 : Les candidats de TA-Search peuvent être décrits selon leur position par rapport aux gènes; une partie du candidat est localisée sur un gène codant, et l'autre partie, sur une région non codante de l'ARNm juste avant un gène. D'autre part, le candidat est localisé dans une région non codante de l'ARNm avant un gène.

Par la suite, le logiciel calcule la présence de la tige-boucle terminatrice sur le segment précédant la poly-T, la concentration de CG et l'énergie de la tige-boucle terminatrice à des fins de sélection. À la fin de cette analyse, le logiciel génère un fichier contenant les candidats retenus et un autre contenant l'ensemble des instructions à soumettre à la méthode d'analyse de repliement en utilisant le masque

de type 2, qui considère comme cofacteur le passage du ribosome tel qu'utilisé dans la section 2.3. Bien que les autres masques puissent être utilisés, à partir des résultats présentés dans le tableau 2.4, le masque de type 2 semble permettre d'identifier adéquatement les changements dans tous les cas testés. Les fichiers de candidats et le script sont présentés en Annexe C.2

### 3.1.2 Identification des changements dans les repliements

La deuxième analyse de TA-Search s'effectue à partir la méthode décrite dans la section 2.2 avec les candidats obtenus par l'analyse du génome. En considérant les deux structures du modèle d'atténuation de la transcription, le changement de repliement recherché par l'analyse pour cette implémentation est associé à la déstabilisation de la tige-boucle terminatrice du modèle antitermineur du mécanisme d'atténuation. Les résultats obtenus par cette analyse reflètent les candidats qui semblent répondre au modèle d'atténuation de la transcription. L'étape suivante consistera à identifier les gènes ciblés pour chaque candidat susceptible d'être régulé par ce mécanisme d'atténuation, ou à établir si les candidats sont impliqués dans la régulation d'autres processus métaboliques.

## 3.2 Recherche de candidats : validation de TA-Search sur des atténuateurs de la transcription déjà identifiés

Pour tester l'implémentation de TA-Search, nous avons sélectionné un ensemble de quatre génomes bactériens : *Escherichia coli* (NC\_000913), *Salmonella Enterica Typhi* (NC\_003198), *Shigella flexneri* (NC\_004741) et *Citrobacter rodentium* (NC\_013716). Le choix de ces génomes est basé sur l'existence de plusieurs études de ce mécanisme chez *E.coli* et *Salmonella typhimurium* qui permettent de corroborer les résultats de TA-Search. Les séquences génomiques ont été récupérées de GenBank ([www.ncbi.nlm.nih.gov/genbank/](http://www.ncbi.nlm.nih.gov/genbank/), 31 mai 2011). Pour ce test, le fichier descripteur est composé d'un fichier contenant les quatre génomes, d'un autre

contenant les vecteurs descripteurs de chaque opéron et de la base de données gene2accession décrite en Annexe D.

TA-Search a été lancé pour chaque génome sur les deux brins, 5'→3' et 3'→5' en cherchant les atténuateurs pour les opérons trp, phe, ilv et thr. Les nombres de candidats pour chaque opéron après l'analyse du génome sont résumés dans le tableau 3.1. Deux résultats ont attiré notre attention dans la prédiction en thr sur les génomes NC\_003198 et NC\_013716 : le logiciel n'a pas réussi à trouver des candidats. En observant les séquences de thrL et les segments d'atténuation notés, nous avons constaté que le motif pour l'opéron thr est légèrement différent dans les deux cas, et qu'en particulier, ces deux séquences partagent le même motif pour thr, plus court que le motif utilisé dans l'analyse originale. En faisant la modification pour le motif thr, le logiciel a généré des candidats pour les deux opérons sans ajouter de résultats supplémentaires pour les candidats des autres opérons (Tableau 3.2). À la fin du processus, après l'analyse du repliement, les nombres des candidats potentiels sont résumés dans le tableau 3.3. Les résultats finaux ont démontré que 90 % des candidats retenus étaient ceux déjà identifiés comme des atténuateurs, et que pour les 10 % restant, la régulation par l'atténuateur correspondait à une voie de synthèse reliée à l'opéron.

Brin	Trp		Phe		Ilv		Thr	
	+	-	+	-	+	-	+	-
NC_000913	2	4	4	6	1	0	1	0
NC_003198	2	5	3	4	0	1	0	0
NC_004741	3	1	1	0	0	1	1	0
NC_013716	2	8	5	0	0	1	0	0

Tableau 3.1 : TA-Search a été lancé avec les paramètres énumérés dans le fichier descripteur sur chaque génome sur les deux brins, 5'→3' et 3'→5'.

Brin	Trp		Phe		Ilv		Thr	
	+	-	+	-	+	-	+	-
NC_000913	2	4	4	6	1	0	1	0
NC_003198	2	5	3	4	0	1	1	0
NC_004741	3	1	1	0	0	1	1	0
NC_013716	2	8	5	0	0	1	1	0

Tableau 3.2 : Une fois changée la définition du motif régulateur de l'opéron thr, l'analyse de génome de TA-Search a donné les candidats pour l'opéron thr dans tous les génomes.

Brin	Trp		Phe		Ilv		Thr	
	+	-	+	-	+	-	+	-
NC_000913	1	1	2	2	1	0	1	0
NC_003198	1	1	2	1	0	1	1	0
NC_004741	0	1	1	0	0	1	1	0
NC_013716	1	3	2	0	0	1	1	0

Tableau 3.3 : Le nombre de candidats proposés par TA-Search pour les quatre génomes analysés.

Parmi les résultats figurent les candidats observés pour l'opéron *trp* chez *E. coli* (NC\_000913). Après avoir effectué l'analyse sur les deux brins du génome avec TA-Search, le premier résultat de l'analyse du génome a généré un premier ensemble de six candidats pour l'opéron *trp*. Après l'analyse du repliement, le logiciel récupère deux candidats au modèle d'atténuation. En observant le résultat du repliement de chaque candidat, on constate que le changement de repliement se fait quand le masque se trouve sur les deux codons *trp* et qu'après le passage du masque sur cette position, le repliement change pour présenter une très forte tige-boucle similaire à la tige-boucle terminatrice du modèle d'atténuation (Figure 3.4 a). Parmi ces deux candidats, un seul a été associé au gène *trpE* sur le brin 3'→5'. Dans les quatre autres cas identifiés dans l'analyse du génome, les résultats de l'analyse du repliement permettent de conclure qu'il n'y a aucun changement de repliement significatif et que le logiciel les a rejetés (Figure 3.4 b).

Un autre exemple de résultat est la prédiction pour l'opéron *trp* en NC\_003198, où le gène leader *trpL* n'est pas noté. Dans ce cas, l'analyse du génome a trouvé sept candidats et, par la suite, l'analyse du repliement a proposé trois candidats. Un de ces trois candidats sur le brin 3'→5' est associé au geneId 1247728, qui correspond au gène *trpE*; soit le premier gène de l'opéron *trp* en NC\_003198.



### 3.3 Recherche des nouveaux candidats sur des génomes complets

À la suite de l'analyse des résultats avec des atténuateurs connus, on a constaté que TA-Search est sensible à l'identification des candidats au modèle d'atténuation en combinant les deux approches, soit l'analyse des éléments génomiques et l'analyse du repliement du candidat.

Par la suite, nous avons procédé à une analyse de quatre génomes qui ont été incorporés récemment à la base de données de GenBank: *Erwinia billingiae* (NC\_014306), *Aeromonas veronii* (NC\_015424), *Yersenia enterocolitica* (NC\_008800) et *Escherichia coli* (NC\_012947). Dans les quatre génomes sélectionnés, la notation était incomplète ou hypothétique. (<http://www.ncbi.nlm.nih.gov/genbank/>, 1<sup>er</sup> août 2011)

TA-Search a été lancé sur les deux brins pour les quatre opérons (trp, phe, ilv et thr), et les résultats de l'analyse du génome sont présentés dans le tableau 3.4. Les résultats de l'analyse du repliement sont présentés dans le tableau 3.5.

Brin	Trp		Phe		Ilv		Thr	
	+	-	+	-	+	-	+	-
NC_008800	3	2	1	11	2	0	0	0
NC_012947	1	1	11	0	0	1	0	1
NC_014306	4	2	10	0	0	0	0	1
NC_015424	1	1	1	3	0	0	0	0

Tableau 3.4 : TA-Search a été lancé sur quatre génomes, dont l'annotation est partielle ou hypothétique. En utilisant les mêmes paramètres que dans la section 2.2, après l'analyse du génome, le logiciel est arrivé à proposer des candidats.

Brin	Trp		Phe		Ilv		Thr	
	+	-	+	-	+	-	+	-
NC_008800	3	2	1	5	2	0	0	0
NC_012947	1	1	4	0	0	1	0	1
NC_014306	3	1	5	0	0	0	0	1
NC_015424	1	1	1	3	0	0	0	0

Tableau 3.5 : TA-Search a été lancé sur quatre génomes, dont l'annotation est partielle ou hypothétique. En utilisant les mêmes paramètres que dans la section 2.2, après l'analyse du repliement, le logiciel est arrivé à proposer des candidats.

En observant les résultats, il est évident que la sensibilité de TA-Search semble être faible pour les opérons *ilv* et *thr*. Il a déjà été mentionné que les motifs qui semblent caractériser ses atténuateurs ont été sélectionnés à partir de la revue littéraire. En analysant plusieurs génomes bactériens, en particulier ceux présentant des motifs complexes comme ceux trouvés dans *thr* et *ilv*, nous avons conclu qu'il existe des sous-motifs du motif consensus original de *thr* et *ilv* qui pourront également être utilisés pour la recherche en cas d'échec ou pour obtenir un plus grand ensemble de résultats. Dans ce cas, nous avons lancé TA-Search pour les deux opérons avec des sous-motifs du motif original. Les résultats sont résumés dans le tableau 3.6.

Brin	Ilv		Thr	
	+	-	+	-
NC_008800	2	0	2	0
NC_012947	0	1	0	1
NC_014306	0	0	1	1
NC_015424	0	0	0	1

Tableau 3.6 : Dans les cas de thr et ilv, les paramètres ont été modifiés pour améliorer la prédiction. Le plus grand changement est lié au motif qui a identifié les codons régulateurs dans chaque cas.

À partir des résultats du tableau 3.5, une recherche sur les gènes sujets à ce mécanisme devrait être effectuée. À titre d'exemple, nous avons retenu les candidats obtenus pour le génome NC\_008800 sur l'opéron trp. Parmi les cinq candidats finaux, quatre sont associés soit à des protéines hypothétiques, soit au gène ilvB, qui ne sont pas liés à la régulation de l'opéron trp. Un seul cas, le candidat sur la position [2419772,2419898] sur le brin 5'→3', est associé au gène identifié par 4715480, qui appartient à trpE, le premier gène de l'opéron trp.

## **Chapitre 4 : Discussion et conclusion**

Un grand nombre de mécanismes de régulation des gènes chez les bactéries se fondent principalement en l'existence de plus d'un repliement possible de l'ARNm. Ces repliements dépendent de la présence d'un métabolite spécifique ou d'autres cofacteurs pouvant interférer avec le repliement de la molécule. En particulier, le mécanisme d'atténuation de la transcription chez les bactéries se fonde sur l'existence de deux structures qui se forment à diverses concentrations d'un acide aminé particulier, et une des deux structures empêche la transcription/traduction du gène qui se trouve après la séquence d'atténuation.

### **4.1 L'utilisation de la prédiction de structures secondaires**

L'objectif principal de ce travail était de trouver une méthode pouvant s'ajouter à la recherche classique en se basant sur l'étude du repliement de la séquence de l'atténuateur et permettant d'ajouter de l'information au moment de choisir les candidats au modèle d'atténuation de la transcription.

Comme il a été expliqué dans la section 1.2, le repliement de l'ARNm peut se voir affecté par différents cofacteurs à diverses étapes de la transcription ou de la traduction. Nous nous sommes attardés à l'étude du repliement et à son application dans la recherche des atténuateurs de la transcription. Dans le mécanisme d'atténuation, les cofacteurs qui semblent avoir plus d'influence sont la concentration d'un acide aminé particulier, les pauses transcriptionnelles provoquées par l'ARNp pendant la transcription, la vitesse de transcription et de traduction et le déplacement du ribosome sur l'ARNm pendant la traduction.

En considérant ces cofacteurs, le problème qui s'est posé était lié à l'implémentation d'une méthode permettant d'étudier le repliement en considérant tous les cofacteurs. En cherchant un logiciel de repliement de structure secondaire de l'ARN, nous n'avons découvert aucun logiciel à même d'incorporer tous les cofacteurs dans la prédiction. Les logiciels comme SFold, MFold, MC-Fold ou RNAsubopt, mentionnés dans le chapitre 2, ne permettent pas de considérer un grand

nombre de cofacteurs de façon simultanée. Dans ce cas, nous avons envisagé la possibilité de réduire le nombre des cofacteurs à considérer dans la méthode.

Par une première analyse, nous avons évalué l'impact des pauses transcriptionnelles sur l'ADN, puis nous avons tenté de les caractériser dans la méthode. En effectuant une revue de la littérature portant sur ce type de pauses, nous avons conclu que l'information disponible, notamment quant à l'atténuateur de l'opéron *trp* chez *E.coli*, ne permettait pas de formuler une généralisation des occurrences des pauses sur les séquences concernées (Platt et Farhham, 1980; Landick et Yanofsky, 1985; Chan et Landick, 1989; Yanofsky, 1984; Landick et Yanofsky, 1987). Nous avons donc pris la décision d'éliminer les pauses transcriptionnelles de cette version de la méthode. Sachant que, chez les bactéries, le processus de transcription/traduction est un processus coordonné, nous avons choisi de considérer le repliement de la séquence comme un processus dans lequel la vitesse de transcription/traduction ne puisse pas être considérée comme un cofacteur influençant substantiellement le repliement et, pour cette raison, il a été éliminé de la méthode.

Par la suite, nous avons analysé l'impact de la concentration d'un acide aminé particulier sur le repliement. La concentration de l'acide est liée à sa disponibilité dans l'étape de traduction; une diminution significative de la concentration provoque une brève pause dans le déplacement du ribosome sur les codons régulateurs qui correspondent au même acide aminé sur la séquence d'ARNm. Cette pause fait en sorte que le segment sous le ribosome ne soit pas disponible à la formation de paires de bases avec le reste de l'ARNm en cours de transcription, en même temps que la traduction s'effectue. On peut en déduire que le déplacement du ribosome sur la séquence d'ARNm est le cofacteur qui affecte directement le repliement de l'ARNm, et non la concentration de l'acide aminé.

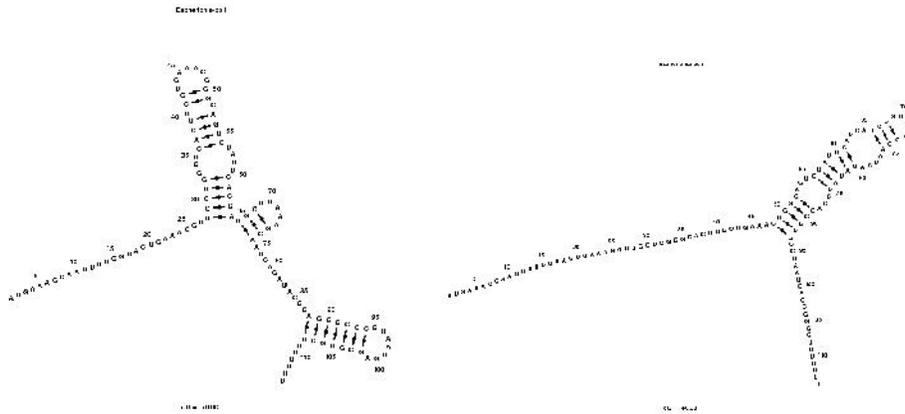
De cette façon, le seul cofacteur que la méthode doit considérer pour la recherche des atténuateurs de transcription est le déplacement du ribosome sur la séquence d'ARNm. Pour incorporer le cofacteur aux logiciels présentés dans le

chapitre 2, nous avons utilisé l'option de « Structural Constraints » proposée par les quatre logiciels.

De façon générale, toute la littérature consultée pour ce travail (voir le chapitre 1) considère le repliement de la séquence d'un atténuateur de la transcription comme un événement décrit par un modèle à deux structures qui représentent les étapes du processus de repliement. Mais, si l'on considère le repliement de l'ARNm comme un processus dynamique, la meilleure solution ne devra pas correspondre à une solution unique; elle devra plutôt être recherchée dans l'ensemble des repliements de la séquence dans l'espace. Ce concept correspond bien à la perspective de ce travail, basé sur la recherche de candidats au modèle d'atténuation en considérant un ensemble de solutions dans l'espace des solutions du repliement de l'ARNm. Ainsi, les logiciels proposés dans la section 2.2 ont été sélectionnés parce qu'ils permettent l'inclusion d'un cofacteur et qu'ils fournissent un sous-ensemble de solutions au repliement de la séquence proposée.

Avant de poursuivre, revenons au modèle d'atténuation proposé par Yanofsky en 1981 (Yanofsky, 1981), selon lequel les deux repliements du modèle d'atténuation présentent seulement des appariements de type canonique (Figure 4.1 a). En considérant son modèle, devenu le modèle consensus du mécanisme, nous nous serions attendu à ce que tous les logiciels utilisés dans l'analyse soient en mesure de tenir compte de ces deux états du modèle consensus, mais seul MC-Fold a pu détecter ce changement (Figure 4.1 b). Parmi les logiciels retenus, on remarque que seul MC-Fold considère dans le repliement la contribution de l'énergie de paires de bases non canoniques. Il semble que ces paires de bases confèrent de la flexibilité au repliement de la séquence de l'atténuateur pour déclencher un changement assez stable pour provoquer soit la déstabilisation de la tige-boucle terminatrice, soit la formation de la tige-boucle antiterminatrice.

a)



b)

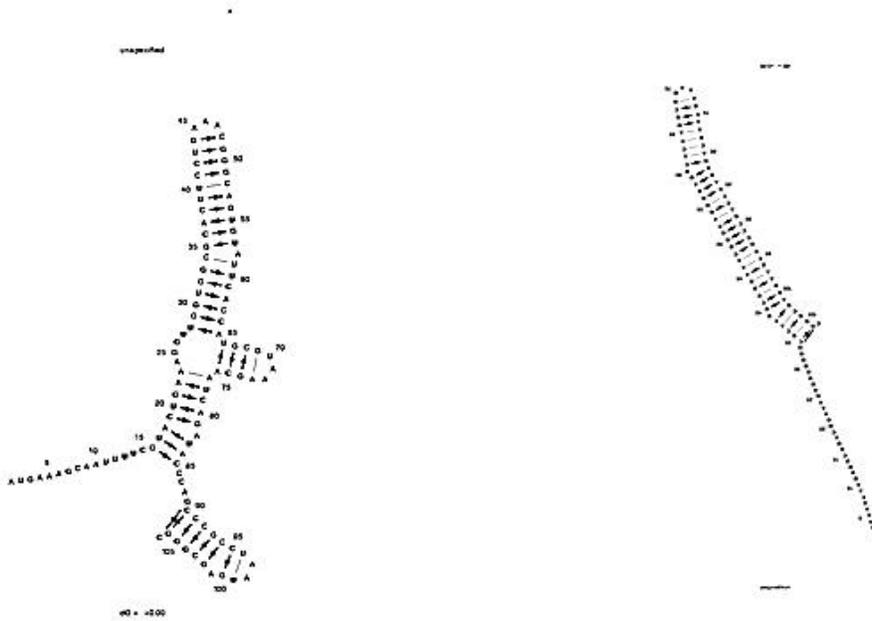


Figure 4.1 a) Modèle consensus de l'atténuateur de la transcription chez l'opéron *trp* chez *E.coli* tel que postulé par Yanofsky en 1981. b) Repliement de la même séquence d'atténuateurs en utilisant MC-Fold et le masque sur le segment contenant les codons régulateurs. Avec MC-Fold, l'apport de bases non canoniques représente 13 % des bases sur le repliement contenant la tige-boucle terminatrice et 34 % sur le repliement de la tige-boucle antiterminatrice.

À ce stade-ci, il convient de vérifier si la méthode proposée ici doit considérer les paires de bases non canoniques ou seulement les paires de bases du modèle consensus de Yanofsky. Dans ce cas, en considérant que plusieurs structures d'ARN déjà répertoriés dans Protein Data Bank présentent un nombre important des paires de bases non canoniques dans les structures, et que dans certains cas, ces appariements remplissent une fonction structurale importante dans la molécule (Stombaugh, Zirbel et al., 2009), et d'après les résultats présentés dans la section 2.2, nous avons choisi de considérer la contribution de bases non canoniques pour la prédiction de structures et d'utiliser néanmoins MC-Fold pour l'implémentation de la méthode.

Avec le premier essai de la méthode sur le repliement de l'atténuateur de l'opéron *trp* chez *E.coli* (NC\_000913) dans la section 2.2, nous avons observé la position du facteur sur la séquence au moment où le changement se produit. Ce dernier s'est produit quand le masque qui simule le ribosome se trouvait sur les codons régulateurs *trp* dans la séquence du gène leader (Figures 2.2 à 2.4). Ce résultat a été observé dans le MFE et dans le 85 % des repliements dans l'ensemble des résultats pour cette position du cofacteur. En poursuivant l'analyse, nous avons constaté que si le masque bloque une autre position à l'extérieur de la zone de codons régulateurs, la formation de la tige-boucle terminatrice est favorisée. Par la suite, grâce aux tests effectués sur un ensemble de séquences déjà documentées par Vitreschak, nous avons observé parmi les six opérons analysés, quatre seulement présentaient un changement de repliement évident.

Parallèlement, nous avons tenté d'appliquer la même analyse à une séquence présentant les mêmes éléments que l'atténuateur *trp* chez *E.coli*, soit une forte tige-boucle terminatrice, et qui pouvait être considéré comme un candidat intéressant. Mais d'après l'analyse, nous avons observé que la formation de la tige-boucle terminatrice est toujours privilégiée, aussi bien dans le MFE que dans l'ensemble de sous-optimaux (Figure 2.6). Dans ce cas, un candidat comme celui-ci, avec les éléments que définit un atténuateur, pouvait ne pas être retenu, considérant qu'un repliement comme celui recherché n'est pas possible en analysant cette séquence.

## 4.2 TA-Search

Avec la méthode destinée à identifier le changement de repliement, TA-Search résulte d'une combinaison de la recherche classique avec la méthode présentée dans le chapitre 2.

Durant l'implémentation de TA-Search sur de nouveaux génomes bactériens, deux problèmes se sont posés. Le premier problème était lié à la distance séparant le candidat du gène dans 3', qui est définie dans chaque vecteur utilisé par l'implémentation. Par définition, les distances ont été établies à partir de la recherche bibliographique sur des atténuateurs relevés dans la littérature. Dans les cas de certains nouveaux génomes, nous avons dû modifier la distance du gène pour obtenir des résultats. Le problème d'augmenter la distance était de générer un plus grand nombre de candidats, ce qui n'a pas été le cas (Tableau 3.6).

Le deuxième problème était lié à la notation de certains nouveaux génomes; TA-Search n'arrivait pas à identifier des candidats suggérés dans la littérature. En ajustant les paramètres des vecteurs de description, un grand nombre de candidats ont été identifiés. Dans le cas de l'opéron thr dans le génome de *Yersenia enterocolitica* (NC\_008800), à la suite de l'analyse de deux candidats prédits, ces derniers ont été écartés. Sachant que le gène de l'opéron thr a été identifié dans le génome de *Yersenia*, nous avons vérifié la notation faite pour le gène sur GenBank. Pour la séquence en aval du gène, nous avons constaté qu'un segment qui correspond à la définition de l'atténuateur se trouve noté de façon différente par rapport au reste des génomes examinés dans ce travail; le potentiel gène leader et le segment d'atténuation sont notés en un seul bloc avec la dénomination de protéine hypothétique. En analysant ce segment au moyen de la méthode définie dans le chapitre 2, nous avons vérifié que le changement de repliement du modèle d'atténuation existe, mais que TA-Search ne peut le prédire en raison de la mauvaise notation. Dans la version implémentée pour ce travail, TA-Search ne considère pas les candidats lorsque la notation du gène leader et le segment d'atténuation sont notés ensemble.

### 4.3 TA-Search et la prédiction chez les opérons leu et his

Durant l'étape de l'implémentation dans TA-Search, nous avons établi certaines limites qui permettent au logiciel de choisir les candidats potentiels. Une de ces limites est liée à la présence de deux repliements du modèle d'atténuation dans l'ensemble des solutions de MC-Fold. Pour renforcer les résultats, nous avons établi que si le changement d'un repliement d'une forme à l'autre se présentait dans 45 % de l'ensemble de sous-optimaux lorsque le masque se trouvait sur les codons régulateurs, le candidat devrait être retenu.

Pour les séquences de l'atténuateur de l'opéron his, en analysant l'ensemble des solutions, on constate que les deux structures sont présentes et que le changement se fait lorsque le masque se trouve sur les codons his. Toutefois, un problème demeure : seulement 3 % des repliements montrent le changement du modèle, la tige-boucle terminatrice et antiterminatrice. Ainsi, le changement du repliement semble être le résultat du hasard davantage que l'influence du cofacteur.

Dans une analyse plus détaillée des atténuateurs des opérons his et leu dans la littérature, nous avons remarqué certains détails où le repliement de ces deux opérons devra inclure autres cofacteurs que TA-Search ne considère pas dans la version présentée dans ce travail.

Avant le modèle consensus de Yanofsky de 1981, Johnston (Johnston et Roth, 1980) avait présenté un modèle de repliement de l'atténuateur his qui comportait deux repliements similaires au modèle de Yanofsky. La différence avec le modèle consensus est que le repliement de his présente trois tiges-boucle, pour la version terminatrice, et un repliement à deux tiges-boucle pour la version antiterminatrice. De plus, le modèle de Johnston considère que la séquence contient deux pauses transcriptionnelles, à la suite de la formation de chaque tige-boucle.

Dans ce travail, l'implémentation de la méthode d'analyse du repliement élimine la présence des pauses transcriptionnelles dans la prédiction, mais il semble que ces pauses doivent être considérées dans le cas de ces deux opérons, si l'on

considère que la présence des pauses en positions spécifiques renforce la disponibilité des segments libres à former des paires de bases.

#### **4.4 Conclusion**

La méthode et l'implémentation proposées dans ce travail permettent d'incorporer des éléments d'analyse supplémentaires dans la recherche de séquences candidates au modèle d'atténuation de la transcription chez les bactéries. De la même façon, la méthode d'analyse du repliement cherche à donner un outil pour généraliser la recherche des éléments de régulation comme les atténuateurs ou les riboswitches, en donnant la possibilité de détecter les changements de repliement qui définissent ce type de régulateur.

Un autre point important de la méthode d'analyse de repliement est qu'elle permet d'identifier spécifiquement les segments de fixation de cofacteurs cofacteurs qui influencent les changements de repliement. La flexibilité d'incorporer des sites de fixation multiples sur le repliement peut contribuer par la même occasion à identifier l'action potentielle de différents cofacteurs sans même spécifier lesquels.

Quant à l'implémentation du logiciel TA-Search, il se base sur la flexibilité donnée par les modèles de recherche en utilisant des vecteurs. Ainsi, la définition d'un élément à rechercher pourra être effectuée à partir d'un vecteur sans modifier le moteur de recherche lui-même. Grâce à ce concept, TA-Search semble pouvoir s'adapter à la recherche des autres régulateurs en tenant en compte de l'information disponible pour définir le vecteur de l'élément à rechercher. Dans cette version, TA-Search a été développé principalement pour retrouver des candidats liés à la transcription d'un gène, mais le logiciel pourra s'adapter aisément à la recherche de candidats liés au processus de traduction.

## Bibliographie

Abreu-Goodger, C. et E. Merino (2005). « RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements », *Nucleic Acids Research*, 33: W690-W692.

Alberts, B., A. Johnson, *et al.* (2004). *Biologie cellulaire de la cellule*, Flammarion.

Atkins, J., R. Gesteland, *et al.* (2011). *RNA worlds : from life's origins to diversity in gene regulation*, New York, Cold Spring Harbor.

Baird, N., N. Kulshina, *et al.* (2010). « Riboswitch function. Flipping the switch or tuning the dimer? », *RNA Biology*, 7(3): 328-332.

Bartkus, J. et J. Calvo (1991). « Transcription Attenuation-Mediated Control of leu Operon Expression: Influence of the Number of Leu Control Codons », *Journal of Bacteriology*, 173(5): 1634-1641.

Bastet, L., A. Dubé, *et al.* (2011). « New insights into riboswitch regulation mechanisms », *Molecular Microbiology*, 80: 1148-1154.

Becker, W., J. Hardin, *et al.* (2008). *The World of the cell Pearson*.

Berg, J., J. Tymoczko, *et al.* (2002). *Biochemistry 5th edition*. New York, W H Freeman.

Blouin, S. et D. Lafontaine (2009). « Riboswitches: Ancient and Promising genetic regulators », *ChemBioChem*, 10: 400-416.

Bolsover, S., J. Hyams, *et al.* (2006). *Biologie cellulaire et moléculaire*. Dunod, Wiley & Sons.

Breaker, R. (2010). « Riboswitch and the RNA World », *CSH Perspectives in Biology*, doi:10.1101/cshperspect.a003566.

Brion, P. et E. Westhof (1997). « Hierarchy and dynamics of RNA folding », *Ann. Rev. Biophys. Biolo. Struc.* 26: 113-137.

Browning, D. et S. Busby (2004). « The regulation of bacterial transcription initiation », *Nat. Rev. Microbiol.*, 2(1): 57-65.

Carter, P. et J. Calvo (1986). « Transcription attenuation in *Salmonella typhimurium*: The significance of rare leucine codons in the leu leader », *PNAS*, 83: 8127-8131.

Chan, C. et R. Landick (1989). « The Salmonella typhimurium his Operon Leader Region Contains an RNA Hairpin-dependent Transcription Pause Site », *The Journal of Biological Chemistry*, 264(34): 20796-20804.

Chen, J. et H. Umbarger (1991). « Specificity of Attenuation Control in the ilvGMEDA Operon of Escherichia coli K-12 », *Journal of Bacteriology*, 173(7): 2328-2340.

Crick, F. (1970). « Central Dogme of Molecular Biology », *Nature*, 227: 561-563.

d'Aubenton-Carafa, Y. et C. Thermes (1990). « Prediction of Rho-independent Escherichia coli Transcription Terminators », *J.Mol.Biol.* 216: 835-858.

Delorme, C. et P. Renault (1999). « Regulation of Expression of the Lactococcus lactis Histidine Operon », *Journal of Bacteriology*, 181(7): 2026-2037.

Ding, Y. et C. Lawrence (2003). « A statistical sampling algorithm for RNA secondary structure prediction », *Nucleic Acid Research*, 31: 7280-7301.

Du, H. et P. Babitzke (2000). « trp RNA-binding attenuation protein-5' stem-loop RNA interaction is required for proper transcription attenuation control of the Bacillus subtilis trpEDCFBA operon », *Journal of Bacteriology*, 182(7): 1819-1827.

Egli, M. et P. Pallan (2007). « Insights from crystallographic studies into the structural and pairing properties of nucleic acid analogs and chemically modified DNA and RNA oligonucleotides », *Annu.Rev.Biophys.Struc*, 36: 281-305.

Elf, J., O. Berg, *et al.* (2001). « Comparison of repressor and transcriptional attenuator systems for control of amino acid biosynthetic operons », *Jour. Mol. Biol.*, 313: 941-954.

Elf, J. et M. Ehrenberg (2005). « What makes ribosome-mediated transcriptional attenuation sensitive to amino acid limitation? », *PLos computational Biology*, 1(1): 14-23.

Ermolaeva, M. et S. Salzberg (2000). « Prediction of Transcription Terminators in Bacterial Genomes », *Jour. Mol.Biol.*, 301: 27-33.

Fisher, R. et C. Yanofsky (1983). « Mutation of the beta Subunit of RNA polymerase Alter Both Transcription Pausing and Transcription Termination in the trp Operon Leader Region in Vitro », *the Journal of Biological Chemistry*, 258(13): 8146-8150.

Fritsch, V. et E. Westhof (2000). « RNA folding : beyond Watson-Crick pairs », *Structure*, 8: R22-R65.

- Gardner, J. (1982). « Initiation, Pausing and Termination of Transcription in the threonine Operon Regulatory Region of Escherichia coli », *The Journal of Biological Chemistry*, 257(7): 3896-3904.
- Gollnick, P. et P. Babitzke (2002). « Transcription attenuation », *Biochimica et Biophysica Acta*, 1577: 240-250.
- Green, N., F. Grundy, *et al.* (2010). « The T box mechanism : tRNA as a regulatory molecule », *FEBS lett*, 584(2): 318-324.
- Grundy, F. et T. Henkin (2006). « From ribosome to riboswitch: Control of gene expression in bacteria by RNA structural rearrangements », *Crit.Rev.in Bioch.and.Mol.Biol.* 41: 329-338.
- Gusarov, I. et E. Nudler (1999). « The Mechanism of Intrinsic Transcription Termination », *Molecular Cell*, 3: 495-504.
- Hall, K. (2008). « RNA in MOTION », *Curr.Opin.Chem.Biol*, 12(6): 612-618.
- Haller, A., M. Souliere, *et al.* (2011). « The Dynamic Nature of RNA as Key to Understanding Riboswitches Mechanism », *Acc.Chem.Res.*, 44(12):1339-1348
- Henkin, T. (1996). « Control of transcription termination in prokaryotes », *Annu.Rev.Genet*, 30: 35-57.
- Henkin, T. et C. Yanofsky (2002). « Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions », *BioEssays*, 24: 700-707.
- Hofacker, I., M. Fekete, *et al.* (2002). « Secondary structure prediction for aligned RNA sequences », *Jour.Mol. Biol*, 319(5): 1059-1066.
- Hofacker, I., W. Fontana, *et al.* (1994). « Fast Folding and comparison of RNA Secondary Structures », *Monatshefte für Chemie*, 125(2): 167-188.
- Hofacker, I. et P. Standler (2006). « Memory efficient folding algorithms for circular RNA secondary structures », *Bioinformatics*, 22: 1172-1176.
- Ishihama, A. (2000). « Functional modulation of Escherichia coli Rna polymerase », *Ann.Rev.Microbiol*, 54: 499-518.
- Ishihama, A. (2010). « Prokaryote genome regulation: multifactor promoters, multitarget regulators and hierarchic networks », *FEMS Microbiology Reviews*, 34(5): 628-645.
- Johnston, H. et J. Roth (1980). « Model for regulation of the histidine operon of Salmonella », *PNAS*, 77(1): 508-512.

Kasai , T. (1974). « Regulation of the expression of histide operon in Salmonella typhimurium », *Nature*, 249(457): 523-527.

Keller, E. et J. Calvo (1979). « Alternative secondary structures of leader RNAs and the regulation of the trp,phe,his,thr and leu operon », *PNAS*, 76(12): 6186-6190.

Kingsford, C. et S. Salzberg (2007). « Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake », *Genome Biol.*, 8(R:22).

Klostermeier, D. (2011). « Single-molecule FRET reveals nucleotide-driven conformational changes in molecular machines and their link to RNA unwinding and DNA supercoiling », *Biochem.Soc.Trans.*, 39(2): 611-616.

Kolter, R. et C. Yanofsky (1982). « Attenuation in amino acid biosynthetic operons », *Ann.Rev.Genetics*, 16: 113-134.

Komissarova, N. et M. Kashlev (2002). « Shortening of RNA:DNA Hybrid in the Elongation Complex of RNA Polymerase IS a Prerequisite for Transcription Termination », *Molecular Cell*, 10: 1151-1162.

Korzheva, N. et A. Mustaev (2001). « Transcription elongation complex: structure and function », *Curr.Opin.Microbiol.*, 4: 119-125.

Kozak, M. (1999). « Initiation of translation in prokaryotes and eukaryotes », *Gene*, 8(234): 187-208.

Kuroda, M. et C. Yanofsky (1984). « Evidence for the transcription secondary structures predicted to regulate transcription attenuation in trp operon », *The Journal of Biological Chemistry*, 259(20): 12838-12843.

Lambert, A. et D. Gautheret (2004). « The ERPIN server: an interface to profile-based RNA motif identification », *Nucleic Acids Research* 32(Web Server issue): W160-W165.

Landick, R. and L. Phung (1990). "Replacement of Escherichia coli trp operon attenuation control codons alters operon expression." *Jour.Mol.Biol.* 216: 25-37.

Landick, R. et C. Yanofsky (1984). « Stability of an RNA secondary structure affects in vitro transcription pausing in the trp operon leader region », *The Journal of Biological Chemistry*, 259(18): 11550-11555.

Landick, R. et C. Yanofsky (1985). « Translation activates the paused transcription complex and restores transcription of trp operon leader region », *PNAS*, 82: 4663-4667.

Landick, R. et C. Yanofsky (1987). « Detection of transcription-pausing in vivo in the trp operon leader region », *PNAS*, 84: 1507-1511.

Lee, F. and C. Yanofsky (1977). "Transcription termination at the trp operon attenuators of *Escherichia coli* and *Salmonella typhimurium*: RNA secondary structure and regulation of termination." *PNAS* 74(10): 4365-4369.

Lee, J. et R. Gutell (2004). « Diversity of base-pair conformations and their occurrence in rRNA structure and RNA structural motifs », *Jour.Mol. Biol.*, 344(5): 1225-1249.

Lemay, J. et D. Lafontaine (2006). « Le riborégulateur adénine: Un nouveau mode de régulation génétique », *Medicine.Science*, 22: 1053-1059.

Lemieux, S. et F. Major (2002). « RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire », *Nucleic Acid Research*, 30(19): 4250-4263.

Leontis, N. et E. Westhof (2001). « Geometric nomenclature and classification of RNA base pairs », *RNA*, 7: 499-512.

Leontis, N. et E. Westhof (2003). « Analysis of RNA motifs », *Current Opinion in Structural Biology*, 13(3): 300-3008.

Lesnik, E. et D. Ecker (2001). « Prediction of rho-independent transcriptional terminators in *Escherichia coli* », *Nucleic Acids Research*, 29(17): 3583-3594.

Lewin, B. (1997). *Genes V*. New York: Oxford, Oxford University Press.

Lodish, H., A. Berk, *et al.* (2000). *Molecular Cell Biology*, W.H. Freeman.

Lyn, S. et J. Gardner (1987). « Specificity of the Attenuation Response of the Threonine Operon of *Escherichia coli* is Determined by the Threonine and Isoleucine Codons in the Leader Transcript », *Jour.Mol.Biol.*, 194: 59-69.

Lyubetskaya, E. et V. Lyubetsky (2003). « Search for alternative RNA secondary structures regulating expression of bacterial genes », *Molecular Biology*, 37(5): 707-715.

Lyubetsky, V. et S. Pirogov (2006). « Model of Gene Expression Regulation in Bacteria via Formation of RNA Secondary Structures », *Molecular Biology*, 40(3): 440.

Mackey, M. (2001). « Dynamic regulation of the tryptophan operon: A modeling study and comparison with experimental data », *PNAS*, 98(4): 1364-1369

- Major, F. et P. Thibault (2007). «RNA tertiary structure prediction». *Bioinformatics: From Genome to Therapies*, 491-539. T. Lengauer. Weinheim, Germany, Wiley-VCH.
- Malys, N. et J. McCarthy (2011). « Translation initiation: variations in the mechanism can be anticipated », *Cell. Mol. Life Sci.*, 68:991-1003.
- Mathews, D. (2006). « Revolutions in RNA Secondary Structure Prediction », *Journal of Molecular Biology*, 359: 526-532.
- Mathews, D. et D. Turner (2002). « Dynalign: an algorithm for finding the secondary structure common to two RNA sequences », *Jour.Mol. Biol*, 317(2): 191-203.
- McGraw, A. et P. Babitzke (2008). « Molecular basis of TRAP-5' SL RNA interaction in the Bacillus subtilis trp operon transcription attenuation mechanism », *RNA*, 15: 55-66.
- Means, J., S. Wolf, *et al.* (2007). « T box riboswitch antiterminator affinity modulated by tRNA structural elements », *Chem.Biol.Drug.Des*, 69(2): 139-135.
- Merino, E. et C. Yanofsky (2008). « Evolution of bacterial trp operons and their regulation », *Current Opinion in Microbiology*, 11: 78-86.
- Mitra, A. et V. Nagaraja (2009). « Occurrence, divergence and evolution of intrinsic terminators across Eubacteria », *Genomics*, 94: 110-116.
- Montagne, R. et R. Batey (2008). « Riboswitches: Emerging themes in RNA structure and function », *Annu.Rev.Biophys*, 37: 117-133.
- Mount, W. (2004). *Bioinformatics: Sequence and Genome Analysis*. New York, Cold Spring Harbor Laboratory Press.
- Mukherjee, S., M. Bansal, *et al.* (2006). « Conformational specificity of non-canonical base pairs and higher order structures in nucleic acids: crystal structure database analysis », *J. Comput Aided Mol Des*, 20: 629-645.
- Muller-Esterl, W. (2007). *Biochimie et biologie moléculaire*, Dunod.
- Naville, M. et D. Gautheret (2009). « Transcription attenuation in bacteria: theme and variations », *Brief.in Funct.Gen.and Proteomiques*, 8(6): 482-492.
- Naville, M. et D. Gautheret (2010). « Premature terminator analysis sheds light on a hidden world of bacterial transcriptional attenuation », *Genome Biol.*, 11(R97).
- Nawrocki, E., D. Kolbe, *et al.* (2009). « Infernal 1.0: Inference of RNA alignments », *Bioinformatics*, 25: 1335-1337.

- Nudler, E. et M. Gottesman (2002). « Transcription termination and anti-termination in E.coli » *Genes to Cells*, 7: 755-768.
- Olson, W., M. Esquerra, *et al.* (2009). « New information content in RNA base pairing deduced from quantitative analysis of high-resolution structures », *Methods*, 47: 177-186.
- Osbourn, A. et B. Field (2009). « Operons », *Cell. Mol. Life Sci.*, 66(23): 3755-3775.
- Oxender, D. et C. Yanofsky (1979). « Attenuation in the Escherichia coli tryptophan operon : Role of RNA secondary structure involving the tryptophan codon region », *Biochemistry*, 76(11): 5524-5528.
- Pan, T. et T. Sosnick (2006). « RNA Folding During Transcription », *Ann. Rev.Biophys.Biolo.Struc.*, 36: 161-175.
- Panina, E. et M. Gelfand (2001). « Regulation of Aromatic Amino Acid Biosynthesis in Gamma-Proteobacteria », *J. Mol.Microbiol.Biotechnol.* 3(4): 529-543.
- Parisien, M. and F. Major (2008). "The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data." *Nature* 452: 51-55.
- Peters, J., A. Vangeloff, *et al.* (2011). « Bacterial transcription terminators: The RNA 3'-end chronicles », *Jour.Mol.Biol.* DOI: 10.10106/j.jmb.2011.03.036
- Platt, T. et P. Farhham (1980). « Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription in vitro », *Nucleic Acids Research*, 9(3): 563-577.
- Ramesh, V. (1993). « NMR evidence for RNA stem-loop structure involved in the transcription attenuation of E.coli trp operon », *Nucleic Acids Research*, 21(23): 5485-5488.
- Reznikoff, W. (1972). « The operon revisited », *Ann.Rev.Genet.*, 6: 133-156.
- Richardson, J. (1993). « Transcription termination », *Critical Reviews on Biochemistry and Molecular Biology*, 28(1): 1-30.
- Roesser, J. et C. Yanofsky (1991). « The effects of leader peptide sequence and length on attenuation control of trp operon of E.coli. », *Nucleic Acids Research*, 19(4):795-800
- Sergarov, A. (2009). « Determination of riboswitch structures . Light at the end of the tunnel? », *RNA Biology*, 7(1): 98-103.

Siebert, S. et R. Backofen (2005). « MARNA : multiple alignment and consensus structure prediction of RNAs based on sequence structure comparison », *Bioinformatics*, 21(16): 3352-3359.

Sigel, R. et S. Gallo (2010). « Shaping RNA structures with metal ions and metal ion complexes », *Chimia*, 64(3): 126-131.

Sosnick, T. et T. Pan (2003). « RNA folding: models and perspectives », *Current Opinion in Structural Biology*, 13: 309-316.

Steitz, T. (2008). « A structural understanding of the dynamic ribosome machine », *Nat.Rev.Mol.Cell.Biol*, 9(3): 242-253.

Stroynowski, I. et C. Yanofsky (1983). « Transcription termination in vitro at the tryptophan operon attenuator is controlled by secondary structures in the leader transcript », *PNAS*, 80: 2206-2210.

Thompson, J., D. Higgins, *et al.* (1994). « CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice », *Nucleic Acid Research*, 11(22): 4673-4680.

Tinocco, I., G. Chen, *et al.* (2010). « RNA reactions one molecule at a time », *Cold Spring Harb Perspect Biol*, 2:a003624.

Tinocco Jr, I. et C. Bustamante (1999). « How RNA Folds », *Journal of Molecular Biology*, 293: 271-281.

Torarinsson, E., J. Havgaard, *et al.* (2007). « Multiple structural alignments and clustering of RNA sequences », *Bioinformatics*, 23(8): 926-932.

Tortora, Funke, *et al.* (2004). *Microbiology: An Introduction*, Pearson Education Inc.

Turnbough, C. (2011). « Regulation of gene expression by reiterative transcription », *Curr.Opin.Microbiol.* 14(2): 142-147.

Uptain, S. et M. Chamberlin (1997). « Basic mechanisms of transcription elongation and its regulation », *Annu.Rev.Biochem.*, 66: 117-172.

Vitreschak, A. et M. Gelfand (2000). « Computer analysis of control signals in bacterial genome, attenuator of operons of aromatic amino acid metabolism », *Molecular Biology*, 34(4): 461-467.

Vitreschak, A. et V. Lyubetsky (2004). « Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis », *FEMS Microbiology letters*, 234: 357-370.

Vitreschak, A., D. Rodionov, *et al.* (2004). « Riboswitches: the oldest mechanism for the regulation of gene expression? », *TRENDS in Genetics*, 20(1): 44-50.

Voet, D et J. Voet (2005). *Biochimie*, De Boeck.

Wachter, A. (2010). « Riboswitch-mediated control of gene expression in eukaryotes », *RNA Biology*, 7(1): 67-76.

Wessler, S. et J. Calvo (1981). « Control of leu Operon Expression in Escherichia coli by a Transcription Attenuation Mechanism », *Journal of Molecular Biology*, 149: 579-597.

Wilson, K. et P. Hippel (1995). « Transcription termination at intrinsic terminators: The role of RNA hairpin », *PNAS*, 92: 8793-8797.

Winkler, W. et R. Breaker (2005). « Regulation of bacterial gene expression by riboswitches », *Annu.Rev.Microbiol.* 59: 487-517.

Xayaphoummine, A. et H. Isambert (2007). « Encoding folding paths of RNA switches », *Nucleic Acids Research*, 35(2): 614-622.

Xiong, J. (2006). *Essential Bioinformatics*. Cambridge, Cambridge University Press.

Yamell, W. S. et J. W. Roberts (1999). « Mechanism of intrinsic transcription termination and antitermination », *Science*, 284: 611-615.

Yanofsky, C. (1981). « Attenuation in the control of expression of bacterial genomes », *Nature*, 26(289): 751-758.

Yanofsky, C. (2000). « Transcription attenuation: once viewed as a novel regulatory strategy », *Journal of Bacteriology*, 182(1): 1-8.

Yanofsky, C. (2004). « The different roles of tryptophan transfer RNA in regulation trp operon expression in E.coli versus B.subtilis », *Trends in Genetics*, 20(8): 367-374.

Yanofsky, C. et R. Fisher (1984). « Analysis of the Requirements for Transcription Pausing in the Tryptophan Operon », *Jour.Mol.Biol.*, 182: 397-409

Yusupova, G., M. Yusupov, *et al.* (2001). « The path of messenger RNA through the ribosome », *Cell*, 106: 233-241.

Zirbel, C., J. Spomer, *et al.* (2009). « Classification and energetics of the base-phosphate interactions in RNA », *Nucleic Acid Research*, 37(15): 4898-4918.

Zucker, M. et P. Stiegler (1981). « Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information », *Nucleic Acid Research* 10(9): 133-148.

## Annexe A

### Search-MFE

Pour récupérer les résultats des repliements soumis à MC-Fold par la méthode décrite dans la section 2.1, nous avons développé un logiciel en Java, Search-MFE, qui cherche dans un ensemble de fichiers de résultats de MC-Fold le repliement MFE. Pour étudier l'effet du repliement en considérant un ou plusieurs cofacteurs, le logiciel récupère les résultats en considérant l'ordre de génération des instructions soumises à MC-Fold. Search-MFE permet de corroborer la progression du cofacteur sur la séquence et de mettre en évidence la présence d'un changement significatif du repliement

Par exemple, dans le cas du ribosome sur l'ARNm de l'atténuateur de l'opéron *trp* chez *Escherichia coli* NC\_00913, la méthode a généré un ensemble d'instructions en utilisant le masque de type 2. Une fois les prédictions terminées, le logiciel cherche les fichiers qui contiennent l'information de l'opéron *trp* pour NC\_000913 en utilisant le masque 2. Dans ce cas-ci, Search-MFE a lu 39 fichiers et il a récupéré la notation parenthésée et l'énergie de la structure MFE de chaque fichier, en considérant le déplacement du masque sur la séquence. Le résultat final de Search-MFE sur l'opéron *trp* (Figure 2.2) permet d'observer en ligne 5 (en rouge) un changement du repliement par rapport aux lignes 1-4. En ligne 9 (en bleu), autre changement est observé, similaire aux repliements des lignes 1-4. Dans ce cas, on a pu observer que lorsque la séquence a changé son repliement, le masque se trouvait sur le segment contenant les deux codons régulateurs *trp* de l'opéron leader. Par la suite, Search-MFE a été lancé sur tous les résultats des séquences du tableau 2.3, puis il a permis d'identifier les séquences où le repliement est affecté, plus précisément où le modèle de consensus d'atténuation se présente.

Dans certains cas, le résultat de Search-MFE ne reflétait pas le changement dans le repliement selon le modèle. Dans ces cas, nous avons effectué une révision manuelle de tous les résultats obtenus avec MC-Fold. Dans le cadre de ce travail, nous avons considéré que si le changement se présentait dans le premier 5 % de l'ensemble des solutions avec une persistance de plus de 45 % dans l'ensemble, la séquence serait un très fort candidat pour le modèle d'atténuation. À partir de ces résultats, nous avons implémenté une modification sur Search-MFE qui permet d'indiquer au logiciel de regarder d'autres résultats sur l'ensemble des solutions (en pourcentage) afin de vérifier si le changement peut être observé.

## Annexe B

### Séquences des opérons bactériens, segments leaders et segments d'atténuation

#### Trp Opéron

>NC\_000913 Escherichia coli str. K-12 substr. MG1655 chromosome, complete genome

ATGAAAGCAATTTTCGTA  
ACTGAAAGGTTGGTGGCGCACTTCTGAAACGG  
GCAGTGTATTCACCATGCGT  
AAAGCAATCAGATACCCAGCCCGCCTAATGAGCGGGCTTTTTTTTGA

>NC\_002655 Escherichia coli O157:H7 str. EDL933 chromosome, complete genome

ATGAAAGCAATTTTCGTA  
ACTGAAAGGTTGGTGGCGCACTTCTGAAACGG  
GCAGTGTATTCACCATGCGT  
AAAGCAATCAGATACCCAGCCCGCCTAATGAGCGGGCTTTTTTTTGA

>NC\_002695 Escherichia coli O157:H7 str. Sakai, complete genome

ATGAAAGCAATTTTCGTA  
ACTGAAAGGTTGGTGGCGCACTTCTGAAACGG  
GCAGTGTATTCACCATGCGT  
AAAGCAATCAGATACCCAGCCCGCCTAATGAGCGGGCTTTTTTTTGAACA  
AAA

>NC\_003197 Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 chromosome, complete genome

ATGGCAGCGACATTTGCATTACACGGTTGGTGGCGCACTTCTGATAGCG  
GGCGGTGTATGAACAGCTGT  
AATCAGCCAAACGATACCCGGCCCGCCTGTTAAGCGGGCTTTTTTTTGA

>NC\_004088 Yersinia pestis KIM 10 chromosome, complete genome

ATGAAGACTTCCCTGATTTCTTACTGCGTTGGTGGCATATCTCCCTCTCT  
CGGGCGATGTAATCACGCA  
TATCCGTCATCAGACAGTGCAGATTGCTTCAGCCCGCTAATAGCGGGTTTT  
TTTATG

>NC\_004337 Shigella flexneri 2a str. 301, complete genome

ATGAAAGCAATTTTCGTA  
ACTGAAAGGTTGGTGGCGCACTTCTGAAACGG  
GCAGTGTATTCACCATGCGT  
AAAGCAATCAGATACCCAGCCCGCCTAATGAGCGGGCTTTTTTTT

>NC\_004741 *Shigella flexneri* 2a str. 2457T, complete genome

ATGAAAGCAATTTTCGTAAGGTTGGTGGCGCACTTCCTGAAACGG  
GCAGTGTATTCACCATGCGT  
AAAGCAATCAGATACCCAGCCCGCCTAATGAGCGGGCTTTTTTTT

>NC\_006840 *Vibrio fischeri* ES114 chromosome I, complete genome

ATGTCACAGCTTAACAGACAACAAGCAATCTTAAAAACATCAGTGGTAAG  
CTCACGATGGTGGCACTACT  
TTGGGTAGTGTGCTAATTTATTTTTGACATAAATCCACTACCCGCA  
ACTAAGCGGGTTTTTTTAT

>NC\_007606 *Shigella dysenteriae* Sd197, complete genome

ATGAAAGCAATTTTCGTAAGGTTGGTGGCGCACTTCCTGAAACGG  
GCAGTGTATTCACCATGCGT  
AAAGCAATCAGATACCCAGCCCGCCTAATGAGCGGGCTTTTTTTT

>NC\_007613 *Shigella boydii* Sb227, complete genome

ATGAAAGCAATTTTCGTAAGGTTGGTGGCGCACTTCCTGAAACGG  
GCAGTGTATTCACCATGCGT  
AAAGCAATCAGATACCCAGCCCGCCTAATGAGCGGGCTTTTTTTT

>NC\_009708 *Yersinia pseudotuberculosis* IP 31758, complete genome

ATGAAACTTCCCTGATTTCTTACTGCGTTGGTGGCATATCTCCCTCTCT  
CGGGCGATGTAATCACGCA  
TATCCGTCATCAGACAGTGCAGATTGCTTCAGCCCGCTAATAGCGGGTTTT  
TTTAT

>NC\_011740 *Escherichia fergusonii* ATCC 35469, complete genome

ATGAAAGCGATTTTCGTAAGGTTGGTGGCGCACTTCCTGAAACGG  
GCAGTGTATTCACCATGCGT  
AAAGCAACCAGATACCCAGCCCGCCTAATGAGCGGGCTTTTTTTTGAAC

>NC\_012125 *Salmonella enterica* subsp. *enterica* serovar Paratyphi C strain  
RKS4594 chromosome, complete genome

ATGGCAGCGACATTTGCATTACACGGTTGGTGGCGCACTTCCTGATAGCG  
GGCGGTGTATGAACAGCTGT  
AATCAGCCAAACGATACCCGGCCCGCCTGTTAAGCGGGCTTTTTTTT

>NC\_013365 *Escherichia coli* O111:H- str. 11128, complete genome

ATGAAAGCAATTTTCGTAAGGTTGGTGGCGCACTTCCTGAAACGG  
GCAGTGTATTCACCATGCGT  
AAAGCAATCAGATACCCAGCCCGCCTAATGAGCGGGCTTTTTTTTGAACA  
AAATTAGA

>NC\_013716 *Citrobacter rodentium* ICC168, complete genome  
 ATGAAAGCAACATTCGTTCTGCACGGTTGGTGGCGCACTTCCTGATCTCG  
 GGCAGTGTATTGTGCATGCT  
 TACCGCACCTGGATACCCGGCCCGCCTGCAAGCGGGCTTTTTTTTGTAGCA

### **Phe Opéron**

>NC\_000913 *Escherichia coli* str. K-12 substr. MG1655 chromosome, complete genome  
 ATGAAACACATAACCGTTTTTCTTCGCATTCTTTTTTACCTTCCCCTGAATGG  
 GAGGCGTTTCGTCGTGTG  
 AAACAGAATGCGAAGACGAACAATAAGGCCTCCCAAATCGGGGGGCCTT  
 TTTTATT

>NC\_003198 *Salmonella enterica* subsp. *enterica* serovar Typhi str. CT18 chromosome, complete genome  
 ATGAAGCTCACCCGGTTTTTCTTCGCATTCTTTTTTATCTTCCCCTGACCGG  
 GAGGCGTTTCGTCATGTG  
 ATAAAGAATGCGAAGACGAACAAGAAGGCCTCCCCACCGGGAGGCCTT  
 TTTTATTG

>NC\_004088 *Yersinia pestis* KIM 10 chromosome, complete genome  
 GTGATTCTCAATAAAGTGTTTTTCTTCGCATTCTTTTTTACCTTCCCCTAAT  
 TGGGAGGCGTTTCGTCGT  
 GTGATAAAGAATGCGAAGACGAATAAATATCCTCCGGCATAGCCGGAGG  
 TTTTTCATATG

>NC\_004741 *Shigella flexneri* 2a str. 2457T, complete genome  
 ATGAAACACATAACCGTTTTTCTTCGCATTCTTTTTTACCTTCCCCTGAATGG  
 GAGGCGTTTCGTCGTGTG  
 AAACAGAATGCGAAGACGAACAATAAGGCCTCCCAAATCGGGGGGCCTT  
 TTTTATTG

>NC\_004337 *Shigella flexneri* 2a str. 301, complete genome  
 ATGAAACACATAACCGTTTTTCTTCGCATTCTTTTTTACCTTCCCCTGAATGG  
 GAGGCGTTTCGTCGTGTG  
 AAACAGAATGCGAAGACGAACAATAAGGCCTCCCAAATCGGGGGGCCTT  
 TTTTATTG

>NC\_007613 *Shigella boydii* Sb227, complete genome  
 ATGAAACACACACCGTTTTTCTTCGCATTCTTTTTTACCTTCCCCTGAATG  
 GGAGGCGTTTCGTCGTGTG  
 AAACAGAATGCGAAGACGAACAATAAGGCCTCCCAAACCGGGGGGCCTT  
 TTTTATTGAT

>NC\_009708 *Yersinia pseudotuberculosis* IP 31758, complete genome  
 GTGATTCTCAATAAAGTGTTTTCTTCGCATTCTTTTTTACCTTCCCCTAAT  
 TGGGAGGCGTTTCGTCGT  
 GTGATAAAGAATGCGAAGACGAACAAGAAAGCCTCCCAAGCGGGAGGCT  
 TTTTTATGTATA

>NC\_011740 *Escherichia fergusonii* ATCC 35469, complete genome  
 ATGAAACAAACACCGTTCCTCTTCGCATTCTTTTTTACCTTCCCCTGAACG  
 GGAGGCGTTTCGTCGTGTG  
 AAACAGAATGCGAAGACGAACAATAAGGCCTCCCAAATCGGGGGGCCTT  
 TTTTATAG

>NC\_013364 *Escherichia coli* O111:H- str. 11128, complete genome  
 ATGAAACACACACCGTTTTCTTCGCATTCTTTTTTACCTTCCCCTGAATG  
 GGAGGCGTTTCGTCGTGTG  
 AAACAGAATGCGAAGACGAACAATAAGGCCTCCCAAATCGGGGGGCCTT  
 TTTTATTGAT

### **Thr Opéron**

>NC\_000913 *Escherichia coli* str. K-12 substr. MG1655 chromosome, complete genome  
 ATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGG  
 TAACGGTGCGGGCTGACGCG  
 TACAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTT  
 TCGACC

>NC\_002695 *Escherichia coli* O157:H7 str. Sakai, complete genome  
 ATGAAACGCATTAGCACCACCATTACCACCACCATCACCACCACCATCAC  
 CATTACCATTACCACAGGTA  
 ACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAGCCCGCACCTG  
 ACAGTGCGGGCTTTTTTTTCG

>NC\_003198 *Salmonella enterica* subsp. *enterica* serovar Typhi str. CT18  
 chromosome, complete genome  
 ATGAACCGCATCAGCACCACCACCATTACCACCATCACCATTACCACAGG  
 TAACGGTGCGGGCTGACGCG  
 TACAGGAAACACAGAAAAAAGCCCGCACCTGAACAGTGCGGGCTTTTTTT  
 TCGACC

>NC\_004088 *Yersinia pestis* KIM 10 chromosome, complete genome  
 ATGAGACAGACAGATAACATGCGATACATCAGCCTGAACACAACGATTA  
 TTACCACCACCGAAACCACAG  
 GTTACGGGGCGGGCTGACGCGTACAGGAAACAATAGAAAAAAGCCCGCA  
 CCTAGACAGTGCGGGCTTTTT  
 TTTTCAATC

>NC\_004741 *Shigella flexneri* 2a str. 2457T, complete genome

ATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGG  
TAACGGTGCGGGCTGACGCG  
TACAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTT  
CGACCA

> NC\_006840 *Vibrio fischeri* ES114 chromosome I, complete genome

ATGTTTATTATCAGCTCTATGATTATTACCACAACAATTACCGGCACCATG  
GTTGCTGGGGTGGGCTGAT  
ACGTATAAGTATATGAAAAAGCCTGCTCTGATTACCAGAGCAGGCTTTTT  
TTTGCAAATTTGTTACT

>NC\_009648 *Klebsiella pneumoniae* subsp. *pneumoniae* MGH 78578 chromosome,  
complete genome

ATGAATCGCATCGGCATGATTACCACCATTATCACCACCACCATTACCAC  
AGGTAACGGTGCGGGCTGAC  
GCGTACAGGAAACACAGAAAAAAGCCCGCACCTGAACAGTGCGGGCTTT  
TTTTTTG

>NC-009708 *Yersinia pseudotuberculosis* IP 31758, complete genome

ATGAGACAGACAGATAACATGCGATACATCAGCCTGAACACAACGATTA  
TTACCACCACCGAAACCACAG  
GTTACGGGGCGGGCTGACGCGTACAGGAAACAACAGAAAAAAGCCCGCA  
CCTAGACAGTGCGGGCTTTTT  
TTTTCA

>NC\_010658 *Shigella boydii* CDC 3083-94 chromosome, complete genome

ATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGG  
TAACGGTGCGGGCTGACGCG  
TACAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTT  
CGACCA

>NC\_011283 *Klebsiella pneumoniae* 342, complete genome

ATGAATCGCATCGGCATGATTACCACCATTATCACCACCACCATTACCAC  
AGGTAACGGTGCGGGCTGAC  
GCGTACAGGAAACACAGAAAAAAGCCCGCACCTGAACAGTGCGGGCTTT  
TTTTTTGAC

>NC\_012125 *Salmonella enterica* subsp. *enterica* serovar Paratyphi C strain  
RKS4594 chromosome, complete genome

ATGAACCGCATCAGCACCACCACCATTACCACCATCACCATTACCACAGG  
TAACGGTGCGGGCTGACGCG  
TACAGGAAACACAGAAAAAAGCCCGCACCTGAACAGTGCGGGCTTTTTTT  
TCGACC

>NC\_013716 *Citrobacter rodentium* ICC168, complete genome  
 ATGAAACGCATCAGCACCACCATTATTACCACCACCACCATTACCACAGG  
 TAACGGTGC GGGCTGACGCG  
 TACAGGAAATACAGAAAAAAGCCCGCACCTGAACAGTGCGGGCTTTTTTT  
 TCGAC

### Ilv Opéron

>NC\_000913 *Escherichia coli* str. K-12 substr. MG1655 chromosome, complete genome  
 ATGACAGCCCTTCTACGAGTGATTAGCCTGGTCGTGATTAGCGTGGTGGT  
 GATTATTATCCCACCGTGCG  
 GGGCTGCACTTGGACGAGGAAAGGCTTAGAGATCAAGCCTTAACGAACT  
 AAGACCCCGCACCGAAAGGT  
 CCGGGGGTTTTTTTTG

>NC\_002695 *Escherichia coli* O157:H7 str. Sakai, complete genome  
 ATGACAGCCCTTCTACGAGTGATTAGCCTGGTCGTGATTAGCGTGGTGGT  
 GATTATTATCCCACCGTGCG  
 GGGCTGCACTTGGACGAGGAAAGGCTTAGAGATCAAGCCTTAACGAACT  
 AAGACCCCGCACCGAAAGGT  
 CCGGGGGTTTTTTTTGACCT

>NC\_003197 *Salmonella enterica* subsp. *enterica* serovar Typhimurium str. LT2 chromosome, complete genome  
 ATGACAGCCCTTCTACGAGTGATTAGCCTGGTCGTGATTAGCGTGGTGGT  
 GATTATTATCCCACCGTGCG  
 GGGCTGCACTTGGACGAGGAAAGGCTTAGAAATCAAGCCTTAACGAACT  
 AAGACCCCGCACCGAAAGGT  
 CCGGGGGTTTTTTTTG

>NC\_003910 *Colwellia psychrerythraea* 34H chromosome, complete genome  
 ATGCACATATTTAACTCAATTATTATAAACCTTCTCGTCGTGGTATTCTT  
 AAATCATCTCGAGGGTTTT  
 GTTGAGCTAAAAAAAAGCATTCAATCAAATAACAAACCCTCGCTTCGCAA  
 GAACCGAGGGTTTTCTTTTT  
 AGT

>NC\_004088 *Yersinia pestis* KIM 10 chromosome, complete genome  
 ATGAAAGCAATTCTCCAAGTGATTAACCTAGTCCTAATTAGCGTGGTGGT  
 GATTATTATCCCACCGTGCG  
 GGGCTGCACTTGGACGAAGAAAGGCATAAAAATCAAGCCTTAATCTCAA  
 GAGAACCCCGCACCGAAAGG  
 TCCGGGGTTTTTTTATT

>NC\_004741 *Shigella flexneri* 2a str. 2457T, complete genome

ATGACAGCCCTTCTACGAGTGATTAGCCTGGTCGTGATTAGCGTGGTGGT  
GATTATTATCCCACCGTGCG  
GGGCTGCACTTGGACGAGGAAAGGCTTAGAGATCAAGCCTTAACGAACT  
AAGACCCCGCACCGAAAGGT  
CCGGGGGTTTTTTTGACCTA

>NC\_007384 *Shigella sonnei* Ss046 chromosome, complete genome

ATGACAGCCCTTCTACGAGTGATTAGCCTAGTCGTGATTAGCGTGGTGGT  
GATTATTATCCCACCGTGCG  
GGGCTGCACTTGGACGAGGAAAGGCTTAGAGATCAAGCCTTAACGAACT  
AAGACCCCGCACCGAAAGGT  
CCGGGGGTTTTTTTG

>NC\_007606 *Shigella dysenteriae* Sd197, complete genome

ATGACAGCCCTTCTACGAGTGATTAGCCTGGTCGTGATTAGCGTGGTGGT  
GATTATTATCCCACCGTGCG  
GGGCTGCACTTGGACGAGGAAAGGCTTAGAGATCAAGCCTTAACGAACT  
AAGACCCCGCACCGAAAGGT  
CCGGGGGTTTTTTTGACCT

>NC\_007613 *Shigella boydii* Sb227, complete genome

ATGACAGCCCTTCTACGAGTGATTAGCCTGGTCGTGATTAGCGTGGTGGT  
GATTATTATCCCACCGTGCG  
GGGCTGCACTTGGACGAGGAAAGGCTTAGAAATCAAGCCTTAACGAACT  
AAGACCCCGCACCGAAAGGT  
CCGGGGGTTTTTTTGA

>NC\_010554 *Proteus mirabilis* HI4320 chromosome, complete genome

ATGAATACTATCGTCCTAGTGATTAGCCTAATTATTGTCAGCGTGGTGGTG  
ATTATTAACCCACTGTGCG  
GGGTTGCACTAGGACGAAGAAAGGCTAACTAACTAGCCGGATCTCAGA  
AAAACCCCGCACCGAAAGGC  
GCGGGGGTTTTTTTGAGCCAAGGATTTTAAAGAG

>NC\_011283 *Klebsiella pneumoniae* 342, complete genome

ATGACAGCCCTTCTACGAGTGATTAGCCTGGTCGTGATTAGCGTGGTGGT  
GATTATTATCCCACCGTGCG  
GGGCTGCACTTGGACGAGGAAAGGCTTAAGAATCAAGCCTTAACGAACT  
AAGACCCCGCACCGAAAGGT  
CCGGGGGTTTTTTTATAAC

>NC\_011740 *Escherichia fergusonii* ATCC 35469 chromosome, complete genome  
 ATGACAGCCCTTCTACGAGTGATTAGCCTGGTCGTGATTAGCGTGGTGGT  
 GATTATTATCCCACCGTGCG  
 GGGCTGCACTTGGACGAGGAAAGGCTTAGAGATCAAGCCTTAACGAACT  
 AAGACCCCCGCACCGAAAGGT  
 CCGGGGGTTTTTTTTTGA

>NC\_012125 *Salmonella enterica* subsp. *enterica* serovar Paratyphi C strain  
 RKS4594 chromosome, complete genome  
 ATGACAGCCCTTCTACGAGTGATTAGCCTGGTCGTGATTAGCGTGGTGGT  
 GATTATTATCCCACCGTGCG  
 GGGCTGCACTTGGACGAGGAAAGGCTTAGAAATCAAGCCTTAACGAACT  
 AAGACCCCCGCACCGAAAGGT  
 CCGGGGGTTTTTTTTTGGCCTGAAAA

>NC\_013716 *Citrobacter rodentium* ICC168, complete genome  
 ATGACAGCCCTTCTACGAGTGATTAGCTTGGTCGTGATTAGCGTGGTGGT  
 GATTATTATCCCACCGTGCG  
 GGGCTGCACTTGGACGAGGAAAGGCTTAGAAATCAAGCCTTAACGAACT  
 AAGACCCCCGCACCGAAAGGT  
 CCGGGGTTTTTTTTTGACCTT

### **Leu Opéron**

>NC\_000913 *Escherichia coli* str. K-12 substr. MG1655 chromosome, complete genome  
 ATGACTCACATCGTTCGCTTTATCGGTCTACTACTACTAAACGCATCTTCT  
 TTGCGCGGTAGACGAGTGA  
 GCGGCATCCAGCATTAAAGCCAGCACGCAGTCAAACAAAAAACCCGCGCC  
 ATTGCGCGGGTTTTTTTATG

>NC\_002655 *Escherichia coli*-O157:H7 str. EDL933 chromosome, complete genome  
 2655  
 ATGACTCACATCGTTCGCTTTATCGGTCTACTACTACTAAACGCATCTTCT  
 TTGCGCGGTAGACGAGTGA  
 GCGGCATCCAGCATTAAAGCCAGCACGCAGTCAAACAAAAAACCCGCGCC  
 ATTGCGCGGGTTTTTTTA

>NC\_003197 *Salmonella enterica* subsp. *enterica* serovar Typhimurium str. LT2  
 chromosome, complete genome  
 ATGTCACATATCGTTCGTTTCACTGGGCTACTACTACTCAACGCATTTATT  
 GTGCGCGGTAGACCGGTGG  
 GCGGCATTCAACATTAAGTCAGCTCGAAGTCAAACAAAAAACCCGCGCCGTT  
 GCGCGGGTTTTTTTATGCCT  
 GAC

>NC\_003198 *Salmonella enterica* subsp. *enterica* serovar Typhi str. CT18  
chromosome, complete genome  
ATGTCATATATCGTTTCGTTTCACTGGGCTACTACTACTCAACGCATTTATT  
GTGCGCGGTAGACCGGTGG  
GCGGCATTCAACATTAAGTCAGCTCGAAGTCAAACAAAACCCGCGCCGTT  
GTGCGGGTTTTTTTATGCCT  
GA

>NC\_003910 gi|71277742:4428233-4428390 *Colwellia psychrerythraea* 34H  
chromosome, complete genome  
ATGACTCACATGACCAAATTGATTAAGCACATTAAAAACAACCTCCTCCT  
CGCACTCATCGTGCGCGGCT  
AGGTTTTGTGCTTCTTCAAAGTTTAAATATTGAGAACTAAGTAGAAAT  
AAAATAAAAACCCGCGCTAAC  
AATCGCGGGTTTTTTTAT

>NC\_004741 *Shigella-flexneri* 2a str. 2457T, complete genome  
ATGACTCACATCGTTTCGCTTTATCGGTCTACTACTACTAAACGCATCTTCT  
TTGCGCGGTAGACGAGTGA  
GCGGCATCCAGCATTAAAGCCAGCACGCAGTCAAACAAAAAACCCGCGCC  
ATTGCGCGGGTTTTTTTAT

>NC\_006840 *Vibrio fischeri* ES114 chromosome I, complete genome  
ATGAAAACAGTATCTTCTTTTGCTCTAAACCTCCTCCTGATCGTGAACATC  
TCACGCGGGTAGGTTATGG  
ACCAAAAGAAACCATAAACCAAATTTTAAAACCCGCTAACCATGGCGGG  
TTTTTTTATGG

>NC\_006905 *Salmonella enterica* subsp. *enterica* serovar Choleraesuis str. SC-B67  
chromosome, complete genome  
ATGTCACATATCGTTTCGTTTCACTGGGCTACTACTACTCAACGCATTTATT  
GTGCGCGGTAGACCGGTGG  
GCGGCATTCAACATTAAGTCAGCTCGAAGTCAAACAAAACCCGCGCCGTT  
GCGCGGGTTTTTTTATGCCT

>NC\_007606 *Shigella-dysenteriae* Sd197, complete genome  
ATGACTCACATCGTTTCGCTTTATCGGTCTACTACTAAACGCATCTTCTTTG  
CGCGGTAGACGAGTGAGCG  
GCATCCAGCATTAAAGCCAGCACGCAGTCAAACAAAAAACCCGCGCCATT  
GCGCGGGTTTTTTTATG

>NC\_007613 *Shigella-boydii* Sb227, complete genome  
ATGACTCACATCGTTTCGCTTTATCGGTCTACTACTACTAAACGCATCTTCT  
TTGCGCGGTAGACGAGTGA  
GCGGCATCCAGCATTAAAGCCAGCACGCAGTCAAACAAAAAACCCGCGCC  
ATTGCGCGGGTTTTTTTAT

>NC\_007384 *Shigella sonnei* Ss046 chromosome, complete genome  
 ATGACTCACATCGTTTCGCTTTATCGGTCTACTACTACTAAACGCATCTTCT  
 TTGCGCGGTAGACGAGTGA  
 GCGGCATCCAGCATTAAAGCCAGCACGCAGTCAAACAAAAAACCCGCGCC  
 ATTGCGCGGGTTTTTTTATGC  
 CCGAAGCGA

>NC\_011283 *Klebsiella-pneumoniae* 342, complete genome  
 ATGATTTCGCACCGCTCGTATCACCAGCCTACTACTACTAAACGCATGCCA  
 TTTGCGCGGTAGGCTGTTGG  
 GCGACGTTTACGCGTTAAGTCATCTTCCAGCAAGACTATAAAACCCGCGCC  
 TTGGCGCGGGTTTTTTTAT

>NC\_011740 *Escherichia-fergusonii*-11740 ATCC 35469, complete genome  
 ATGACTCACATCGTTTCGTTTTAACGGTCTACTACTACTAAACGCATCTTCT  
 TTGCGCGGTAGACTGGTGA  
 GCGGCATTCAGAATTAAGTCAGCACTCAGTCAAACAAAAAACCCGCGCC  
 ATTGCGCGGGTTTTTTTAT

>NC\_013716 *Citrobacter-rodentium* ICC168, complete genome  
 ATGATCCGTATTACCCGCTTCAACGGTCTACTACTACTAAACGCATTTTCG  
 GTGCGCGGTAGACTGGTGA  
 ACGGCATTCAGCATTAAAGTCGACGACAGTCAGATAAAAAAACCCGCGCC  
 GCTGCGCGGGTTTTTTTATGC  
 C

### His Opéron

>NC\_000913 *Escherichia-coli* str. K-12 substr. MG1655 chromosome, complete genome  
 ATGACACGCGTTCAATTTAAACACCACCATCATCACCATCATCCTGACTA  
 GTCTTTCAGGCGATGTGTGC  
 TGGAAGACATTCAGATCTTCCAGTGGTGCATGAACGCATGAGAAAGCCCC  
 CGGAAGATCACCTTCCGGGG  
 GCTTTTTTATTGCG

>NC\_002695 *Escherichia coli* O157:H7 str. Sakai, complete genome  
 ATGACACGCGTTCAATTTAAACAGCACCATCATCACCATCATCCTGACTA  
 GTCTTTCAGGCGATGTGTGC  
 TGGAAGACATTCAGATCTTCCAGTGGTGCATGAACGCATGAGAAAGCCCC  
 CGGAAGATCACCTTCCGGGG  
 GCTTTTTTATTGCGCGGTTG

>NC\_003198 *Salmonella-enterica* subsp. *enterica* serovar Typhi str. CT18  
chromosome, complete genome  
ATGACACGCGTTCAATTTAAACATCACCATCATCCTGACTAGTCTTTCAGG  
CGATGTGTGCTGGAAGACA  
TTCAGATCTTCCAGTGGTGCATGAACGCATGAGAAAGCCCCCGGAAGATC  
ATCTTCCGGGGGCTTTTTTT  
TTG

>NC\_003910 *Colwellia-psychrerythraea* 34H chromosome, complete genome  
ATGATTATTAACCTGAATCGCCGACATCATCACCATCACTACAACCTAGCC  
CATTACAGGTTTACTTAATAA  
GTAAATTACTGAAGGGCAACTGGCCTTTCAGTGGTTTAAAAAAGCATAAT  
ACATCCTCTTAATAGAAGTA  
ATTGTAATTGCTAAACCCCCGAAAGAGTCACGACTCTCTCGGGGGTTTTTC  
GCATTTT

>NC\_007613 *Shigella-boydii* Sb227, complete genome  
ATGACACGCGTTCAATTTAAACACCACCATCATCACCATCATCCTGACTA  
GTCTTTCAGGCGATGTGTGC  
TGGAAGACATTCAGATCTTCCAGTGGTGCATGAACGCATGAGAAAGCCCC  
CGGAAGATCACCTTCCGGGG  
GCTTTTTTATT

>NC\_012125 *Salmonella-enterica*-P subsp. *enterica* serovar Paratyphi C strain  
RKS4594 chromosome, complete genome  
ATGACACGCGTTCAATTTAAACACCACCATCATCCTGACTAGTCTTTCAGG  
CGATGTGTGCTGGAAGACA  
TTCAGATCTTCCAGTGGTGCATGAACGCATGAGAAAGCCCCCGGAAGATC  
ATCTTCCGGGGGCTTTTTTT  
TGG

>NC\_011283 *Klebsiella-pneumoniae* 342, complete genome  
ATGAACCGCGTTCAATTTAAACACCACCATCATCACCATCATCCTGACTA  
GTCTTTCAGGCGATGTGTGC  
TGGAAGACATTTGGATCTTCCAGTGGTGCATGAACGCGCAGAGAGCCCCC  
GGAAGATTCACCTTCCGGGGG  
CTTTTTTTTGG

>NC\_004337 *Shigella-flexneri* 2a str. 301, complete genome  
ATGACACGCGTTCAATTTAAACACCACCATCATCCTGACTAGTCTTTCAGG  
CGATGTGTGCTGGAAGACA  
TTCAGATCTTCCAGTGGTGCATGAACGCATGAGAAAGCCCCCGGAAGATC  
ACCTTCCGGGGGCTTTTTTA  
TTGCGCGGTTGTT

>NC\_013716 *Citrobacter-rodentium* ICC168, complete genome

ATGACACGCGTTCAGTTTAAACACCACCATCATCACCATCATCCTGACTA  
GTCTTTCAGGCGATGTGTGC  
TGGAAGACGTTTCAGATCTTCCAGTGGTGCATGAACGCATGAGAAAGCCCC  
CGGAAGATCGTCTTCCGGGG  
GCTTTTTTTTGCTGC

>NC\_011740 *Escherichia-fergusonii* ATCC 35469, complete genome

ATGACACGCGTTC AATT TAAACACCACCATCATCACCATCATCCTGACTA  
GTCTTTCAGGCGATGTGTGC  
TGGAAGACATTCAGATCTTCCAGTGGTGCATGAACGCATGAGAAAGCCCC  
CGGAAGATCACCTTCCGGGG  
GCTTTTTTATTGC

>NC\_007384 *Shigella-sonnei* Ss046, complete genome

ATGACACGCGTTC AATT TAAACACCACCATCATCACCATCATCCTGACTA  
GTCTTTCAGGCGATGTGTGC  
TGGAAGACATTCAGATCTTCCAGTGGTGCATGAACGCATGAGAAAGCCCC  
CGGAAGATCACCTTCCGGGG  
GCTTTTTTATTGC

## **Annexe C**

### **1 – Sélection des éléments et valeurs pour l'implémentation de TA-Search sur le mécanisme d'atténuation de la transcription chez les bactéries**

La sélection des éléments et des valeurs a été effectuée à partir de la revue de la littérature (voir le chapitre 2). Les éléments répertoriés pour tous les opérons sont :

Codons d'initiation

Codons stop

Codons régulateurs pour chaque opéron

Segment de haute concentration en T, poly-T

Distance maximale entre le codon d'initiation et le codon de stop pour le gène leader

Distance maximale entre le codon d'initiation du gène leader et le segment poly-T

Distance maximale entre le dernier nucléotide du segment poly-T et le premier gène de l'opéron

Pour déterminer les valeurs, nous avons retenu les atténuateurs déjà répertoriés pour chaque opéron, puis nous avons déterminé les distances maximales en utilisant l'alignement de séquences (Figure C.1). Parallèlement, divers paramètres de classification ont été établis pour l'implémentation telle que:

Le nombre de fois où les motifs de régulation sont présentés dans le gène leader

La fenêtre de recherche pour la poly-T

La limite de l'énergie de repliement de la tige-boucle terminatrice

La concentration en CG de la tige-boucle terminatrice

La limite pour le score de d'Aubenton-Carafa-Lesnik

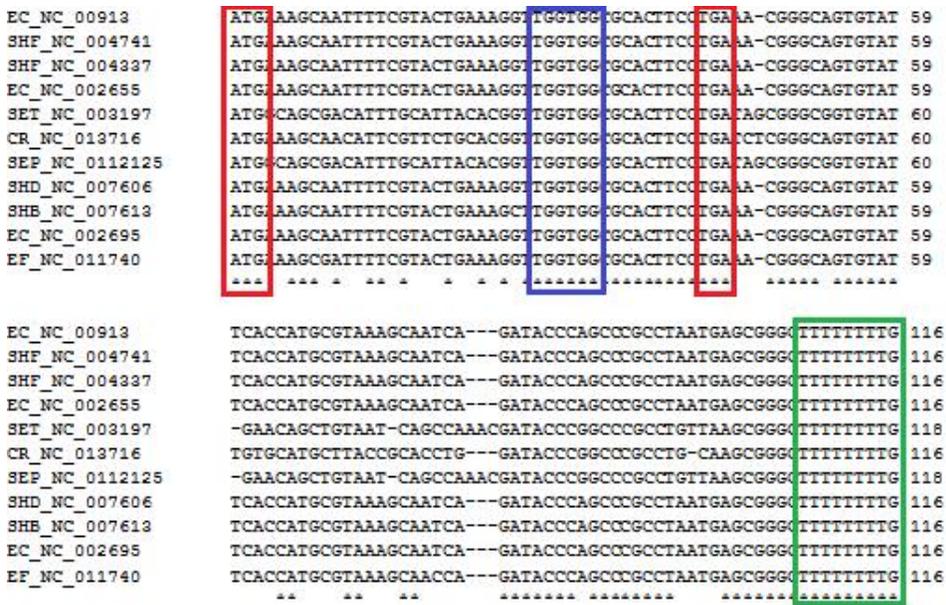


Figure C.1 : L’alignement des séquences de trpL et le segment d’atténuation jusqu’à trpE permet d’identifier la persistance des codons d’initiation et stop (rouge), les codons régulateurs (bleu) et la poly-T (vert).

Les valeurs de ces éléments et les paramètres pour chaque type d’opéron inclus dans le fichier descripteur sont résumés dans le tableau C.1.

	Trp	Phe	Leu	His	Ilv	Thr
Motif	W	FFXFFF	L	H	LLXVIXLVVIXVVVIII	TTITTTITTT
Repeat	2	1	4	5	1	1
Distan.Init/Stop	70	90	90	80	100	95
Distan.Init/ T	160	180	180	180	180	180
Fenêtre	15	15	15	15	15	15
Énergie Cutoff	-12	-12	-12	-12	-12	-12
Concentration de GC	0.15	0.15	0.15	0.15	0.15	0.15
Distan. gene en 3’	200	250	250	150	250	250
Limite de Lesnik	3.6	3.6	3.6	3.6	3.6	3.6

Tableau C.1 : Les valeurs des éléments qui définissent chaque vecteur pour chaque opéron sont le résultat de l’analyse de la littérature, l’alignement de séquences et les paramètres qui permettent de limiter la recherche des candidats.

## 2 - Description du fichier de candidats après l'analyse génomique

À la fin de l'analyse génomique, le logiciel génère le fichier contenant les candidats. Chaque candidat est décrit à partir de:

- la position du codon d'initiation, le codon stop;
- la position de la poly-T;
- la séquence en nucléotides du candidat;
- le pourcentage en contenu de CG de la tige-boucle terminatrice;
- l'identification du gène proche en 3';
- le score de d'Aubenton Carafa-Lesnik pour la poly-T;
- la distance du candidat au gène proche en 3';
- la séquence de la potentielle tige-boucle terminatrice;
- la notation parenthésée du repliement de la tige-boucle terminatrice;
- l'énergie de la tige-boucle terminatrice calculée par MC-Fold;
- le score de Carafa pour la tige-boucle terminatrice.

Les candidats sont ordonnés de façon ascendante par la distance du candidat au gène dans 3', et dans le cas où deux candidats présentent une même distance, le score de d'Aubenton-Carafa-Lesnik est ajouté à la distance pour ordonner les candidats.

L'exemple suivant présente les candidats tels que rapportés par l'analyse génomique pour l'opéron trp sur le brin 3'→5' en NC\_000913.

## CAND\_NC\_000913\_W\_N.txt

2220217 2220168 2220099 2220090

ATGGAAGACATTTTCCACTCCTTGTGGTGGCGAAAAAATATCACAACTAGAATAGCGCC  
 TTGTATAACAATTCAGCGAGTGAATATTTTGAAAATTCAGACAATAAACTGAGCATTCTTT  
 TTCTCT 0.2222222222222222 946682 4.722254244000001 80.0  
 AUUUUGAAAAUUCAGACAAUAAACUGAGCA .....((((((((.....))))))))) -14.05  
 65.8211013567543

2781557 2781484 2781414 2781405

GTGTCACTACCAAAGAAAAAGGGGCTACGTTTTTCATGTAACCCCTTGATTTATTTGGTGGG  
 CTGGCAGTGAATGATTGGTGTGTAATTGGTTGAAATATTCATCTAATGTCTGGTTCAA  
 CTTTCCTTAGCTGCCTGAAAGTTTTTTGAT 0.4166666666666667 947250 4.899401244000001  
 80.0  
 UGUCUGGUUCAACUUUCCUUAGCUGCCUGA ..((((..((((..))))))))) -22.73  
 118.4691732577067

3396647 3396628 3396548 3396539

TTGCTTACCTGGTGGCGCTGAAATACCAGCTTTTCCGCAACCTCGCATTATGGCAGCAGGC  
 GCTGGTCGTCATGTTGCTTTCGCTGGTGGTGGATATTATTGTTTTCT 0.3333333333333333  
 947753 4.063724244 140.0  
 UCGUCAUGUUGCUUUCGUGGUGGUGGAUA (((((((((((.....))))))))).. -20.66  
 61.95614806051367

3396677 3396628 3396548 3396539

TTGGCGTACGCGTATTGGCGATGAGCATCATTGCTTACCTGGTGGCGCTGAAATACCAGC  
 TTTCCGCAACCTCGCATTATGGCAGCAGGCGCTGGTCGTCATGTTGCTTTCGCTGGTGGT  
 GGATATTATTGTTTTCT 0.3333333333333333 947753 4.063724244 140.0  
 UCGUCAUGUUGCUUUCGUGGUGGUGGAUA (((((((((((.....))))))))).. -20.66  
 61.95614806051367

Un deuxième fichier contenant l'ensemble des instructions pour MC-Fold est généré par TA-Search. Les paramètres choisis pour MC-Fold sont :

Séquence = [séquence du candidat]

Top = 1000

Espace d'exploration = 20%

Numéro maximum de « bulges » = 5

Masque = [option 2]

De cette façon, pour chaque séquence du candidat, un ensemble des instructions est généré en utilisant le masque de type 2, qui parcourt la séquence en bloquant la séquence de 5'→3', tel que défini dans le chapitre 2 (Figure 2.1).

## Annexe D

### Base de données gene2accession

Notation de génomes telle que présentée en NCBI (<ftp://ftp.ncbi.nih.gov/gene/>) , les éléments soulignés sont utilisés pour l'implémentation de TA-Search :

tax\_id:

the unique identifier provided by NCBI Taxonomy  
for the species or strain/isolate

GeneID:

the unique identifier for a gene

status:

status of the RefSeq if a refseq, else '-'

RNA nucleotide accession.version:

may be null (-) for some genomes

RNA nucleotide gi:

the gi for an RNA nucleotide accession, '-' if not applicable

protein accession.version:

will be null (-) for RNA-coding genes

protein gi:

the gi for a protein accession, '-' if not applicable

genomic nucleotide accession.version:

may be null (-)

genomic nucleotide gi:

the gi for a genomic nucleotide accession, '-' if not applicable

start position on the genomic accession:

position of the gene feature on the genomic accession,  
'-' if not applicable  
position 0-based

end position on the genomic accession:

position of the gene feature on the genomic accession,  
'-' if not applicable  
position 0-based

orientation:

orientation of the gene feature on the genomic accession,  
'?' if not applicable

assembly:

the name of the assembly  
'-' if not applicable

**OXFORD UNIVERSITY PRESS LICENSE  
TERMS AND CONDITIONS**

Dec 05, 2011

---

---

This is a License Agreement between Maria Abella ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

**All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.**

License Number	2802720781150
License date	Dec 05, 2011
Licensed content publisher	Oxford University Press
Licensed content publication	Briefings in Functional Genomics
Licensed content title	Transcription attenuation in bacteria: theme and variations:
Licensed content author	Magali Naville, Daniel Gautheret
Licensed content date	03/01/2010
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	Riboswitches : le cas des atténuateurs de la transcription du type terminator/antiterminator chez les bactéries
Publisher of your work	n/a
Expected publication date	Dec 2011
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Total	0.00 USD

[Terms and Conditions](#)

**STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL  
FROM AN OXFORD UNIVERSITY PRESS JOURNAL**

1. Use of the material is restricted to the type of use specified in your order details.
2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.
3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.
4. No alteration, omission or addition is made to the material without our written consent.

Permission must be re-cleared with Oxford University Press if/when you decide to reprint.

5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.

6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from [www.oxfordjournals.org](http://www.oxfordjournals.org) Should there be a problem clearing these rights, please contact [journals.permissions@oxfordjournals.org](mailto:journals.permissions@oxfordjournals.org)

7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

12. Other Terms and Conditions:

v1.4

**If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK500678254.**

**Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.**

**Make Payment To:**

**Copyright Clearance Center**

**Dept 001**

**P.O. Box 843006**

**Boston, MA 02284-3006**

**For suggestions or comments regarding this order, contact RightsLink Customer Support: [customercare@copyright.com](mailto:customercare@copyright.com) or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.**

**Gratis licenses (referencing \$0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.**

# NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Dec 05, 2011

This is a License Agreement between Maria Abella ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

**All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.**

License Number	2802720020209
License date	Dec 05, 2011
Licensed content publisher	Nature Publishing Group
Licensed content publication	EMBO reports
Licensed content title	The G[middot]U wobble base pair
Licensed content author	Gabriele Varani and William H McClain
Licensed content date	Jul 1, 2000
Type of Use	reuse in a thesis/dissertation
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
Figures	figure 1 Watson-Crick GC and AU base pairs
Author of this NPG article	no
Your reference number	figure 1
Title of your thesis / dissertation	Riboswitches : le cas des atténuateurs de la transcription du type terminator/antiterminator chez les bactéries
Expected completion date	Dec 2011
Estimated size (number of pages)	100
<b>Total</b>	<b>0.00 USD</b>
Terms and Conditions	

## Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to reproduce this material for this purpose, and for no other use, subject to the conditions below:

1. NPG warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to Nature Publishing Group and does not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested indicates that it was reprinted or adapted by NPG with permission from another source, then you should also seek permission from that source to reuse the material.
2. Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to the work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version. Where print permission has been granted for a fee, separate permission must be obtained for any additional, electronic re-use (unless, as in the case of a full paper, this has already been accounted for

during your initial request in the calculation of a print run). NB: In all cases, web-based use of full-text articles must be authorized separately through the 'Use on a Web Site' option when requesting permission.

3. Permission granted for a first edition does not apply to second and subsequent editions and for editions in other languages (except for signatories to the STM Permissions Guidelines, or where the first edition permission was granted for free).
4. Nature Publishing Group's permission must be acknowledged next to the figure, table or abstract in print. In electronic form, this acknowledgement must be visible at the same time as the figure/table/abstract, and must be hyperlinked to the journal's homepage.
5. The credit line should read:  
Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)  
For AOP papers, the credit line should read:  
Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

**Note: For republication from the *British Journal of Cancer*, the following credit lines apply.**

Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication) For AOP papers, the credit line should read:  
Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

6. Adaptations of single figures do not require NPG approval. However, the adaptation should be credited as follows:

Adapted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

**Note: For adaptation from the *British Journal of Cancer*, the following credit line applies.**

Adapted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

7. Translations of 401 words up to a whole article require NPG approval. Please visit <http://www.macmillanmedicalcommunications.com> for more information. Translations of up to a 400 words do not require NPG approval. The translation should be credited as follows:

Translated by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication).

**Note: For translation from the *British Journal of Cancer*, the following credit line applies.**

Translated by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

We are certain that all parties will benefit from this agreement and wish you the best in the use of this material. Thank you.

Special Terms:

v1.1

**If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK0. Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.**

**Make Payment To:  
Copyright Clearance Center**

**Dept 001**  
**P.O. Box 843006**  
**Boston, MA 02284-3006**

For suggestions or comments regarding this order, contact RightsLink Customer Support: [customercare@copyright.com](mailto:customercare@copyright.com)  
or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

**Gratis licenses (referencing \$0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.**

---

---