

الجمهورية الجزائرية الديمقراطية الشعبية  
République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي  
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique



جامعة الإخوة منتوري قسنطينة I  
Frères Mentouri Constantine I University  
Université Frères Mentouri Constantine I

Université Frères Mentouri Constantine  
Faculté des Sciences de la Nature et de la Vie  
Département de biologie appliquée

جامعة الاخوة منتوري قسنطينة  
كلية علوم الطبيعة والحياة  
قسم بيولوجية التطبيقية

Mémoire présenté en vue de l'obtention du diplôme de Master

Domaine : Sciences de la Nature et de la Vie  
Filière : Sciences biologiques  
Spécialité : *Bioinformatique*

N° d'ordre :  
N° de série :

Intitulé :

---

**Essai de définition des motifs structuraux de la famille  
des ARNnc**

---

Présenté par : CHEKIREB Rayen et MEROUANI Aicha

En date du 18/06/2023

Jury :

|              |                         |                                           |
|--------------|-------------------------|-------------------------------------------|
| Président    | KELLO Kamel             | Université Frères Mentouri, Constantine 1 |
| Examinatrice | GHERBOUDJ Amira         | Université Frères Mentouri, Constantine 1 |
| Encadreur    | HAMIDECHI M. Abdelhafid | Université Frères Mentouri, Constantine 1 |

Année universitaire  
2022 – 2023

## *Remerciements*

Toute notre gratitude et nos remerciements vont à Allah le plus puissant, qui nous a accordé la force, le courage et la volonté nécessaires pour élaborer ce modeste travail.

Nous remercions chaleureusement tous les membres du jury d'avoir bien accepté d'examiner notre travail.

Nous tenons à exprimer notre profonde gratitude envers notre encadreur, le Professeur HAMIDECHI Mohamed Abdelhafid, pour sa guidance, son expertise et son soutien tout au long de ce mémoire. Ses conseils éclairés et sa disponibilité ont grandement contribué à l'aboutissement de ce travail. Nous sommes reconnaissantes de l'opportunité qui nous a été donnée de travailler sous sa supervision. Enfin, nous remercions également l'ensemble de nos enseignants ayant contribué dans nos apprentissages durant toute notre formation de Bioinformatique au sein du département de Biologie Appliquée.

## *Dédicace*

*À ma famille, à mon père Houcine, ma mère Halima, ma sœur Maroua, mes deux frères Skander et Tyed à mes amis Aicha, Imen, Lyna et Hamza .*

*Je tiens à vous dédier ce mémoire avec une profonde gratitude et un amour infini. Votre soutien inconditionnel, votre encouragement constant et votre présence bienveillante ont été des piliers essentiels tout au long de ce parcours académique.*

*Que cette dédicace soit le témoignage de ma gratitude éternelle envers vous tous. Vous êtes les véritables héros de cette réussite et je suis fière de pouvoir partager ce mémoire avec vous.*

*Avec tout mon amour et ma reconnaissance.*

**RAYEN**

## *Dédicace*

*A l'homme de ma vie, mon exemple éternel, celui qui s'est toujours sacrifié pour me voir réussir , à toi mon père MED SALAH .*

*A ma maman KARIMA , pour son amour et qu'elle m'a toujours accordé en témoignage de ma reconnaissance envers sa confiance, ses sacrifices .*

*A la mémoire de ma grand-mère MMA que dieu la garde dans son vaste paradis.*

*A ma sœur ARIDJ et mes frères CHARIB et HAMZA , mon grand père et ceux qui ont partagés avec moi tous les émotions lors de la réalisation de ce travail ils m'ont chaleureusement supporté et encouragé tout au long de mon parcours .*

*A RAYANE , chère amie avant d'être binome .*

*Je tiens à vous dédier ce mémoire avec une profonde gratitude et un amour infini*

**ARCHA**

## Table des matières :

|                                                                               |   |
|-------------------------------------------------------------------------------|---|
| 1. Introduction .....                                                         | 1 |
| 2. Recherche bibliographique.....                                             | 2 |
| 2.1. Les ARN non codants.....                                                 | 2 |
| 2.1.1 Différents types d'ARN non codants.....                                 | 2 |
| 2.1.2 Caractéristiques des longs ARN non codants.....                         | 3 |
| 2.1.3 Fonction des longs ARN non codants.....                                 | 4 |
| 2.1.4 Les longs ARN non codants dérivés des séquences<br>amplificatrices..... | 5 |
| 2.2. Notion de motifs.....                                                    | 7 |
| 2.2.1 Importance des motifs.....                                              | 8 |
| 2.2.2 Le motif nucléique.....                                                 | 9 |

|         |                                                        |    |
|---------|--------------------------------------------------------|----|
| 2.2.2.1 | La séquence consensus.....                             | 9  |
| 2.2.2.2 | Table de fréquence, matrice de pondération.....        | 10 |
| 2.2.3   | Le motif protéique.....                                | 11 |
| 2.2.4   | Description des motifs par une séquence consensus..... | 11 |
| 2.2.5   | Les logiciels de recherche des motifs consensus.....   | 12 |
| 2.2.6   | Les bases de données de motifs.....                    | 12 |
| 2.2.6.1 | Les bases de motif nucléiques.....                     | 12 |
| 2.2.6.2 | Les bases de motif protéiques.....                     | 13 |
| 2.3.    | Alignement.....                                        | 13 |
| 2.3.1   | Définition.....                                        | 13 |
| 2.3.2   | Alignement global.....                                 | 14 |
| 2.3.3   | Alignement local.....                                  | 14 |
| 2.3.4   | Alignement multiple.....                               | 15 |
| 2.3.5   | Le score.....                                          | 16 |
| 2.4.    | L'ARN.....                                             | 17 |
| 2.4.1   | Définition.....                                        | 17 |
| 2.4.2   | ARNr.....                                              | 18 |
| 2.4.3   | ARNr16s.....                                           | 19 |
| 2.4.4   | ARNt.....                                              | 20 |
| 3.      | Matériel et méthodes.....                              | 24 |
| 3.1     | Les séquences des ARNr.....                            | 24 |
| 3.2     | Alignement multiple de séquences.....                  | 25 |
| 3.3     | Détection et définition des motifs.....                | 26 |
| 4.      | Résultat.....                                          | 26 |
| 4.1     | L'alignement.....                                      | 26 |
| 4.2     | Détection et définition de motif.....                  | 27 |
| 5.      | Conclusion.....                                        | 34 |
| 6.      | Les références.....                                    | 35 |

## **Liste des figures :**

|    |                                                                            |    |
|----|----------------------------------------------------------------------------|----|
| 1. | Figure 01 : Exemple d'alignement multiple avec un motif.....               | 7  |
| 2. | Figure 02 : Alignement global .....                                        | 13 |
| 3. | Figure 03 : Alignement local.....                                          | 14 |
| 4. | Figure 04 : Exemple d'un score.....                                        | 15 |
| 5. | Figure 05 : Un schéma montrant la structure de base d'un ribosome.....     | 17 |
| 6. | Figure 06 : Schéma illustrant la forme en trèfle d'une molécule d'ARNt.... | 20 |

### **Liste des tableaux :**

1. Tableau 01 : L'alphabet IUPAC .....9
2. Tableau 02 : Expression régulière d'une séquence consensus du doigt de zinc réalisé à partir d'un alignement multiple .....10
3. Tableau 03 : Les bases de motif nucléiques .....11
4. Tableau 04 : Les bases de motif protéiques.....12

|                                                                     |    |
|---------------------------------------------------------------------|----|
| 5. Tableau 05 : Les séquences d'ARNr 16s.....                       | 22 |
| 6. Tableau 06 : Les séquences d'ARNt.....                           | 23 |
| 7. Tableau 07 : Les outils utilisés et leur description .....       | 25 |
| 8. Tableau 08 : Une portion d'alignement des séquences d'ARNr16s... | 27 |

### **Liste des abréviations :**

**ANRIL** : Antisense Non-Coding RNA in the INK4 Locus

**ARNlnc** : ARN long non codant

**ARNnm** : ARN non codant

**ARNr** : ARN ribosomique

**ARNsn** : Small nuclear RNA (petits ARN nucléaires)

**ARNsno** : Small nucleolar RNA (petits ARN nucléolaires)

**ARNt** : ARN de transfert

**GAS5** : Growth Arrest-Specific 5

**HOTAIR** : HOX transcript antisense RNA

**IUPAC** : Union internationale de chimie pure et appliquée

**LINE** : Long Interspersed Nuclear Elements (éléments nucléaires intercalés longs)

**miRNA** : MicroRNA (micro-ARN)

**pARNi** : small interfering RNA (petits ARN interférents)

**piRNA** : Piwi-Associated small RNA (piARN)

**SINE** : Short Interspersed Nuclear Elements (éléments nucléaires intercalés courts)

**siRNA** : Small interfering RNA (petits ARN interférents)

## **Résumé**

Les ARN non codants (ARNnc) sont des transcrits non traduits qui jouent des rôles clés et variés dans la cellule. Notre travail décrit une analyse basée sur la construction et le calcul de matrices (comptage, fréquence et poids-position) pour définir les signatures moléculaires existantes au sein d'une même famille d'ARNnc. L'objectif de ce travail est de détecter un motif structural d'ARNr16s et d'ARNt à partir d'un



alignement multiple, et d'en calculer le score correspondant. L'analyse des scores calculés permet de confirmer la présence de motifs communs et de classer des séquences génomiques au sein de familles de gènes leur correspondant.

**Mots clés :** ARNnc, Matrice Poids-Position, Alignement Multiple, Motif.

### **Summary:**

Non-coding RNAs (ncRNAs) are untranslated transcripts that play key and varied roles in the cell. Our work describes an analysis based on the construction and calculation of matrices (counting, frequency and weight-position) to define the existing molecular signatures within the same ncRNA family. The aim of this work is to detect a structural motif of 16s rRNA and tRNA from a multiple alignment, and to calculate the corresponding score. The analysis of the calculated scores makes it possible to confirm the presence of common motifs and to classify genomic sequences within families of genes corresponding to them.

**Keywords :** ncRNA, Position Weight Matrix, Multiple Sequence Alignment, Motif.

## المخلص :

الحمض النووي الريبسي غير المشفر (ncRNAs) عبارة عن نصوص غير مترجمة تلعب أدواراً رئيسية ومتنوعة في الخلية. يصف عملنا تحليلاً يعتمد على بناء وحساب المصفوفات (العد والتكرار وموضع الوزن) لتحديد التواقع الجزئية الموجودة داخل نفس عائلة ncRNA. الهدف من هذا العمل هو اكتشاف الحافز البنيوي للـ 16 rRNA و tRNA من محاذاة متعددة ، ولحساب النتيجة المقابلة. يتيح تحليل الدرجات المحسوبة تأكيد وجود العناصر المشتركة وتصنيف التسلسلات الجينومية داخل عائلات الجينات المقابلة لها.

*الكلمات المفتاحية: ARNnc , مصفوفة وزن الموضع , توازن متعدد , نمط.*



## INTRODUCTION

L'avènement spectaculaire de la biotechnologie a permis l'acquisition de données génomiques (et protéomiques) massives impossibles à gérer ou à utiliser sans systèmes informatiques pertinents.

Parmi les challenges à relever pour la bioinformatique, est notamment la classification et l'annotation de toutes ces données. Cela se manifeste dans le domaine agroalimentaire, environnemental et surtout de santé publique.

La classification des données issues du Next Generation Sequencing (NGS) est l'un des défis majeurs qui fait face aux bioinformaticiens car nécessite des applications de plus en plus précises et crédibles afin de faire face à un tel tsunami génomique. Ainsi, les bases de données spécifiques et généralistes sont dans l'obligation de structurer leurs données en les classifiant dans des ébiologiquement distinctes afin d'assurer une meilleure gestion voire une exploitation plus efficace.

Dans ce travail, nous proposons une approche basée sur les calculs statistiques programmés en langage python, afin de définir des signatures (motifs) moléculaires représentatives de familles de gènes RNA 16S et RNAt prises comme exemple de travail.

# 1. RECHERCHE BIBLIOGRAPHIQUE

## 1.1 Les ARN non codants :

La séquence non codante est une région d'un gène qui ne détermine pas directement la séquence en acides aminés de la protéine correspondante. Un ARN non codant (ARNnm), produit lors de la transcription de l'ADN, n'est pas traduit en protéine par les ribosomes.

Il existe aujourd'hui une grande variété d'ARN non codants, et il semble que la majorité des génomes soient transcrits en ARN, tandis que seulement 1,2 % de l'ADN du génome humain, par exemple, est traduit en protéines.

La partie non codante du génome est activement transcrite, donnant lieu à la production de milliers d'ARN non codants de différentes longueurs, qui sont capables de réguler les réseaux de gènes.

Les ARN non codants peuvent être classés en fonction de leur taille. Les petits ARN non codants ont une taille inférieure à 200 nucléotides et comprennent les petits ARN inférieurs (siRNA), les micro-ARN (miRNA) et les piARN (piwi-Associated small RNA).

D'autre part, les transcrits de plus de 200 nucléotides sont appelés ARN longs non codants (ARNlnc).

### 1.1.1 Différents types d'ARN non codants

Les ARN ribosomiques sont des ARN qui forment les ribosomes et qui sont associés à des protéines. Chez les eucaryotes, on distingue différents types d'ARNr tels que l'ARNr 5S, 5.8S, 18S et 28S.

Les ARN de transfert sont des ARN impliqués dans la traduction et qui se lient spécifiquement à un codon de trois nucléotides, en associant un acide aminé spécifique déterminé par le code génétique.

Les SINE sont des ARN transcrits à partir de petits éléments nucléaires intercalés, qui peuvent exploiter le système de réplication des LINE. Chez l'homme, les séquences Alu sont les principaux SINE produits.

Les petits ARN nucléaires (ARNsn) sont des ARN présents dans le noyau des eucaryotes et qui sont impliqués dans l'épissage des ARNm.

Les petits ARN nucléolaires (ARNsno) sont des ARN présents dans le nucléole et qui contribuent à la maturation des ARNr.

Les petits ARN interférents (pARNi) sont des ARN dérivés d'un génome étranger qui, par appariement complémentaire, induisent la dégradation des ARNm correspondants.

Les micro-ARN (miARN) sont des ARN codés par la cellule elle-même (contrairement aux pARNi) et qui régulent l'expression de certains gènes en modulant la traduction des ARNm complémentaires.

Les larges ARN non codants intergéniques (Linc RNA) sont des ARN de grande taille qui peuvent réguler l'expression des gènes en s'ancrant probablement à des régulateurs transcriptionnels complémentaires situés au niveau des promoteurs. Un exemple d'un tel ARN est MALAT1.

Les ARN interagissant avec les protéines piwi sont des ARN qui inhibent la réplication des rétroéléments endogènes.

### 1.1.2 Caractéristiques des longs ARN non codants

Les longs ARN non codants sont des ARN d'une taille généralement supérieure à environ 200 nucléotides qui ne possèdent que peu ou pas de capacité à coder pour des protéines. Le potentiel codant de ces ARN peut être évalué en analysant divers paramètres tels que les cadres de lecture, les codons d'initiation et de terminaison, ainsi que les sites d'épissage. La classification des longs ARN non codants peut également se baser sur leur localisation dans le génome. Par rapport au gène codant le plus proche, un long ARN non codant peut être intragénique (partiellement chevauchant les séquences d'un gène codant) ou intergénique (entre deux gènes codants). De plus, la transcription de l'ARN non

codant peut se faire dans le même sens ou dans le sens contraire de l'ARN messenger du gène de référence.

Il est important de souligner que les longs ARN non codants sont généralement moins conservés que les gènes codants lorsqu'on les compare entre différentes espèces de vertébrés. Cependant, lorsqu'on étudie des espèces étroitement liées comme la souris et l'homme, on observe souvent des séquences orthologues. Il convient de noter que tant que nous ne comprendrons pas les mécanismes de signalisation utilisés par les longs ARN non codants pour interagir avec d'autres molécules, il sera difficile d'évaluer le degré de conservation de leurs parties fonctionnelles élémentaires.

Il est intéressant de noter que certains longs ARN non codants partagent certaines caractéristiques structurelles avec les ARN messagers. Une fraction du transcriptome non codant est composée d'ARN multi-exoniques qui subissent des processus d'épissage, tout comme les ARNm. Ces ARN non codants multi-exoniques sont également polyadénylés et possèdent une coiffe à leur extrémité 5'. D'autre part, il existe des longs ARN non codants sans introns qui sont transcrits de manière bidirectionnelle à partir de sites d'initiation généralement situés dans des régions amplificatrices. Certains de ces sites amplificateurs produisent également des ARN non codants multi-exoniques, tout comme les ARN non codants issus de régions promotrices.

À l'heure actuelle, il est difficile de déterminer si les longs ARN non codants dérivés de régions amplificatrices ont des fonctions distinctes en fonction de leur structure bidirectionnelle ou multi-exonique. De plus, il a été observé que les promoteurs de gènes codants peuvent également produire des longs ARN non codants dans le sens opposé de la transcription du gène associé [4,5]. Au-delà de leur simple caractéristique de transcription, la fonction de ces longs transcrits n'est pas encore connue.

### 1.1.3 Fonctions des longs ARN non codants

Les longs ARN non codants ont été associés à diverses fonctions, mais leur rôle précis dans les processus physiopathologiques et le maintien de l'homéostasie cellulaire reste à déterminer. Cependant, l'étude de systèmes modèles a permis de comprendre certains mécanismes moléculaires liés à cette nouvelle classe de molécules. En particulier, les longs ARN non codants jouent un rôle important en tant que régulateurs de la transcription.

Un grand nombre de longs ARN non codants sont impliqués dans le remodelage de la chromatine. Ils agissent en ciblant des complexes modificateurs de la chromatine vers des régions spécifiques du génome [3]. Par exemple, les ARN non codants HOTAIR (HOX transcript antisense RNA) et ANRIL (antisens non-coding RNA in the INK4 locus) [6] se lient au complexe répresseur 2 du polycomb (PRC2) et initient l'inhibition de la transcription des gènes en induisant la production de marques répressives (H3K27Me3) sur la chromatine cible. Les longs ARN non codants peuvent également réguler la transcription en interagissant directement avec des composants essentiels de la machinerie transcriptionnelle, en déplaçant des éléments fonctionnels tels que la polymérase

II ou des facteurs de transcription. Ils peuvent également occuper des sites régulateurs spécifiques sur l'ADN, empêchant ainsi la liaison d'activateurs de la transcription. Par exemple, le long ARN non codant GAS5 (growth arrest-specific 5) imite le site de liaison du récepteur aux glucocorticoïdes et réprime les événements transcriptionnels associés à ces hormones [7].

De plus, certains longs ARN non codants contribuent à la diversité en s'appariant aux ARN messagers, bloquant les sites d'épissage alternatif et favorisant l'apparition préférentielle de certains variants [8]. Une autre fonction des longs ARN non codants est liée à leur capacité à être reconnus par des microARN. Ils peuvent agir comme des leurres, permettant aux ARN messagers ciblés par les microARN d'échapper à la dégradation. Ainsi, le trio ARN messager, microARN et long ARN non codant forme un réseau compétitif (competitive endogenous network), où l'expression d'un des éléments du réseau influence les niveaux des deux autres composants [9].



#### 1.1.4 Les longs ARN non codants dérivés de séquences amplificatrices

Une sous-classe importante de longs ARN non codants est constituée des longs ARN non codants dérivés de séquences amplificatrices (enhancer-associated long noncoding RNA) [10,11]. Les séquences amplificatrices sont des éléments clés du génome qui permettent une expression spécifique des gènes dans les cellules et les tissus. Elles contrôlent également la temporalité de l'expression pendant la différenciation cellulaire et le développement embryonnaire. On estime qu'il existe environ un million de séquences amplificatrices dans le génome humain, ce qui suggère que chaque gène codant est sous le contrôle de nombreuses séquences amplificatrices, mettant en évidence la complexité énorme de la régulation de l'expression génique [12].

Les séquences amplificatrices régulent l'expression des gènes codants adjacents en agissant en cis, mais elles peuvent également réguler l'expression de gènes distants en trans. Lorsqu'elles sont activées, ces séquences amplificatrices s'associent physiquement à plusieurs régions du génome et coordonnent ainsi l'expression de gènes situés dans des régions différentes. La transcription à partir de séquences amplificatrices activées a été démontrée pour la première fois dans des neurones par séquençage à haut débit d'ARN [13], et cette observation a été confirmée dans de nombreux types cellulaires et espèces. La transcription de séquences amplificatrices semble donc être une caractéristique d'un état d'activation, avec des régions génomiques présentant des marques épigénétiques spécifiques, telles que H3K4Me1, acquérant ensuite des marques d'activation, telles que H3K27Ac [14 ,15].

Le mécanisme d'action des longs ARN non codants dérivés de séquences amplificatrices est encore l'objet de recherches intenses. Cependant, il semble que ces ARN jouent un rôle crucial dans la formation de boucles de chromatine qui permettent la mise en contact des séquences amplificatrices avec les sites d'initiation de la transcription de leurs gènes cibles [16,17]. En d'autres termes, les longs ARN non codants initient et probablement stabilisent l'association des séquences amplificatrices avec les promoteurs cibles, favorisant ainsi la transcription des gènes. Des expériences d'inactivation, où ces ARN sont dégradés par des oligonucléotides

spécifiques, ont confirmé l'importance des longs ARN non codants dérivés de séquences amplificatrices, entraînant l'extinction de l'expression des gènes cibles [18].

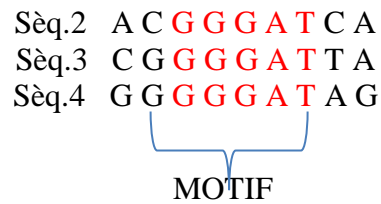
De manière très intéressante, la grande majorité des polymorphismes mononucléotidiques identifiés dans des études d'associations pangénomiques liées à des traits phénotypiques pathologiques chez l'homme se trouvent dans des séquences amplificatrices [19,20]. Il est donc probable que bon nombre de ces mutations ponctuelles modifient l'expression ou la fonction des transcrits non codants dérivés des séquences amplificatrices. De plus, les polymorphismes associés à des pathologies spécifiques d'un organe affectent principalement des séquences amplificatrices actives spécifiques à cet organe [20]. Par exemple, les polymorphismes associés à des changements phénotypiques dans les paramètres électrocardiographiques se trouvent préférentiellement dans des séquences amplificatrices actives dans le cœur. Cette observation souligne l'importance cruciale de la caractérisation fonctionnelle des longs ARN non codants pour notre compréhension des pathologies humaines et notre capacité à développer des traitements personnalisés.

## 1.2 Notion de motif

Le motif est un terme utilisé en bioinformatique pour désigner un ensemble de résidus soit 20 nucléotides ou sept aminoacides qui sont retrouvés dans un ensemble de séquences alignées préalablement et présentant une similarité entre celles-ci. C'est un court segment continu ou non permettant de caractériser un ensemble de séquences non ambiguës et peu dégénérées.

Le motif spécifie généralement une fonction conservée au cours de l'évolution. Pour déterminer un motif dans un ensemble de séquences (nucléiques ou protéiques) on doit réaliser un alignement multiple [Gérard C et al ,2006]. Un ensemble de motifs différents séparés par des régions variables compose un pattern. La Fig.1 représente un motif dans des séquences alignées.

Sèq. 1 A A G G G A T C C



**Figure1** : Exemple d'alignement multiple avec un motif

### 1.2.1 Importance des motifs

Il existe plusieurs raisons pour rechercher les motifs à travers les séquences car ils sont généralement impliqués dans des systèmes de régulation ou définissent des fonctions biologiques. Parmi ces raisons, on peut citer :

- la détermination de la fonction d'une nouvelle séquence (par exemple en localisant un ou plusieurs motifs répertoriés dans des bases de motifs)
- l'identification dans une séquence nucléique de régions codantes (par exemple en repérant les codons d'initiation et de terminaison, les sites d'épissages et les zones de fixation des ribosomes)

- la recherche d'un motif particulier dans une séquence (par exemple en identifiant sur une séquence les sites de coupures d'enzymes de restriction ou des promoteurs Spécifiques etc...)

- l'extraction à partir des banques de données (par exemple extraire des séquences possédant le même signal de régulation ou la même signature protéique pour effectuer des études comparatives ultérieures).

- la recherche d'un site de fixation de facteur de transcription dans une séquence.

Il est donc évident qu'il existe des niveaux de complexité très différents dans la définition des motifs. Certains sont précis et non ambigus comme les sites de reconnaissance des enzymes de restrictions ou comme certains motifs de protéases. D'autres peuvent être beaucoup plus flous et complexes comme les motifs consensus liés à des familles de protéines telles que les facteurs de transcription.

Dans ce cas, la difficulté est souvent de savoir quel motif utiliser et quelle est la pertinence de la définition.

## 1.2.2 Le motif nucléique

À partir de l'analyse d'un alignement multiple des séquences nucléotidiques connues, on peut définir un motif nucléique. Cette analyse permet de déterminer la variabilité des nucléotides à chaque position de l'alignement. En réalisant un alignement multiple des séquences, on obtient une séquence appelée séquence consensus.

### 1.2.2.1 La séquence consensus

La séquence consensus est une représentation des bases les plus fréquentes pour chaque position dans un alignement multiple. Dans le cas de séquences très spécifiques, cette séquence simple peut être suffisante pour décrire de manière satisfaisante une région active. Cependant, dans la plupart des cas, comme pour les facteurs de transcription, la séquence consensus ne suffit pas à identifier les sites biologiquement actifs, car elle ne correspond pas nécessairement au signal le plus fréquent. Dans certains cas, la séquence consensus elle-même peut ne pas être présente en tant que signal actif.

Cela est principalement dû au fait que l'on considère l'indépendance entre les positions lors de la construction du consensus, et celui-ci représente seulement un résumé des séquences réellement actives. Pour atténuer ce problème, il est possible de tolérer un nombre maximal d'erreurs par rapport à la séquence consensus dans la définition du motif. On parle alors d'éloignement ou de distance par rapport à la séquence consensus (Mengeritsky et Smith, 1987).

La séquence consensus est construite à partir de l'alphabet IUPAC ou en retenant une seule base, la plus fréquente, pour chacune des positions de la séquence. L'alphabet d'IUPAC est représenté dans le tableau suivant :

**Tableau 1** : l'alphabet IUPAC

|   |   |          |
|---|---|----------|
| A | A | Adénine  |
| T | T | Thymine  |
| C | C | Cytosine |
| G | G | Guanine  |

|   |            |                                      |
|---|------------|--------------------------------------|
| R | A, G       | Purines                              |
| Y | T, C       | Pyrimidines                          |
| W | A T        | Weak hydrogen bonding (2 liaisons)   |
| S | C G        | Strong hydrogen bonding (3 liaisons) |
| M | A C        | Groupe amino (amine)                 |
| K | T G        | Groupe keto (cétone)                 |
| H | A, C ou T  | Non G                                |
| B | C, G ou T  | Non A                                |
| V | A, C ou G  | Non T                                |
| D | A, G ou T  | Non C                                |
| N | A, G, C, T | N'importe nucléotide                 |

#### 1.2.2.2 Table de fréquences, matrice de pondération

Pour exprimer l'ambiguïté et la complexité d'un motif, on peut également déduire de l'alignement des séquences une table de fréquences en comptabilisant les occurrences de chaque base à chaque position du motif. En d'autres termes, on définit à partir d'un échantillon donné, la probabilité d'apparition des bases pour chaque position du motif.

Il est possible ensuite, pour augmenter la fiabilité des probabilités, de considérer des critères supplémentaires, intrinsèques aux séquences, comme la thermodynamique liée au motif étudié ou la fréquence attendue des bases selon la région où se trouve le motif. On peut ainsi, considérer que l'apparition d'une cytosine est plus significative que l'apparition d'une guanine dans une zone riche en guanine.

La transformation de la table des fréquences en tenant compte éventuellement de critères supplémentaires donne naissance à une matrice de pondération (weight matrix). Celle-ci est généralement construite en prenant le logarithme de la fréquence de chaque base à chaque position pour optimiser les différences contenues dans la table des fréquences. Pour prendre en compte des critères supplémentaires comme le pourcentage des bases de la région étudiée, chacune des valeurs logarithmiques pourra être divisée par la fréquence génomique de la base observée.

On trouvera dans la littérature plusieurs exemples et méthodes de génération de matrices de fréquence ou de pondération (Bucher, 1990 ; Stormo, 1990).

#### 1.2.3 Le motif protéique

Le motif protéique est un élément structural que l'on retrouve dans tous les membres d'une famille de protéine. Il contient des résidus essentiels à une fonction conservée, et ne sont pas nécessairement consécutifs (il s'agit ainsi d'un "pattern"), mais assez proches dans la structure 3D, car ils participent à la même fonction (site actif d'une enzyme, site de fixation d'un récepteur...). Le tableau suivant est un exemple sur motif du doigt de Zinc (un élément structural des facteurs de transcription)

**Tableau 2** : Expression régulière d'une séquence consensus du doigt de Zinc réalisée à partir d'un alignement multiple

|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Y | K | C | T | - | - | V | C | R | - | - | - | K | D | I | S | S | S | E | S | L | R | T | H | M | F | K | - | H | H |
| L | K | C | S | V | P | G | C | K | - | - | - | R | S | F | R | K | K | R | A | L | R | I | H | V | S | - | - | - | H |
| F | E | C | N | - | - | M | C | G | - | - | - | Y | H | S | Q | D | R | Y | E | F | S | S | H | I | T | R | - | E | H |
| Y | T | C | G | - | - | Y | C | T | E | D | S | P | S | F | P | R | P | S | L | L | E | S | H | I | S | - | - | M | H |
| Y | K | C | E | F | A | D | C | E | - | - | - | K | A | F | S | N | A | S | D | R | A | K | H | Q | N | - | - | T | H |
| F | V | C | H | W | Q | D | C | S | R | E | L | R | P | F | K | A | Q | Y | M | L | V | V | H | M | R | - | - | - | H |
| F | R | C | S | - | - | E | C | S | - | - | - | R | S | F | T | H | N | S | D | L | T | A | H | M | K | - | - | - | H |
| C | K | C | E | T | E | N | C | N | - | - | - | L | A | F | T | T | A | S | N | M | R | L | H | F | R | - | - | A | H |
| Y | R | C | S | Y | E | D | C | Q | - | - | - | T | V | S | P | T | W | T | A | L | Q | T | H | L | K | - | - | - | H |
| F | R | C | V | - | - | W | C | K | - | - | - | Q | S | F | P | T | L | E | A | L | T | T | H | M | D | S | - | K | H |

Séquence consensus : x(2)-C-x(2,4)-C-x(15,12)- H-x(2,5)-H.

#### 1.2.4 Description des motifs par une séquence consensus

Pour établir une séquence consensus, on peut réunir toutes les séquences appartenant à une même famille (par exemple, les cytochromes ou les kinases). On recherche ensuite les zones spécifiques qui peuvent être considérées comme caractéristiques de ces séquences, ceci en s'aidant des données disponibles dans la littérature et si possible d'experts de la famille considérée. Les motifs ainsi obtenus sont alors systématiquement recherchés dans une banque de séquences protéiques pour estimer leur fiabilité qui repose sur le nombre de faux positifs et de faux négatifs identifiés.

Une bonne définition doit minimiser ces deux nombres. C'est une des méthodes qu'utilise Amos Bairoch pour constituer la banque de motifs protéiques PROSITE (Bairoch, 1993).

#### 1.2.5 Les logiciels de recherche des motifs consensus

Si le motif est défini par une séquence consensus, il existe des méthodes et logiciels pour rechercher dans une séquence personnelle la présence de tel ou tel motif. C'est le

cas du logiciel **ScanProsite** qui permet de rechercher la présence dans une séquence d'un motif défini dans la banque PROSITE. Pour les domaines protéiques définis dans PRODOM, le logiciel de recherche proposé est Blast.

### 1.2.6 Les bases de données de motifs

Selon les types des motifs, il existe deux types de bases de motifs : celles qui s'intéressent aux motifs nucléiques et les autres sont spécifiques pour les motifs protéiques.

#### 1.2.6.1 Les bases de motif nucléiques

Le tableau suivant représente quelques bases de données pour les motifs nucléiques

**Tableau 3** : les bases de motif nucléiques

| Bases de données                          | Adresse URL                                                                                                                                   |
|-------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| TRANSFAC database                         | <a href="http://www.biobaseinternational.com/pages/index.php?id=transfac">http://www.biobaseinternational.com/pages/index.php?id=transfac</a> |
| JASPAR database                           | <a href="http://jaspar.genereg.net/">http://jaspar.genereg.net/</a>                                                                           |
| SCPD (Promoter Database of S. cerevisiae) | <a href="http://rulai.cshl.edu/SCPD/">http://rulai.cshl.edu/SCPD/</a>                                                                         |

#### 1.2.6.2 Les bases de motif protéiques

Le tableau suivant représente quelques bases de données pour les motifs protéiques.

**Tableau 4** : les bases de motif protéiques

| Bases de données | Adresse URL |
|------------------|-------------|
|                  |             |

|         |                                                                                                                       |
|---------|-----------------------------------------------------------------------------------------------------------------------|
| PROSITE | <a href="https://prosite.expasy.org">https://prosite.expasy.org</a>                                                   |
| CATH    | <a href="https://www.cathdb.info">https://www.cathdb.info</a>                                                         |
| PRODOM  | <a href="http://prodom.prabi.fr/prodom/current/html/home.php">http://prodom.prabi.fr/prodom/current/html/home.php</a> |

## 1.3 Alignement

### 1.3.1 Définition

L'alignement des séquences est un processus qui permet de comparer deux (alignement par paires) ou plusieurs (alignement multiple) séquences de nucléotides ou d'acides aminés afin de déterminer leur degré de similitude. Lorsqu'il existe une similitude significative entre deux séquences génétiques ou protéiques, cela indique une évolution à partir d'une séquence ancestrale commune. Les séquences apparentées de cette manière sont qualifiées d'homologues, et la similitude conservée entre ces séquences au fil de l'évolution est appelée homologie.

Pour comparer deux séquences, l'une des étapes clés consiste à les aligner l'une avec l'autre afin de révéler les similitudes existantes. Les alignements reposent sur le principe de la présence de quelques résidus identiques, situés à des positions correspondantes. Il existe deux types d'alignements couramment utilisés : l'alignement global et l'alignement local.

L'alignement global est réalisé sur l'ensemble des séquences, c'est-à-dire qu'il cherche à aligner l'intégralité des séquences, même si cela inclut des régions peu similaires. L'objectif est d'obtenir un alignement global qui couvre l'ensemble des séquences étudiées.

En revanche, l'alignement local se concentre sur l'alignement d'une partie spécifique des séquences. Il permet de trouver des régions de similitude plus petites et plus spécifiques au sein des séquences globales. L'alignement local est particulièrement utile lorsque l'on souhaite identifier des motifs conservés ou des régions fonctionnelles communes entre les séquences.

Ces deux types d'alignements sont des outils essentiels en bioinformatique et en biologie moléculaire pour étudier les relations évolutives entre les séquences, identifier des homologies et déduire des informations fonctionnelles à partir de similitudes structurales ou séquentielles.

### 1.3.2 Alignement global (Global alignment)

L'algorithme de Needleman-Wunsch est couramment utilisé pour réaliser un alignement global entre deux séquences d'ADN ou de protéines. Lors de l'alignement de deux séquences protéiques (figure 2), l'algorithme recherche une correspondance maximale entre les deux séquences, en alignant autant de résidus identiques que



possible. Il est également possible d'introduire des gaps (ou trous) dans l'alignement, soit au milieu, soit à une extrémité, afin d'optimiser l'alignement global.

L'objectif principal de l'algorithme de Needleman-Wunsch est de trouver l'alignement optimal qui maximise la similarité entre les séquences en tenant compte des similarités locales et des gaps nécessaires. Cela permet de mettre en évidence les régions conservées et de déterminer les motifs structurels et fonctionnels importants dans les séquences.

L'algorithme de Needleman-Wunsch repose sur la construction d'une matrice de scores, où chaque position de la matrice représente une paire de résidus des séquences à aligner. En utilisant des scores de similarité et des pénalités pour les gaps, l'algorithme remplit la matrice en calculant le score optimal pour chaque position. En fin de compte, l'alignement optimal est obtenu en remontant la matrice de scores à partir de la dernière position.

En alignant les séquences de manière globale avec l'algorithme de Needleman-Wunsch, on peut mettre en évidence les similitudes structurelles et fonctionnelles entre les séquences, ce qui permet d'obtenir des informations précieuses sur leur évolution et leur rôle biologique.

```
L G P S S K Q T G K G S - S R I W D N
| | | | | | | | | | | | | | | |
L N - I T K S A G K G A I M R L G D A
```

**Figure 2.** Alignement global : les résidus identiques sont liés par une ligne verticale.

### 1.3.3 Alignement local

L'algorithme de Smith-Waterman (figure 3) est couramment utilisé pour réaliser des alignements locaux entre les séquences d'ADN et de protéines. Contrairement à l'alignement global, l'alignement local se concentre uniquement sur les fragments similaires présents dans les séquences, tandis que le reste des séquences n'est pas pris en compte.

L'objectif principal de l'algorithme de Smith-Waterman est de trouver les fragments conservés ou les régions similaires les plus significatives entre les séquences. Il permet d'identifier des motifs importants, des domaines structuraux ou des régions fonctionnelles spécifiques dans les séquences d'ADN ou de protéines.

L'algorithme de Smith-Waterman repose sur une approche dynamique de programmation, tout comme l'algorithme de Needleman-Wunsch. Il utilise une matrice de scores pour évaluer les similitudes locales entre les résidus des séquences à aligner. En utilisant des scores de similarité et des pénalités pour les gaps, l'algorithme construit une matrice de scores et détermine le score optimal pour chaque position de la matrice.

La particularité de l'algorithme de Smith-Waterman réside dans sa capacité à trouver les meilleurs alignements locaux en trouvant les fragments similaires les plus significatifs, même si ceux-ci ne sont pas présents dans des régions consécutives des séquences. L'algorithme explore toutes les possibilités d'alignements locaux en tenant compte des scores de similarité et des pénalités.

En réalisant des alignements locaux avec l'algorithme de Smith-Waterman, il est possible de découvrir des fragments conservés importants dans les séquences d'ADN et de protéines. Cela permet d'identifier des motifs, des domaines fonctionnels ou des régions structurelles clés, ce qui facilite la compréhension des relations évolutives et des fonctions biologiques des séquences.

```

----- G K G -----
----- | G K G -----

```

**Figure 3.** Alignement local : juste le plus long fragment des résidus identique est pris en considération, les tirets sont utilisés pour indiquer les fragments qui ne sont pas pris en considération.

### 1.3.4 Alignement multiples

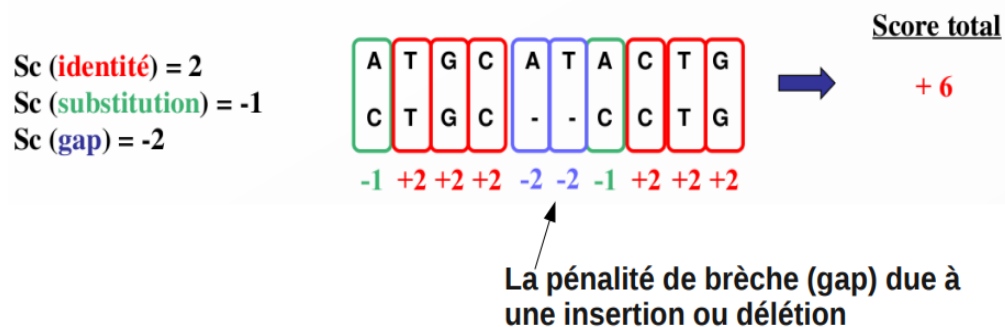
Un alignement multiple consiste à comparer plusieurs séquences. On parvient à ce type d'alignement en considérant successivement tous les alignements possibles deux à deux. En pratique, on compare une séquence à toutes les autres pour tenter de déterminer la voie la plus probable dans l'évolution, étant donné les probabilités de

différentes substitutions possibles. Plus on ajoute de séquences à l'alignement multiple, plus le modèle acquiert de la précision sur l'histoire évolutive des familles des séquences comparées. Cet alignement permet de montrer que certains résidus sont identiques dans toutes les séquences. Tout résidu ou toute séquence courte identique dans toutes les séquences d'un groupe donné est dite conservée. Les alignements multiples donnent en général une meilleure évaluation de la similitude que les alignements deux à deux et permettent d'identifier des membres ayant une relation lointaine dans une famille de gènes, qui n'auraient pas été révélés par un alignement deux à deux.

Cet alignement consiste à superposer chaque résidu d'une séquence avec ceux d'une ou plusieurs autres séquences. De plus, pour construire un alignement optimal, on aura souvent besoin d'ajouter des indels (gaps).

### 1.3.5 Le score

Le score de l'alignement est la somme de toutes les positions 2 à 2.



**Figure 04** : Exemple d'un score

Pour calculer le score d'un alignement de séquences protéiques, on associe à chaque paire de résidus alignés le score correspondant dans la matrice de substitutions (BLOSUM62, PAM250...).

## 1.4 L'ARN

### 1.4.1 Définition

L'ARN (Acide Ribonucléique) partage de nombreuses similitudes avec l'ADN, mais il présente également des différences notables :

**Composition chimique :** La principale différence réside dans le sucre utilisé. L'ADN contient du désoxyribose, tandis que l'ARN utilise du ribose. Cela est dû à la présence d'un groupe hydroxyle supplémentaire sur le ribose de l'ARN.

**Bases nucléotidiques :** Alors que l'ADN utilise les bases nucléotidiques Adénine (A), Thymine (T), Cytosine (C) et Guanine (G), l'ARN remplace la Thymine par l'Uracile (U). Ainsi, les bases nucléotidiques de l'ARN sont Adénine (A), Uracile (U), Cytosine (C) et Guanine (G)

**Structure :** L'ADN se présente généralement sous forme de double brin, où les brins complémentaires s'apparient (A avec T et C avec G). En revanche, l'ARN est généralement simple brin, bien qu'il puisse s'apparier avec un autre ARN complémentaire.

**Types d'ARN :** Trois principaux types d'ARN ont été identifiés : l'ARN messager (ARNm), l'ARN ribosomique (ARNr) et l'ARN de transfert (ARNt). L'ARNm est responsable de la transmission de l'information génétique du noyau cellulaire aux ribosomes pour la synthèse des protéines. L'ARNr constitue une composante essentielle des ribosomes, tandis que l'ARNt transporte les acides aminés nécessaires à la construction des protéines. Ces dernières années, d'autres types d'ARN ont été découverts, chacun ayant des rôles spécifiques dans la régulation génétique et d'autres processus cellulaires.

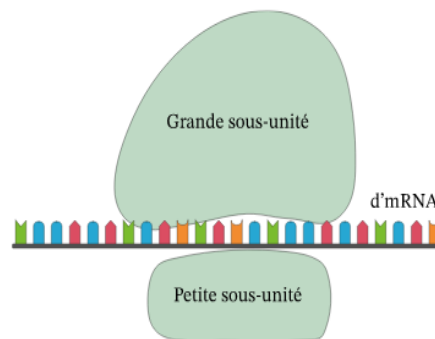
**Virus :** Chez certains virus, le matériel génétique est constitué d'ARN plutôt que d'ADN. Ces virus à ARN utilisent leur ARN pour stocker l'information génétique et se répliquer.

Ces différences entre l'ARN et l'ADN sont essentielles pour les fonctions biologiques spécifiques qu'ils remplissent dans les cellules. L'ARN joue un rôle crucial dans la

synthèse des protéines, la régulation génétique et d'autres processus cellulaires importants.

#### 1.4.2 ARNr

L'ARNr est un type d'ARN non codant qui est l'un des principaux composants des ribosomes.



**Figure 5** : Un schéma montrant la structure de base d'un ribosome.

Les ARNr (ARN ribosomiques) jouent un rôle essentiel au sein des ribosomes, des organites cellulaires responsables de la traduction de l'information génétique. Les ribosomes permettent la synthèse des protéines à partir de l'ARNm, qui est lui-même transcrit à partir d'une portion spécifique du génome. Les ARNr constituent à la fois l'ossature et le cœur du ribosome.

En plus des ARNr, le ribosome est composé d'environ cinquante protéines appelées protéines ribosomiques. Les ARNr sont produits à partir de gènes codés dans l'ADN. Ils sont d'abord transcrits sous forme de précurseurs plus longs, qui subissent ensuite des processus de maturation et de clivage pour donner naissance aux différents ARNr fonctionnels.

Chez les eucaryotes, cette maturation et ce clivage des ARNr se déroulent principalement dans une région spécifique du noyau cellulaire appelée nucléole. Au sein

du ribosome, les ARNr adoptent une structure tridimensionnelle compacte, repliée sur eux-mêmes. Cette conformation en forme de boucle confère une certaine stabilité aux ARNr, contrairement aux ARNm qui ont généralement une durée de vie plus courte.

Les ARNr assurent plusieurs fonctions cruciales dans les ribosomes. Ils facilitent l'assemblage des différentes sous-unités du ribosome, participent à la fixation de l'ARNm et permettent le positionnement précis des acides aminés nécessaires à la synthèse protéique. Ainsi, les ARNr sont essentiels à la traduction de l'information génétique en protéines et jouent un rôle central dans la biologie cellulaire.

| ARNr                                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Chez les procaryotes                                                                                                                                                                                                                       | Chez les eucaryotes                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| <p>-Grande sous-unité du ribosome (50S) :</p> <p>-ARNr 23S (2300 nucléotides).</p> <p>-ARNr 5S (120 nucléotides) : il n'est pas associé au grand ARNr.</p> <p>-Petite sous-unité du ribosome (30S)</p> <p>-ARNr 16S (1500 nucléotides)</p> | <p>-Grande sous-unité du ribosome (60S) : elle est composée de 45 protéines plus</p> <p>-ARNr 28S (4800 nucléotides) pour les animaux et 25S pour les végétaux.</p> <p>-ARNr 5.8S (160 nucléotides) : Dans le ribosome, l'ARNr 5.8S est accroché au 28S</p> <p>-ARNr 5S (120 nucléotides) : il n'est pas accroché au 28S ni au 5.8S</p> <p>-Petite sous-unité du ribosome (40S) : composée de 33 protéines plus</p> <p>-ARNr 18S (1900 nucléotides)</p> |

Les ARNr 28S ; 5.8S et 18S sont synthétisés dans le nucléole tandis que l'ARN 5S est synthétisé à l'extérieur du nucléole, mais toujours dans le noyau.

Dans les cellules eucaryotes, la production des ribosomes se déroule au sein d'une petite structure située à l'intérieur du noyau cellulaire, appelée nucléole. Tout

comme l'ARNm, l'ARNr est synthétisé par transcription à partir de l'ADN. Cependant, contrairement à l'ARNm, l'ARNr ne subit pas de traduction pour être converti en protéines.

Le nucléole est responsable de la synthèse des ARNr, et le taux de production de ces molécules est élevé, atteignant plusieurs centaines de milliers de molécules par heure. En effet, l'ADN des cellules eucaryotes contient généralement environ 600 copies des gènes nécessaires à la production des ARNr.

Les polypeptides ribosomiques, les composants protéiques des ribosomes, sont quant à eux synthétisés dans le cytoplasme de la cellule. Ensuite, ils se déplacent vers le nucléole en traversant l'enveloppe nucléaire. C'est dans le nucléole que les polypeptides ribosomiques s'assemblent avec les ARNr pour former les ribosomes fonctionnels. Cette étape d'assemblage est cruciale pour la formation complète des ribosomes, qui seront ensuite exportés vers le cytoplasme pour participer à la synthèse protéique.

#### 1.4.3 ARNr16s

L'ARNr16s est une molécule d'ARN ribosomal présente chez les bactéries, ainsi que chez certains autres organismes procaryotes et eucaryotes. Il joue un rôle essentiel dans la synthèse des protéines et est impliqué dans le processus de traduction.

L'ARNr16s fait partie de la petite sous-unité ribosomique (30S chez les bactéries) et est composé d'une séquence d'environ 1500 bases nucléotidiques. Il est hautement conservé dans l'évolution, ce qui signifie que certaines régions de cet ARN restent relativement inchangées au fil du temps. Cependant, il présente également des régions hypervariables, qui sont plus spécifiques à chaque espèce bactérienne et permettent d'identifier et de classer les différentes espèces.

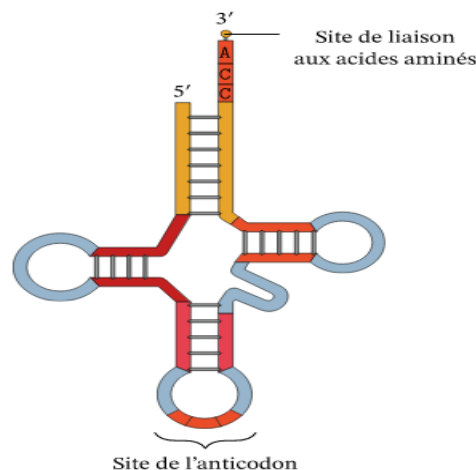
En raison de cette combinaison de conservation et de variation, l'ARNr16s est souvent utilisé comme marqueur moléculaire pour étudier la diversité et la classification des bactéries

#### 1.4.4 ARNt

L'ARNt, ou ARN de transfert, joue un rôle essentiel dans la traduction de l'information génétique en protéines. Il s'agit d'une molécule adaptatrice qui transporte les acides aminés spécifiques vers les ribosomes lors de la synthèse protéique.

Tout comme l'ARNr, les molécules d'ARNt sont synthétisées par transcription d'un segment d'ADN. Cependant, contrairement à l'ARNm, elles ne sont pas traduites en protéines elles-mêmes. Leur fonction principale est de fournir les acides aminés correspondants aux ribosomes lors de la traduction de l'ARNm.

La structure de l'ARNt est caractéristique et présente une forme en feuille de trèfle, comme illustré dans la figure 06. Cette structure en trèfle est cruciale car elle permet à l'ARNt de se lier à la fois à l'acide aminé spécifique et à l'ARNm pendant le processus de traduction. Cette interaction entre l'ARNt, l'acide aminé et l'ARNm assure la précision et l'exactitude de la séquence d'acides aminés lors de la synthèse protéique.



**Figure 6 :** Schéma illustrant la forme en trèfle d'une molécule d'ARNt

La structure de la molécule d'ARNt est caractérisée par des régions qui se replient sur elles-mêmes et forment des appariements de bases, créant ainsi des boucles ou des bras qui confèrent à la molécule sa forme en trèfle. Deux régions clés dans cette structure sont importantes pour la traduction : l'anticodon et le site de liaison aux acides aminés, comme indiqué dans la figure 6.

L'anticodon est une séquence de trois nucléotides spécifiques présente sur la molécule d'ARNt. Lorsque le ribosome se déplace le long de l'ARNm pendant la traduction, chaque codon de l'ARNm est associé à une molécule d'ARNt portant un



anticodon complémentaire au codon. Cette reconnaissance de séquence entre l'anticodon de l'ARNt et le codon de l'ARNm assure l'appariement correct des acides aminés pendant la synthèse protéique.

Le site de liaison aux acides aminés est une autre région de l'ARNt où l'acide aminé spécifique correspondant est attaché. Chaque type d'ARNt est spécifique à un acide aminé particulier, et il existe une enzyme appelée aminoacyl-ARNt synthétase qui catalyse cette liaison entre l'ARNt et l'acide aminé correspondant. Ainsi, l'ARNt joue un rôle crucial en transportant spécifiquement les acides aminés nécessaires vers les ribosomes pour la construction de la protéine en cours de synthèse.

## 2. MATERIEL ET METHODES

### 2.1 Les séquences des ARNr 16S et ARNt :

Les séquences des ARNr 16s et ARNt de cinq entérobactéries (voir tableau 5) ont été téléchargées à partir du National Center for Biotechnology Information (NCBI). Pour cela, une recherche a été effectuée sur le site web de NCBI en utilisant les noms spécifiques des cinq bactéries. Les critères de recherche ont été restreints pour obtenir les séquences les plus récentes et les plus complètes disponibles.

Le tableau suivant présente les séquences d'ARNr 16s de cinq bactéries téléchargées, avec leurs identifiants (ID) respectifs et leurs tailles :

**Tableau 05** : Caractéristiques des séquences d'ARNr16s de cinq entérobactéries

| Bactérie                                   | Taille (Pb) | Gène ID    |
|--------------------------------------------|-------------|------------|
| <i>Escherichia coli</i>                    | 1450        | 219722938  |
| <i>Salmonella (enterica subsp.salamae)</i> | 1448        | 343205887  |
| <i>Salmonella bongari</i>                  | 1443        | 636560064  |
| <i>Proteus terrae</i>                      | 1444        | 1150406765 |
| <i>Klebsiella pneumoniae</i>               | 1452        | 645322555  |

Les séquences d'ARNt téléchargées sont présentées dans les tableaux suivants, avec leurs identifiants (codes) et leurs tailles respectives.

**Tableau 06** : Caractéristiques des séquences d'ARNt de cinq entérobactéries

| <b>Bactérie</b>             | <b>Taille (Pb)</b> | <b>Gène ID</b> |
|-----------------------------|--------------------|----------------|
| <i>Salmonella bangori</i>   | 1775               | 66756751       |
| <i>Klebsiella pasteurii</i> | 1781               | 772276553      |
| <i>Escherichia coli</i>     | 1794               | 947297         |
| <i>Proteus mirabilis</i>    | 1787               | 6802570        |
| <i>Shigella flexneri</i>    | 1783               | 772276653      |

Les séquences des ARNr 16s et ARNt correspondantes ont été téléchargées au format FASTA à l'aide de l'interface de téléchargement fournie par NCBI. Les séquences obtenues ont été enregistrées localement dans un répertoire dédié pour une utilisation ultérieure dans l'alignement multiple, et la définition des motifs.

## 2.2 Alignement multiple de séquences :

Après avoir téléchargé les séquences d'ARNr 16s et ARNt des cinq entérobactéries sélectionnées, un alignement multiple a été réalisé pour permettre une comparaison et une analyse plus approfondies. Pour cela, le logiciel T-COFFEE (Version 12.52.6cbe486) (Notredame et al. 2000) a été utilisé.

T-Coffee (Tree-based Consistency Objective Function for alignment Evaluation) est un logiciel utilisé pour effectuer l'alignement multiple de séquences biologiques. Il utilise une approche basée sur des arbres phylogénétiques pour améliorer la qualité de l'alignement en prenant en compte les informations évolutives.

Le logiciel a été exécuté en utilisant les paramètres par défaut, garantissant ainsi un alignement des séquences optimal.

Une fois l'alignement multiple effectué, les séquences ont été ajustées les unes par rapport aux autres, permettant ainsi de visualiser les positions conservées et variables au niveau de chaque position de cet alignement multiple.

### 2.3 Détection et définition des motifs :

Notre approche est basée sur une adaptation d'un code informatique pour mettre en œuvre une méthode de calcul des différentes matrices.

La première matrice à calculer est une matrice de comptage, où les lignes représentent les positions de l'alignement multiple et les colonnes représentent les quatre nucléotides.

Une deuxième matrice est ensuite calculée pour représenter les fréquences de chaque nucléotide dans chaque colonne.

Une troisième matrice, appelée matrice de pondération (Matrice Poids-Position), est construite en calculant le logarithme de la fréquence de chaque base à chaque position. Cette étape est réalisée avec la formule suivante :  $\log_2 (f(x)/0,25)$ .

Enfin, à partir de cette dernière matrice, le motif s'affiche et son score est calculé.

Un ensemble d'outils, dont JavaScript, Vue.js, Node.js, Yarn et Tauri Apps, ont été utilisés pour créer une application de détection des motifs.

Grâce à JavaScript, nous avons pu développer la logique et les fonctionnalités essentielles de l'application. Vue.js, un framework JavaScript a permis de créer une interface utilisateur réactive et conviviale, offrant une expérience utilisateur fluide.

Node.js, a joué un rôle crucial en tant qu'environnement d'exécution côté serveur, permettant d'effectuer des opérations de traitement des données et de communiquer avec

d'autres composants de l'application. Yarn a été utilisé comme gestionnaire de paquets pour gérer les dépendances du projet de manière efficace et maintenable.

Enfin, Tauri Apps a été intégré dans le processus de développement pour créer une application de bureau à partir du code JavaScript et Vue.js. Tauri Apps fournit les outils nécessaires pour emballer l'application en une application native pour la plateforme Windows.

**Tableau 07** : les outils utilisés et leur description

| Outil      | Description                                                                                                                                                                                                                                                                                                             |
|------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| JAVASCRIPT | JavaScript est un langage de programmation utilisé principalement pour le développement web. Il permet d'ajouter des fonctionnalités interactives aux pages web.                                                                                                                                                        |
| Vue.js     | Vue.js est un framework JavaScript pour créer des interfaces utilisateur réactives. Il facilite la création d'applications web modulaires en utilisant des composants réutilisables.                                                                                                                                    |
| Node.js    | Node.js est un environnement d'exécution côté serveur pour JavaScript. Il permet d'exécuter du code JavaScript sur le serveur, ce qui est utile pour créer des applications web, des API et des outils en ligne de commande.                                                                                            |
| Yarn       | Yarn est un gestionnaire de paquets JavaScript utilisé pour gérer les dépendances des projets de manière efficace et fiable.                                                                                                                                                                                            |
| Tauri Apps | Tauri Apps est un framework qui permet de créer des applications de bureau multiplateformes à partir de code JavaScript, HTML et CSS. Il offre un accès aux fonctionnalités système du système d'exploitation hôte, permettant de créer des applications de bureau performantes avec une expérience utilisateur native. |

### 3. RESULTAT

#### 3.1 L'alignement Multiple :

L'alignement obtenu à l'aide de T-COFFEE pour les séquences d'ARNr16s est partiellement représenté (voir annexe) pour les cinq séquences 16S provenant d'entérobactéries :

```
NR_119278.1 AGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAAT
NR_146019.1 AGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAAT
NR_116124.1 AGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAAT
NR_024570.1 AGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAAT
NR_044372.1 AGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAAT

cons *****

NR_119278.1 TACTGGGCGTAAAGCGCACGCAGGCGGTCTGTCAAGTCCGATGTGAAATCCCCGGGCTCAACCTGGGAA
NR_146019.1 TACTGGGCGTAAAGCGCACGCAGGCGGTCAATTAAGTCAGATGTGAAAGCCCCGAGCTTAACTGGGAA
NR_116124.1 TACTGGGCGTAAAGCGCACGCAGGCGGTCTGTCAAGTCCGATGTGAAATCCCCGGGCTCAACCTGGGAA
NR_024570.1 TACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGCTCAACCTGGGAA
NR_044372.1 TACTGGGCGTAAAGCGCACGCAGGCGGTCTGTCAAGTCCGATGTGAAATCCCCGGGCTCAACCTGGGAA

cons ***** * *****

NR_119278.1 CTGCATTCGAAACTGGCAGGCTAGAGTCTTGTAGAGGGGGGTAGAATTCAGGTGTAGCGGTGAAATGC
NR_146019.1 TTGCATCTGAAACTGGTGGCTAGAGTCTTGTAGAGGGGGGTAGAATTCACGTGTAGCGGTGAAATGC
NR_116124.1 CTGCATCCGAAACTGGCAGGCTTGAGTCTCGTAGAGGGGGGTAGAATTCAGGTGTAGCGGTGAAATGC
NR_024570.1 CTGCATCTGAACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAATTCAGGTGTAGCGGTGAAATGC
NR_044372.1 CTGCATTCGAAACTGGCAGGCTTGAGTCTTGTAGAGGGGGGTAGAATTCAGGTGTAGCGGTGAAATGC

cons ***** ** *****

NR_119278.1 GTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACAAAGACTGACGCTCAGGTGCGAAA
NR_146019.1 GTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACAAAGACTGACGCTCAGGTGCGAAA
NR_116124.1 GTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAA
NR_024570.1 GTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAA
NR_044372.1 GTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACAAAGACTGACGCTCAGGTGCGAAA

cons *****
```



### 3.2 Détection et définition de motif :

Nous avons sélectionné une portion de l'alignement de 25 nucléotides pour l'analyser dans notre application. Le tableau 08 représente la portion de l'alignement des séquences d'ARNr 16s

**Tableau 08** : une portion d'alignement des séquences d'ARNr 16 S prise au sein de l'alignement multiple

```
G C G A T G A G G T T A A T A A C C T C A T C G A
G T G T T A A G A T T A A T A C T C T T A G C A A
G C G A C A A G G T T A A T A A C C T T G T T G A
G G A G T A A A G T T A A T A C C T T T G C T C A
G T G T T G A G G T T A A T A A C C G C A G C A A
```

Les résultats obtenus pour les différentes étapes de calculs et de définition des motifs sont :

La matrice de comptage :

| Comptage    | Fréquences |    |    |    | Poids-Position |   |   |   |
|-------------|------------|----|----|----|----------------|---|---|---|
|             | A          | T  | C  | G  | A              | T | C | G |
| Position    |            |    |    |    |                |   |   |   |
| Position 1  | 0          | 0  | 0  | 5  |                |   |   |   |
| Position 2  | 0          | 2  | 2  | 1  |                |   |   |   |
| Position 3  | 1          | 0  | 1  | 3  |                |   |   |   |
| Position 4  | 2          | 2  | 0  | 1  |                |   |   |   |
| Position 5  | 0          | 4  | 1  | 0  |                |   |   |   |
| Position 6  | 3          | 0  | 0  | 2  |                |   |   |   |
| Position 7  | 4          | 1  | 0  | 0  |                |   |   |   |
| Position 8  | 1          | 0  | 0  | 4  |                |   |   |   |
| Position 9  | 1          | 0  | 0  | 4  |                |   |   |   |
| Position 10 | 0          | 5  | 0  | 0  |                |   |   |   |
| Total       | 40         | 37 | 21 | 27 |                |   |   |   |



La matrice de fréquence :

| Comptage    | Fréquences |     | Poids-Position |     |
|-------------|------------|-----|----------------|-----|
| Position    | A          | T   | C              | G   |
| Position 1  | 0          | 0   | 0              | 1   |
| Position 2  | 0          | 0.4 | 0.4            | 0.2 |
| Position 3  | 0.2        | 0   | 0.2            | 0.6 |
| Position 4  | 0.4        | 0.4 | 0              | 0.2 |
| Position 5  | 0          | 0.8 | 0.2            | 0   |
| Position 6  | 0.6        | 0   | 0              | 0.4 |
| Position 7  | 0.8        | 0.2 | 0              | 0   |
| Position 8  | 0.2        | 0   | 0              | 0.8 |
| Position 9  | 0.2        | 0   | 0              | 0.8 |
| Position 10 | 0          | 1   | 0              | 0   |

La matrice Poids-Position :

| Comptage    | Fréquences  |             |      | Poids-Position |       |
|-------------|-------------|-------------|------|----------------|-------|
| Position    | A           | T           | C    | G              | Motif |
| Position 1  | -10         | -10         | -10  | <b>2</b>       | G     |
| Position 2  | -10         | <b>0.68</b> | 0.68 | -10            | T     |
| Position 3  | -10         | -10         | -10  | <b>1.3</b>     | G     |
| Position 4  | <b>0.68</b> | 0.68        | -10  | -10            | A     |
| Position 5  | -10         | <b>1.7</b>  | -10  | -10            | T     |
| Position 6  | <b>1.3</b>  | -10         | -10  | 0.68           | A     |
| Position 7  | <b>1.7</b>  | -10         | -10  | -10            | A     |
| Position 8  | -10         | -10         | -10  | <b>1.7</b>     | G     |
| Position 9  | -10         | -10         | -10  | <b>1.7</b>     | G     |
| Position 10 | -10         | <b>2</b>    | -10  | -10            | T     |
| Score       |             |             |      |                | 38.42 |

À partir de cette dernière matrice, le motif et le score sont affichés.

Le motif : GTGATAAGGTTAATAACCTTACCTTATCAA

Le score : 38,42

L'insertion d'une nouvelle séquence d'ARN pour test a donné le résultat suivant :

Sequence 1

Sequence X

Score: 25.48

Longueur de la séquence : - 25 +

Supprimer une séquence Ajouter une séquence

| Comptage    | Fréquences |      |      |      | Poids-Position |
|-------------|------------|------|------|------|----------------|
|             | A          | T    | C    | G    |                |
| Position    |            |      |      |      |                |
| Position 1  | -10        | -10  | -10  | 2    | G              |
| Position 2  | -10        | 0.68 | 0.68 | -10  | T              |
| Position 3  | -10        | -10  | -10  | 1.3  | G              |
| Position 4  | 0.68       | 0.68 | -10  | -10  | A              |
| Position 5  | -10        | 1.7  | -10  | -10  | T              |
| Position 6  | 1.3        | -10  | -10  | 0.68 | A              |
| Position 7  | 1.7        | -10  | -10  | -10  | A              |
| Position 8  | -10        | -10  | -10  | 1.7  | G              |
| Position 9  | -10        | -10  | -10  | 1.7  | G              |
| Position 10 | -10        | 2    | -10  | -10  | T              |
| Score       |            |      |      |      | 38.42          |

Le calcul de score à partir de la matrice poids-position donne un score similaire au premier score. Nous pouvons considérer que cette séquence test fait partie des ARNr16s.

L'insertion d'une deuxième séquence d'ARN pour test a donné :

Sequence 1

Sequence X

Score: -88.88

Longueur de la séquence : - 25 +

Supprimer une séquence Ajouter une séquence

| Comptage    | Fréquences |      |      |      | Poids-Position |
|-------------|------------|------|------|------|----------------|
|             | A          | T    | C    | G    |                |
| Position    |            |      |      |      |                |
| Position 16 | 1.3        | -10  | 0.68 | -10  | A              |
| Position 17 | -10        | -10  | 1.7  | -10  | C              |
| Position 18 | -10        | -10  | 1.7  | -10  | C              |
| Position 19 | -10        | 1.7  | -10  | -10  | T              |
| Position 20 | -10        | 1.3  | 0.68 | -10  | T              |
| Position 21 | 1.3        | -10  | -10  | 0.68 | A              |
| Position 22 | -10        | 0.68 | -10  | 0.68 | T              |
| Position 23 | -10        | 0.68 | 1.3  | -10  | C              |
| Position 24 | 0.68       | -10  | -10  | 0.68 | A              |
| Position 25 | 2          | -10  | -10  | -10  | A              |
| Score       |            |      |      |      | 38.42          |

Le score n'est pas proche du premier score, ce qui signifie que cette séquence test ne fait pas partie des ARNr16 S.

# CONCLUSION

Les ARNnc constituent une large famille génomique au sein des trois phyla (Archéobactéries, Bactéries et Eucaryotes) avec des proportions et des rôles tout à fait différents.

Les données relatives à leurs séquences nécessitent un traitement spécifique pour les classer au sein de familles bien distinctes sur la base de motifs et de signatures moléculaires spécifique et relatives à chacune de ces familles ; ce qui constituera une organisation bien meilleure des bases de données génomiques et une exploration plus facile pour l'utilisateur de ces bases de données.

Notre application a permis, sur la base de calculs tirés des résultats d'alignements multiples, de proposer une définition non ambiguë du motif des familles analysées par notre étude.

Cependant, nous envisageons, en termes de perspectives, d'améliorer cette application par des mesures de paramètres de précision et d'estimer surtout si des faux positifs ou des faux négatifs sont insérés dans les réponses de notre application, car leur simple présence va contribuer à une mauvaise classification des séquences génomiques d'intérêt et par voie de conséquence cela nuira à l'annotation post-séquençage des génomes analysés.

## REFERENCES BIBLIOGRAPHIQUES

1. Cécile Dumas, « Génome: pour en finir avec l'ADN "poubelle" » [archive], sur [www.sciencesetavenir.fr](http://www.sciencesetavenir.fr), 10 septembre 2012 (consulté le 4 avril 2020).
2. Mattick JS. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* 2003 ; 25 :930-939. [[CrossRef](#)] [[PubMed](#)] [[Google Scholar](#)]
3. Wapinski O, Chang HY. Long noncoding RNAs and human disease. *Trends Cell Biol* 2011 ; 21 : 354–361. [[CrossRef](#)] [[PubMed](#)] [[Google Scholar](#)]
4. Orom UA, Shiekhattar R. Long noncoding RNAs usher in a new era in the biology of enhancers. *Cell* 2013 ; 154 : 1190–1193. [[CrossRef](#)] [[PubMed](#)] [[Google Scholar](#)]
5. Preker P, Nielsen J, Kammler S, et al. RNA exosome depletion reveals transcription upstream of active human promoters. *Science* 2008 ; 322 : 1851–1854. [[CrossRef](#)] [[PubMed](#)] [[Google Scholar](#)]
6. Mathieu EL, Belhocine M, Dao LTM, et al. Rôle des longs ARN non codants dans le développement normal et pathologique. *Med Sci (Paris)* 2014 ; 30 : 790–796. [[CrossRef](#)] [[EDP Sciences](#)] [[PubMed](#)] [[Google Scholar](#)]
7. Kotake Y, Nakagawa T, Kitagawa K, *et al.* Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. *Oncogene* 2011 ; 30 : 1956–1962. [[CrossRef](#)] [[PubMed](#)] [[Google Scholar](#)]
8. Kino T, Hurt DE, Ichijo T, *et al.* Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci Signal* 2010; 3 : ra8. [[PubMed](#)] [[Google Scholar](#)]
9. Zong X, Tripathi V, Prasanth KV. RNA splicing control: yet another gene regulatory role for long nuclear noncoding RNAs. *RNA Biol* 2011 ; 8 : 968–977. [[CrossRef](#)] [[PubMed](#)] [[Google Scholar](#)]
10. Salmena L, Poliseno L, Tay Y, *et al.* A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language. *Cell* 2011 ; 146 : 353–358. [[CrossRef](#)] [[PubMed](#)] [[Google Scholar](#)]
11. Orom UA, Derrien T, Beringer M, et al. Long noncoding RNAs with enhancer-like function in human cells. *Cell* 2010 ; 143 : 46–58. [[CrossRef](#)] [[PubMed](#)] [[Google Scholar](#)]

12. Lam MT, Li W, Rosenfeld MG, Glass CK. Enhancer RNAs and regulated transcriptional programs. *Trends Biochem Sci* 2014 ; 39 : 170–182. [[CrossRef](#)] [[PubMed](#)] [[Google Scholar](#)]
13. Xie W, Ren B. Developmental biology. Enhancing pluripotency and lineage specification. *Science* 2013 ; 341 : 245–247. [[Google Scholar](#)]
14. Kim TK, Hemberg M, Gray JM, *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* 2010 ; 465 : 182–[PubMed](#)] [[Google Scholar](#)]
15. Heintzman ND, Hon GC, Hawkins RD, *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 2009 ; 459 : 108–112. [[CrossRef](#)] [[PubMed](#)] [[Google Scholar](#)]
16. Kaikkonen MU, Spann NJ, Heinz S, *et al.* Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Mol Cell* 2013 ; 51 : 310–325. [[CrossRef](#)] [[PubMed](#)] [[Google Scholar](#)]
17. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature* 2012 ; 489 : 109–113. [[CrossRef](#)] [[PubMed](#)] [[Google Scholar](#)]
18. Zhang Y, Wong CH, Birnbaum RY, *et al.* Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* 2013 ; 504 : 306–310. [[CrossRef](#)] [[PubMed](#)] [[Google Scholar](#)]
19. Li W, Notani D, Ma Q, *et al.* Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* 2013 ; 498 : 516–520. [[CrossRef](#)] [[PubMed](#)] [[Google Scholar](#)]
20. Vernot B, Stergachis AB, Maurano MT, *et al.* Personal and population genomics of human regulatory variation. *Genome Res* 2012 ; 22 : 1689–1697. [[CrossRef](#)] [[PubMed](#)] [[Google Scholar](#)]

**Année universitaire : 2022-2023**

**Présenté par : CHEKIREB Rayen et MEROUANI Aicha**

## Essai de définitions des motifs structuraux de la famille des ARNnc

### Mémoire pour l'obtention du diplôme de Master en BIOINFORMATIQUE

Les ARN non codants (ARNnc) sont des transcrits non traduits qui jouent des rôles clés et variés dans la cellule. Notre travail décrit une analyse basée sur la construction et le calcul de matrices (comptage, fréquence et poids-position) pour définir les signatures moléculaires existantes au sein d'une même famille d'ARNnc. L'objectif de ce travail est de détecter un motif structural d'ARNr16s et d'ARNt à partir d'un alignement multiple, et d'en calculer le score correspondant. L'analyse des scores calculés permet de confirmer la présence de motifs communs et de classer des séquences génomiques au sein de familles de gènes leur correspondant.

**Mots clés :** ARNnc, Matrice Poids-Position, Alignement Multiple, Motif.

**Président :** KELLOU Kamel Université Frères Mentouri, Constantine 1

**Examineur :** GHERBOUDJ Amira Université Frères Mentouri, Constantine 1

**Encadreur :** HAMIDECHI Med Abdelhafid Université Frères Mentouri, Constantine 1