Study of circular code motifs in nucleic acid sequences

Karim El Soufi

UNIVERSITY OF STRASBOURG

DOCTORALE SCHOOL «Mathematics, Information Sciences and Engineering»

ICube

Thesis presented by:

Karim EL SOUFI

defended on : **24 January 2017**

to obtain: **Doctor of the university of Strasbourg**

Discipline / Speciality: Computer science

Study of circular code motifs in nucleic acid sequences

THESIS supervised by :

MICHEL Christian	Professor, University of Strasbourg

RAPPORTEURS :

COMET Jean-Paul	Professor, University of Nice Sophia Antipolis
FIMMEL Elena	Professor, Hochschule Mannheim

OTHER MEMBERS OF THE JURY :

COLLET Pierre	Professor, University of Strasbourg, jury president
STRÜNGMANN Lutz	Professor, Hochschule Mannheim
THOMPSON Julie	Director of research, University of Strasbourg

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE «Mathématique, Sciences de l'Informatique et de l'Ingénieur»

ICube

THÈSE présentée par :

Karim EL SOUFI

soutenue le : **24 Janvier 2017**

pour obtenir le grade de: **Docteur de l'université de Strasbourg**

Discipline / Spécialité: Informatique

Étude des motifs de code circulaire dans les séquences d'acides nucléiques

THÈSE dirigée par :

MICHEL Christian	Professeur, Université de Strasbourg

RAPPORTEURS :

COMET Jean-Paul	Professeur, Université Nice Sophia Antipolis
FIMMEL Elena	Professeur, Hochschule Mannheim

AUTRES MEMBRES DU JURY :

COLLET Pierre	Professeur, Université de Strasbourg, président du jury
STRÜNGMANN Lutz	Professeur, Hochschule Mannheim
THOMPSON Julie	Directeur de recherche, Université de Strasbourg

Abstract

Le travail effectué dans cette thèse présente une nouvelle approche de la théorie du code circulaire dans les gènes qui a été initiée en 1996. Cette approche consiste à analyser les motifs construits à partir de ce code circulaire. Ces motifs particuliers sont appelés motifs de code circulaire. Ainsi, nous avons développé des algorithmes de recherche pour localiser les motifs de code circulaire dans les séquences d'acides nucléiques afin de leurs trouver une signification bioinformatique. En effet, le code circulaire X identifié dans les gènes est un ensemble de trinucléotides qui a la propriété de retrouver, synchroniser et maintenir la phase de lecture. De plus, il possède la propriété d'autocomplémentarité $\mathcal{C}$: les trinucléotides de X sont complémentaires entre eux, c'est-à-dire $X = \mathcal{C}(X)$. Enfin, il a la propriété C^3 : les ensembles de trinucléotides $\mathcal{P}(X)$ et $\mathcal{P}^2(X)$ par permutation de X d'un et de deux nucléotides, respectivement, sont également des codes circulaires et de plus complémentaires entre eux, c'est-à-dire $\mathcal{C}(X_1) = X_2$ et $\mathcal{C}(X_2) = X_1$.

Notre travail de recherche s'est donc intéressé à la recherche de motifs du code circulaire X dans des séquences d'ADN ou d'ARN. Nous avons commencé notre analyse avec le centre de décodage du ribosome (ARNr) qui est une région majeure dans le processus de traduction des gènes aux protéines. Puis, nous avons étendu les résultats obtenus avec le ribosome aux ARN de transfert (ARNt) pour étudier les interactions ARNr-ARNt. Enfin, nous avons généralisé la recherche de motifs de code circulaire X dans l'ADN aux chromosomes d'eucaryotes complets.

La théorie du code circulaire a contribué à l'analyse du centre de décodage du ribosome, en particulier à sa structure primaire. De fa con surprenante, les nucléotides universellement conservés A1492 et A1493 dans tous les ARNr de bactéries, d'archées, d'eucaryotes nucléaires et de chloroplastes appartiennent à des motifs de code circulaire $m_{AA}(X)$. Le nucléotide conservé G530 dans les ARNr des bactéries et des archées est également inclus dans des motifs de code circulaire $m_G(X)$. Le développement d'un algorithme de recherche des motifs de code circulaire associé à l'alignement global des séquences multiples permet d'identifier les motifs de code circulaire $m_G(X)$ dans les ARNr nucléaires et chloroplastes, résultat qui ne peut pas être obtenu part des méthodes classiques

de bioinformatique. Par ailleurs, un nouveau motif de code circulaire $m(X)$ universellement conservé dans les sept organismes étudiés est identifié gr^ace à la théorie du code circulaire. La visualisation spatiale des trois motifs $m_{AA}(X)$, $m_G(X)$ et $m(X)$ montrent qu'ils appartiennent au centre de décodage du ribosome dans tous les ARNr étudiés de bactéries, d'archées, d'eucaryotes nucléaires et de chloroplastes. En conclusion, la fonction biologique du centre de décodage du ribosome qui a été attribuée à un nombre restreint de nucléotides, précisément les nucléotides A1492, A1493 et G530, peut maintenant être associée à des motifs de code circulaire comportant au moins deux et jusqu'à cinq trinucléotides successifs.

Nous identifions également de nouvelles propriétés de cette théorie du code circulaire avec des analyses statistiques de motifs du code circulaire X de grande taille dans les génomes des eucaryotes. Pour la première fois, les régions non-codantes (en dehors des gènes) sont étudiées avec cette théorie du code circulaire. Les motifs du code circulaire X de grande taille de longueurs $l \geq 15$ trinucléotides et de cardinalité (composition) supérieure à 10 trinucléotides ont la plus grande fréquence d'apparition dans les génomes des eucaryotes par rapport (i) aux 23 motifs de codes circulaires bijectifs de grande taille, (ii) aux deux motifs de codes circulaires permutés de grande taille ; (iii) aux motifs aléatoires de grande taille obtenus avec des codes aléatoires (non circulaires). Les plus longs motifs du code circulaire X sont identifiés dans les génomes des eucaryotes, par exemple un motif X dans une région non-codante du génome *Solanum pennelli*i avec une longueur de 155 trinucléotides (465 nucleotides) associé à une probabilité d'occurrence de 10^{-71} , deux motifs X dans des régions non-codante du génome Salmo salar avec des longueurs de 118 trinucléotides (354 nucléotides) avec une probabilité d'occurrence de 10^{-52} , etc. Le plus longs motif du code circulaire X dans le génome humain se trouve dans une région non-codante du chromosome 13 avec une longueur de 36 trinucléotides et une probabilité d'occurrence de 10^{-11}.

Les motifs du code circulaire X dans les régions non-codantes des génomes sont probablement des vestiges évolutifs de gènes primitifs utilisant le code circulaire pour la traduction des gènes aux protéines. Cependant, les études statistiques réalisées dans ce travail de thèse montrent que les motifs X apparaissent préférentiellement dans les gènes avec une proportion de motifs X (de longueurs l supérieure à 10 trinucléotides et de cardinalité supérieure à 5 trinucléotides) égal à 8 dans les gènes / régions non-codantes pour 138 génomes eucaryotes complets. Ce facteur de 8 est également retrouvé avec les motifs X dans les gènes / régions non-codantes des 24 chromosomes humains. D'un point de vue biologique, cette

propriété peut s'expliquer par le fait que les mutations (substitution, insertion et délétion de nucléotides) sont plus fréquentes dans les régions non-codantes par rapport aux gènes.

L'existence du code circulaire X dans les gènes est un problème ouvert depuis sa découverte en 1996. Nous montrons dans la suite de notre travail que le concept de code circulaire dans les régions d'ADN de faible complexité existe également avec les codes circulaires unitaires (UCC) de dinucléotides, trinucléotides et tétranucléotides générant des motifs UCC de dinucléotides répétés (Di^+ motifs), de trinucléotides répétés (motifs Tri^+) et de tétranucléotides répétés (motifs $Tetra^+$) dans les génomes d'eucaryotes. Plus précisément, 12 codes UCC de dinucléotides sont "strong comma-free" et quatre d'entre eux $\{AT\}$, $\{CG\}$, $\{GC\}$ et $\{TA\}$ sont auto-complémentaires. Egalement, 48 codes UCC de trinucléotides sont "strong comma-free" et 12 codes UCC de trinucléotides sont "comma-free". Enfin, 180 codes UCC de tétranucléotides sont "strong comma-free", 60 codes UCC de tétranucléotides sont "comma-free" et 12 codes "strong comma-free" $\{AATT\}$, $\{ACGT\}$, $\{AGCT\}$, $\{CATG\}$, $\{CCGG\}$, $\{CTAG\}$, $\{GATC\}$, $\{GGCC\}$, $\{GTAC\}$, $\{TCGA\}$, $\{TGCA\}$ et $\{TTAA\}$ sont en plus autocomplémentaires. Ainsi, les motifs Di^+, Tri^+ et $Tetra^+$ permettent de retrouver une phase de lecture modulo 2, modulo 3 et modulo 4, respectivement, dans les régions non-codantes des eucaryotes. De plus, les propriétés $\mathcal{C}^2$, $\mathcal{C}^3$ et $\mathcal{C}^4$ permettent également de retrouver les phases décalées et la propriété d'autocomplémentarité permet l'appariement des deux brins de l'ADN dans les régions non codantes des eucaryotes. Un motif UCC et son motif UCC complémentaire ont la même distribution dans les génomes des eucaryotes, à la fois selon leur nombre d'apparition et leur nombre total de nucléotides. Cette propriété est observée avec les motifs Di^+, Tri^+ et $Tetra^+$. De plus, pour les motifs Tri^+ et $Tetra^+$, un motif UCC et son motif UCC complémentaire ont des occurrences croissantes inversement proportionnel à leur nombre de liaisons hydrogène.

De manière surprenante, on observe une rareté des trinucléotides répétés (motifs Tri^+) dans les grands génomes d'eucaryotes par rapport aux motifs Di^+ et $Tetra^+$. Ce résultat statistique est obtenu avec des mesures de moyenne et de médiane, et confirmé par deux tests statistiques (un test paramétrique t de Student pour échantillon apparié et un test non-paramétrique W de Wilcoxon signé). Ainsi, les codes circulaires unitaires de trinucléotides associés aux trinucléotides répétés dans les génomes des eucaryotes pourraient avoir contribué à la formation du code circulaire X dans les gènes. Une conséquence d'une telle hypothèse serait la persistance de certaines propriétés statistiques des trinucléotides répétés

du code circulaire X. De fa con inattendue, des paires de trinucléotides identiques (14 trinucléotides parmi 20) du code circulaire X sont préférentiellement utilisées dans les gènes des eucaryotes. Pour la première fois depuis 20 ans, la théorie du code circulaire dans les gènes est étendue aux génomes dans ce travail de thèse. Ainsi, le code circulaire pourrait être une structure mathématique des gènes mais également des génomes.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Christian Michel for his patience, motivation, and immense knowledge in the field. His guidance helped me in all the time of research and writing of this thesis.

My regards to the jury members, Prof. Jean-Paul Comet and Dr. Julie Thompson, for showing interest in my work, and Profs. Pierre Collet, Elena Fimmel and Lutz Strüngmann for their continued interest. The valuable feedback that they offered was instrumental to the enhancement of my work.

I would like to thank the Lebanese Association for Scientific Research (LASeR) for their help in financing my thesis and my team Complex Systems and Translational Bioinformatics (CSTB) for their support. A special gratitude to my friends and loved ones who offered assistance when I needed it most.

Last but not the least, I would like to thank my parents for whom without I would not be here. Therefore, I dedicate this thesis to them.

Contents

Introduction **1**

Circular code . . . 1

Biological context . . . 2

Thesis structure . . . 3

1 Biological environment **5**

1.1 Introduction . . . 6

1.2 Nucleic acid sequence . . . 6

1.3 Ribosome . . . 8

1.3.1 Translation . . . 8

1.3.2 Biomolecular structure . . . 9

1.4 Biological databases . . . 11

1.4.1 Crystallographic database . . . 12

1.4.2 Sequence database . . . 13

1.5 Summary . . . 15

2 Code Theory **17**

2.1 Introduction . . . 18

2.2 History of codes . . . 18

2.2.1 Diamond code . . . 18

2.2.2 Comma-free code . . . 19

2.2.3 Genetic code . . . 21

2.3 Circular code . . . 21

2.3.1 Circular code motifs . . . 25

2.3.2 Unitary circular codes of dinucleotides, trinucleotides and tetranucleotides . . . 26

2.4 Summary . . . 29

3 Data and Methods **31**

3.1 Introduction . . . 33

3.2 Data acquisition . . . 33

3.2.1 Ribosomal data . . . 33

3.2.2 Genomic database 34
3.3 Circular code motifs 36
3.3.1 Search algorithms 36
3.3.1.1 Biinfinite word 36
3.3.1.2 Sets algorithm 37
3.3.1.3 Frames algorithm 41
3.3.2 Definition of random code motifs 42
3.3.3 Definition of the 23 bijective circular codes motifs 43
3.3.3.1 Bijective transformation circular codes 43
3.3.3.2 Main properties of the 23 bijective transformation circular codes 47
3.3.4 Statistical analysis of X circular code motifs 47
3.3.4.1 Coverage of X motifs in tRNA 47
3.3.4.2 Expectation of the occurrence number of a motif . . 48
3.3.4.3 Ratio of X motifs in coding and non-coding regions . 48
3.4 Ribosome study tools 50
3.4.1 Multiple sequence alignment 50
3.4.2 Molecule viewer 50
3.5 Repeated motifs 52
3.5.1 Dinucleotide unitary circular code motifs 52
3.5.2 Trinucleotide unitary circular code motifs 52
3.5.3 Tetranucleotide unitary circular code motifs 53
3.5.4 Statistical analysis of repeated motifs 53
3.5.4.1 Occurrence number of unitary circular code motifs . 53
3.5.4.2 Base number of unitary circular code motifs 55
3.5.4.3 Total base number of unitary circular code motifs . . 55
3.6 Occurrence number of trinucleotide pairs 56
3.7 Summary 59

4 Results and Discussion **61**
4.1 Introduction 63
4.2 X circular code motifs in the ribosome 63
4.2.1 X circular code motifs in the ribosomal decoding center . . . 63
4.2.1.1 The conserved A1492 and A1493 nucleotides 63
4.2.1.2 The conserved G530 nucleotide 68
4.2.2 Spatial study of X circular code motifs near the ribosomal decoding center 73

4.2.2.1 A conserved X motif near the ribosomal decoding center . . . 73
4.2.2.2 Conserved X motifs in the rRNA of prokaryotes . . . 74
4.2.2.3 Conserved X motifs in the rRNA of eukaryotes . . . 78
4.2.3 X circular code motifs in prokaryotic tRNAs . . . 80
4.2.3.22 Summary of X circular code motifs in tRNA sequences 99
4.2.3.23 Coverage of X circular code motifs in prokaryotic tRNAs . . . 100
4.3 Analysis of X circular code motifs in eukaryotic genomes . . . 100
4.3.1 Occurrence of large randoms code motifs in eukaryotic genomes 101
4.3.2 Occurrence of large motifs from X, X_1, X_2 and the 23 bijective transformations of X in eukaryotic genomes . . . 102
4.3.3 Largest X motifs in eukaryotic genomes . . . 105
4.3.4 Largest X motifs in *Homo sapiens* . . . 108
4.3.5 X motifs in coding regions versus non-coding regions in eukaryotic genomes . . . 108
4.3.6 X motifs in coding regions versus non-coding regions in *Homo sapiens* . . . 110
4.4 Analysis of unitary circular code motifs in eukaryotic genomes . . . 111
4.4.1 Occurrence of repeated dinucleotides in eukaryotic genomes . 112
4.4.2 Occurrence of repeated trinucleotides in eukaryotic genomes . 112
4.4.3 Occurrence of repeated tetranucleotides in eukaryotic genomes 114
4.4.4 Largest repeated motifs in eukaryotic genomes . . . 116
4.4.5 Scarcity of repeated trinucleotides in eukaryotic genomes . . 119
4.5 Identical trinucleotide pairs of the X circular code in eukaryotic gene sequences . . . 120
4.6 Summary . . . 123

5 Conclusion **125**

References **130**

List of Figures

1.1 Nucleotide. 6
1.2 Phosphodiester bond. 7
1.3 DNA structure. 7
1.4 Ribosome during translation. 9
1.5 Decoding center of the ribosome. 10
1.6 Partial 3D structure of *E. coli* ribosome. 11
1.7 Complete 3D structure of *E. coli* ribosome. 12
1.8 Protein Data Base (PDB) growth. 13

2.1 Diamond code. 19
2.2 Circular code definition. 23
2.3 Circular code example. 23
2.4 Circular code window. 24

3.1 ClustalX output. 50
3.2 Modified multiple sequence alignment output. 51
3.3 Jmol. 51

4.1 3D structure of the ribosome of *Escherichia coli*. 65
4.2 3D structure of the ribosome of *Thermus thermophilus*. 66
4.3 3D structure of the ribosome of *Pyrococcus furiosus*. 67
4.4 3D structure of the ribosome of *Saccharomyces cerevisiae*. 68
4.5 3D structure of the ribosome of *Triticum aestivum*. 69
4.6 3D structure of the ribosome of *Homo sapiens*. 71
4.7 3D structure of the ribosome of *Spinacia oleracea*. 72
4.8 Conserved X motifs in *Escherichia coli*. 75
4.9 Conserved X motifs in *Thermus thermophilus*. 75
4.10 Conserved X motifs in *Pyrococcus furiosus*. 77
4.11 Conserved X motifs in *Saccharomyces cerevisiae*. 77
4.12 Conserved X motifs in *Triticum aestivum*. 78
4.13 Conserved X motifs in *Homo sapiens*. 78

4.14 Occurrence numbers of the large motifs from X, X_1, X_2 and the 23 bijective transformations of X. 102
4.15 Occurrence numbers, by length, of the large motifs from X, X_1, X_2 and the top six bijective transformations of X. 103
4.16 Occurrence numbers, by cardinality, of the large motifs from X, X_1, X_2 and the top six bijective transformations of X. 104
4.17 Base ratio of coding/non-coding regions and base ratio of X motifs in the 138 eukaryotic genomes. 105
4.18 Base ratio of coding/non-coding regions and base ratio of X motifs in *Homo sapiens* genome. 108
4.19 Occurrence numbers of repeated dinucleotides in eukaryotic genomes. 113
4.20 Base numbers in repeated dinucleotides in eukaryotic genomes. . . 113
4.21 Occurrence numbers of repeated trinucleotides in eukaryotic genomes. 114
4.22 Base numbers in repeated trinucleotides in eukaryotic genomes. . . 115
4.23 Occurrence numbers of repeated tetranucleotides in eukaryotic genomes. 116
4.24 Base numbers in repeated tetranucleotides in eukaryotic genomes. . 117
4.25 Preference of identical trinucleotide pairs in gene sequences. 121

List of Tables

2.1 Trinucleotide occurrences. 22
2.2 Circular code statistics. 25

3.1 Protein Data Base (PDB) entries. 34
3.2 Genomes characteristics. 35
3.3 Bijective transformation circular codes 46

4.1 X motifs containing the A1492 and A1493 nucleotides. 64
4.2 X motifs containing the G530 nucleotide. 70
4.3 Conserved motif in the decoding center. 73
4.4 Prokaryotic conserved motifs near the decoding center. 74
4.5 Eukaryotic conserved motifs near the decoding center. 79
4.27 The coverage of X motifs in tRNA sequences. 101
4.28 Top 20 largest X motifs in the 138 eukaryotic genomes. 106
4.29 Largest X motifs in *Homo sapiens* chromosomes. 107
4.30 Base ratio of coding/non-coding regions and base ratio of X motifs in the 138 eukaryotic genomes. 109
4.31 Base ratio of coding/non-coding regions and base ratio of X motifs in the *Homo sapiens* genome. 110
4.32 Largest repeated motifs in eukaryotic genomes. 117
4.33 Correlation matrix of the base number of all the repeated trinucleotides in eukaryotic genomes. 118
4.34 Scarcity of repeated trinucleotides. 119

Introduction

We offer here a general introduction of this thesis, realized at The Engineering Science, Computer Science and Imaging (ICube) laboratory. Created in 2013, the laboratory brings together researchers of the University of Strasbourg, the CNRS (French National Center for Scientific Research), the ENGEES and the INSA of Strasbourg in the fields of engineering science and computer science. With around 580 members, ICube is a major driving force for research in Strasbourg.

The work done in this thesis presents a new direction for circular code identified in 1996 by analysing the motifs constructed from circular code. These particular motifs are called circular code motifs. We applied search algorithms to locate circular code motifs in nucleic acid sequences in order to find biological significance. We start with an overview of circular code and its property of coding frame retrieval for genes. Afterwards we present the biological environment in which we applied our work. Finally, we show the structure in which this thesis is presented.

Circular code

The genetic code consists of 64 trinucleotides $\{AAA, \ldots, TTT\}$, called codons. Each codon encodes for one of the 20 amino acids used in the synthesis of proteins (translation). Most of the amino acids are encoded by more than one codon. Some of the codons have a special purpose. The ATG codon serves as the starting point of translation while also encoding the amino acid methionine at the same time. While the following three trinucleotides $\{TAA, TAG, TGA\}$, called stop codons, are an exception, they do not encode for an amino acid they signal the end of a translation process.

The genetic code can be expressed as either ribonucleic acid (RNA) codons or deoxyribonucleic acid (DNA) codons. RNA codons occur in messenger RNA (mRNA) and are the codons that are actually read during translation. Each mRNA molecule acquires its sequence of nucleotides by transcription from the corresponding gene. Genes are DNA sequences which are read modulo 3 letters among the three possible frames. As such, only one frame, called reading

frame, which begins with a start codon and ends with a stop codon, codes the corresponding protein sequence according to the genetic code.

However, this does not mean a translation starts at every ATG codon, even though it accounts for most start codons, it could be one of the following $\{TTG, GTG, CTG\}$. Additional requirements need to be present when assigning a start codon, such as the ribosome biding side, the shine-delgarno sequence. This short sequence needs to be located 7 to 13 bases upstream of the start codon. Add to that the importance of maintaining the correct reading frame.

All this indicates that the procedure of maintaining the correct reading frame of genes is far more complex. It was theorized that there are sets of trinucleotides called circular codes X which have the property of reading frame retrieval, synchronization and maintenance (Michel, 2012). Furthermore, there are circular codes which have in addition the C self-complementary property, i.e. the trinucleotides of X are complementary to each other, i.e. $X = \mathcal{C}(X)$. Finally, there are self-complementary circular codes X which have in addition the C^3 property, i.e. the permuted trinucleotide sets $\mathcal{P}(X)$ and $\mathcal{P}^2(X)$ of X by one and two nucleotides, respectively, are also trinucleotide circular codes and complementary to each other, i.e. $\mathcal{C}(X_1) = X_2$ and $\mathcal{C}(X_2) = X_1$. In 1996, a C^3 self-complementary trinucleotide circular code X has been identified in genes (reading frame of mRNAs) simultaneously in eukaryotes and prokaryotes (Arquès and Michel, 1996).

Biological context

Our work revolves around searching for X circular code motifs in DNA or RNA sequences. A circular code motif is basically a sequence of nucleotides where its trinucleotides belong to the circular code. In our biological environment we approached the matter by addressing the very specific and narrow, then moving to the more general and broader look. The first part of the study was confined to the ribosome, more specifically, we started with the decoding center of the ribosome. A region that is very important to the translation process. Afterwards, we included the transfer RNA in the ribosome in the study, in order to give a wider look at the circular code motifs in the ribosome and whether a possible interaction exist. Later in our work, to address the presence of X circular code motifs in DNA, we searched the entire database of complete eukaryotic chromosomes. This involved a huge amount of data and computation. Our results from circular code were thoroughly compared to those of random generated codes in

order to establish the significance of our findings.

Thesis structure

This thesis is structured into an introduction, four chapters for the main body and a conclusion.

The first chapter will present the biological environment we are working in. Our main focus lies in the nucleic acid sequence, we will be explaining what is a sequence made of. We worked in two different biological contexts, the first was the ribosome, in order to highlight the importance of our findings we explaining briefly the translation process and shed some light on the 3D structure of the ribosome. The second phase of our work involved the sequence of complete eukaryotic chromosomes, which was acquired from the RefSeq database.

The second chapter serves as an introduction of circular code, beginning with the root that lead to its discovery. This takes us back to the race of cracking the genetic code, as many code theories fell in light of reality once the genetic code was finally cracked. We will explain how the circular code came to be after the discovery of the genetic code, its features and significance. Finally, we present what are circular code motifs, which serve as the main study course of this thesis.

The data and methods are presented in the fourth chapter. We present the data obtained from the databases mentioned before. A new algorithm was written to extract motifs from sequences, this algorithm is versatile as it can take any code as a parameter. Multiple sequence alignments were modified to allow us to switch the attention onto circular code motifs, this combination allowed us to find interesting results in the ribosome. The chapter also explains how the huge data from the complete eukaryotic chromosomes was approached. We show the various codes we used and the statistical methods employed to compare them.

Our results are presented, detailed and discussed in the fourth chapter. The amount of data that we extracted was just huge. This extensive chapter shows the discoveries found in the decoding center of the ribosome, while also raising some questions about several interesting motifs found in the ribosome and offer an intriguing take on the structure of the tRNA. We also present an deep statistical analysis of the significance and uniqueness of the X circular code, discovered in 1996 (Arquès and Michel, 1996), when compared to other codes, whether they are circular, bijective transformation or randomly generated.

Finally, we give the conclusions from our various findings while providing

some theories, as our work raises questions on the importance and role of the circular code while highlighting new properties.

1

Biological environment

1.1 Introduction 6
1.2 Nucleic acid sequence 6
1.3 Ribosome 8
1.3.1 Translation 8
1.3.2 Biomolecular structure 9
1.4 Biological databases 11
1.4.1 Crystallographic database 12
1.4.2 Sequence database 13
1.5 Summary 15

1.1 Introduction

In this chapter we will define the biological environment in which we are working.

This thesis can be divided into two different sub studies with respect to the biological environment. In the first half of this study we searched for circular code motifs in ribosomes, the cellular protein factory. The study focused on the presence of circular code motifs in important areas of the ribosome. To accomplish this the data used includes the 3D structure of the ribosome and a spatial examination of these motifs. The second part of our work involves a more general study and a wider scope. The sequences of complete eukaryotic chromosomes were retrieved and searched for circular code motifs.

We will give brief description of the nucleic acid sequence, which constitute our base target of research, and then introduce the ribosome while explaining its function during translation. Finally, we explain the nature of the data we are using and its source.

1.2 Nucleic acid sequence

Nucleosides that have one or more phosphate groups attached to the sugar are called nucleotides, those containing ribose (OH) are known as ribonucleotides while those containing deoxyribose (H) are known as deoxyribonucleotides (Figure 1.1). Nucleobases are nitrogen-containing rings linked to a sugar within a nucleoside, historically called simply as *bases*. These bases are grouped into two families depending on strong resemblance, Cytosine (C), Thymine (T), and Uracil (U) are called pyrimidines while guanine (G) and adenine (A) are purines. Each nucleotide is named after the base it contains.

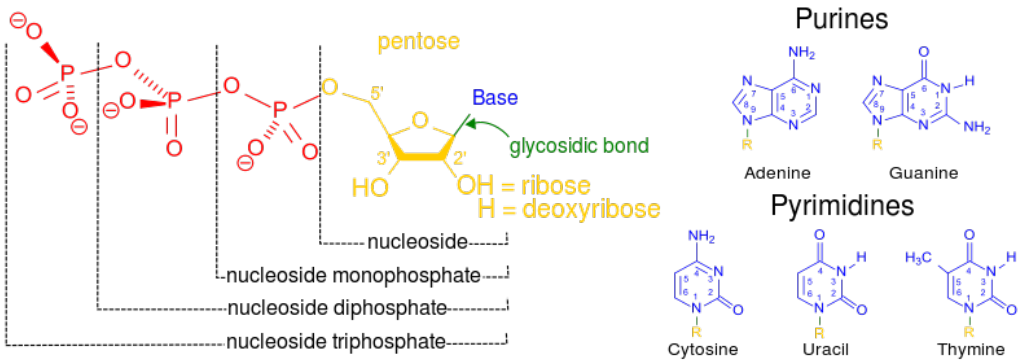

Figure 1.1: Structural elements of a nucleotide: nucleoside in green, bases in blue (with their groups) and the different phosphate groups in red. (By Boris [Public domain], via Wikimedia Commons).

Nucleotides are linked together by the formation of covalent bonds between the phosphate group attached to the sugar of a nucleotide and the hydroxyl group on the sugar of the next nucleotide (Figure 1.2). These links will form a nucleic acid polymer. Nucleotides are therefore responsible for the storage and retrieval of biological information.

Nucleic acids differ in the type of sugar contained in their sugar backbone. Those based on the sugar ribose are known as ribonucleic acids (RNA), and contain the bases A, G, C, and U. They are generally a single-stranded polynucleotide chain. While based on deoxyribose are known as deoxyribonucleic acids (DNA), and contain the bases A, G, C, and T (T is chemically similar to the U in RNA). They are a double helix composed of two polynucleotide chains that run in opposite directions and are held together by hydrogen bonds between the bases of the two chains (Figure 1.3).

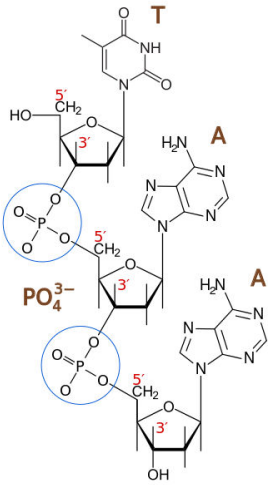

Figure 1.2: An example of phosphodiester bonds (PO_4^{3-}) between Thymine (T) and two molecules of Adenine (A). (By G3pro [Public domain], from Wikimedia Commons).

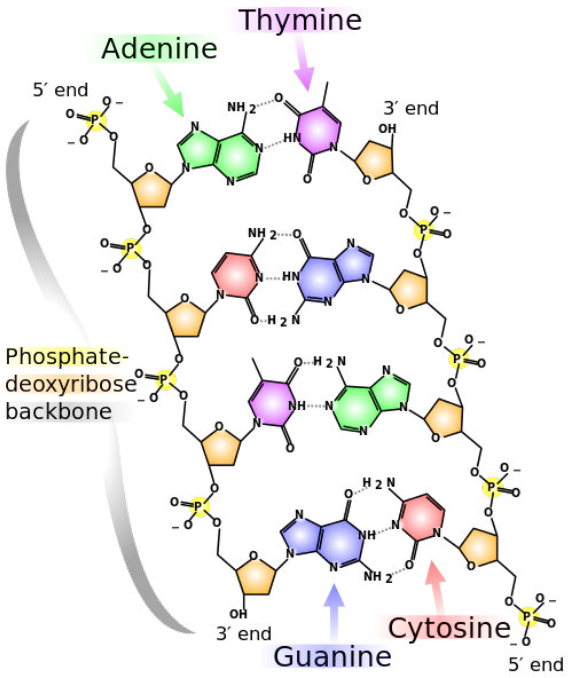

Figure 1.3: Chemical structure of DNA. The two chains run in opposite directions. (By Madeleine Price Ball [Public domain], via Wikimedia Commons).

These two types of nucleic acid hold the genetic information in all life forms, but they differ in roles. On one hand the DNA is more stable because of the double helix structure making it ideal for long term keeping, on the other hand

the RNA is a transient carrier of information.

1.3 Ribosome

The ribosome, a cellular organelle, is a complex macromolecule consisting of RNAs and proteins. It is responsible for the synthesis of the cell protein by translating specific genetic information that is encoded in the deoxyribonucleic acid (DNA) of the cell genome and transferred to the ribosome by messenger RNA (mRNA).

A ribosome is composed of two subunits, a large subunit and a small subunit. Each subunit is an assembly of ribosomal RNAs (rRNAs) and ribosomal proteins. The small subunit is responsible for initiation, identification of the correct reading frame and encoding of the genetic code. The main chemical reaction of protein synthesis, peptide bond formation, occurs in the large subunit. A ribosome contains three transfer RNA (tRNA) binding sites: A-site (aminoacyl), P-site (peptidyl), and E-site (exit).

1.3.1 Translation

Translation is the process of adding one amino acid after the other to the growing polypeptide chain until the protein synthesis in done. Initially the two before mentioned subunits of the ribosome come together on an mRNA at it's 5' end.

At first, the aminoacyl tRNA carrying the amino acid binds to the A-site where the decoding center containing the universally conserved nucleotides G530, A1492 and A1493 of the smaller rRNA subunit is tasked with distinguishing cognate from non-cognate tRNAs by anticodon-codon interactions with the mRNA codon (Wilson, 2014). After the aminoacyl-tRNA binds to the corresponding codon on the mRNA, a peptide-bond forms between the carboxyl end of the polypeptide chain at the P-site and the new arrived amino-acid at the A-site. This reaction is catalysed by an enzymatic site in the large subunit. Consequently, the larger subunit shifts relatively to the smaller subunit, thus moving the tRNAs in the A and P sites to the P and E sites respectively. Following this, the small subunit will move exactly three nucleotides along the mRNA, aligning itself with the large subunit. This will incur a reset, where the tRNA at the E-site is ejected while the A-site is empty for a new tRNA. Given that the mRNA is being translated in the direction of the 3' from the 5', means that the protein is formed first from its N-terminal end all the way to its C-terminus.

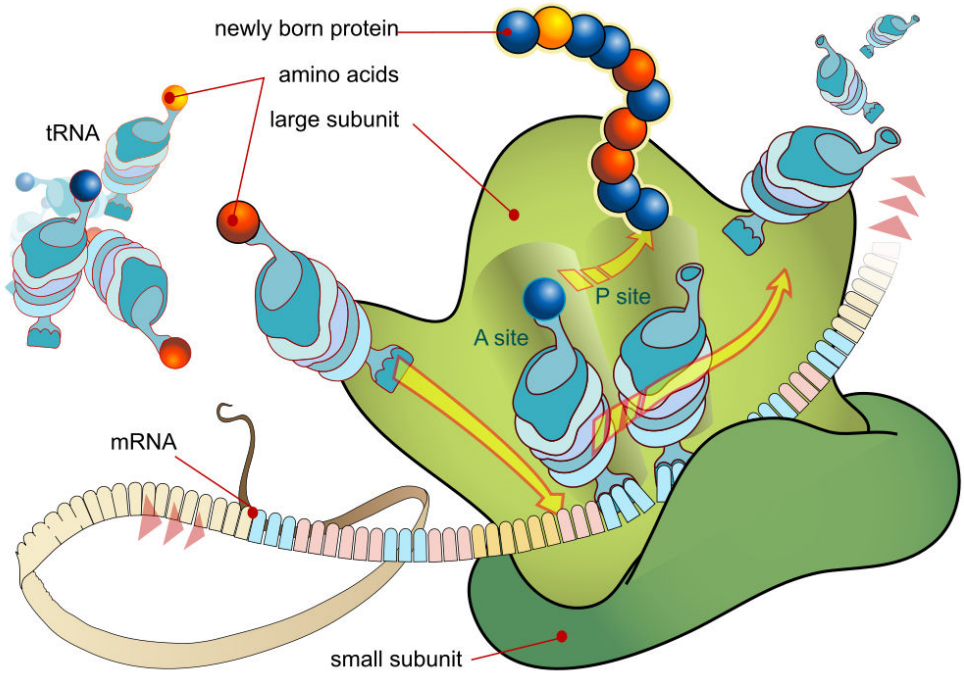

Figure 1.4: The mechanics of ribosome during translation. The start of the aminoacyl tRNA at the A-site, then the transition to the P-site and the exit through the E-site. Showing the position of mRNA with respect to the smaller ribosomal subunit and the various tRNAs. (By LadyofHats [Public domain], via Wikimedia Commons).

When synthesis of the protein is finished, the two subunits of the ribosome separate. Translation speed varies between domains, between six and nine amino acids per seconds for eukaroytes, while it is between seventeen and twenty-one for prokaryotes (Reuveni et al., 2011).

1.3.2 Biomolecular structure

Understanding the functionality of the ribosome was accomplished by determining its three-dimensional structure. While several methods exist to determine the structure of a protein, two are most commonly used, with a third method steadily on the rise (Figure 1.8).

The first of which is X-ray crystallography, where the protein is purified and crystallized. Given the tiny wavelength of X-rays (0.1 nanometre), scientists can then probe the structure of very small objects at the atomic level. Interpreting the resulting map of electron density determines the location of each atom. The second method is Nuclear Magnetic Resonance (NMR) spectroscopy, where a protein is purified, placed in a strong magnetic field then subjugated to a blast of radio waves. These steps will align the atoms according to the magnetic field,

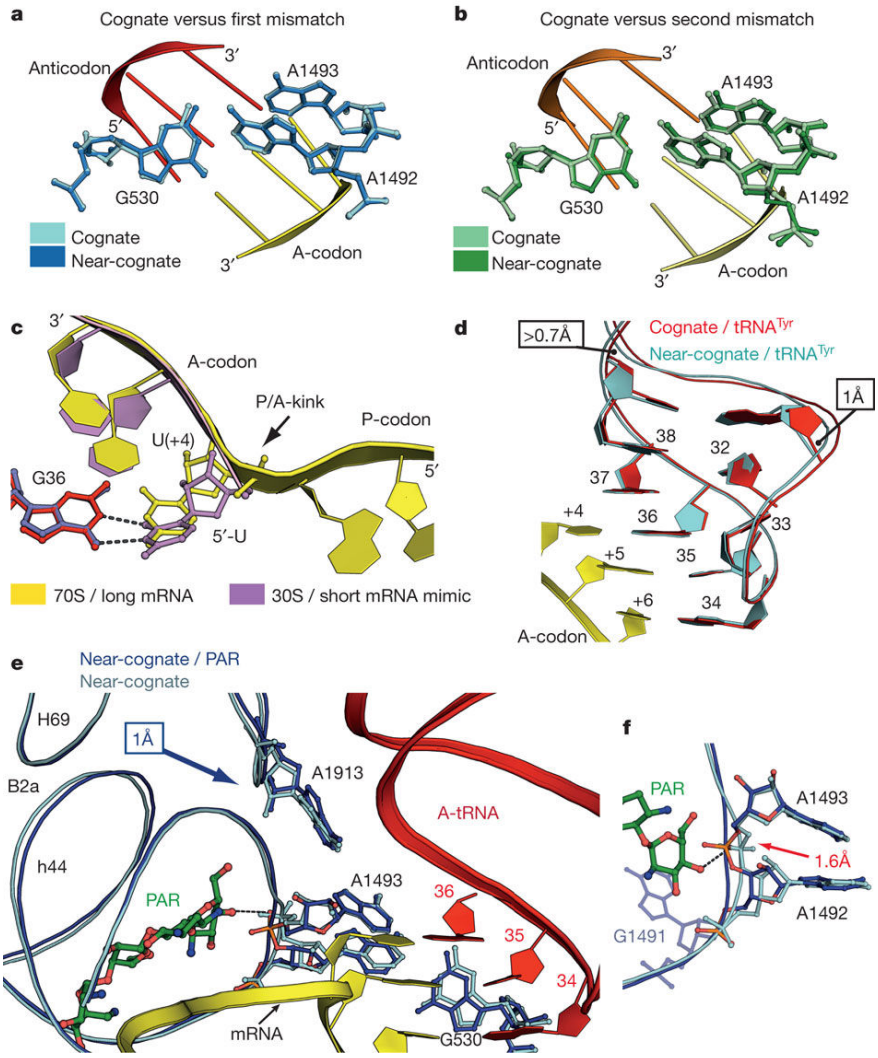

Figure 1.5: (a), (b), The overall conformations of universally conserved G530, A1492 and A1493 of 16S rRNA in the cognate structures are identical to those in the near-cognate models when the mismatches are at the first (a) or second (b) codon-anticodon positions. (c), differences between the position of the first uridine in the UUU codon base-paired to the GAG anticodon of $tRNA_2^{Leu}$ from the 70S structure and from the 30S model. (d), Comparison of the anticodon loops of $tRNA^{Tyr}$ in the cognate (red) and near-cognate (cyan) states. (e), Rearrangements of rRNA helices h44 and H69 in the near-cognate state upon binding of the aminoglycoside paromomycin (PAR). The near-cognate structures with $tRNA^{Tyr}$ are shown. (f), Magnified view of the changes in the A1493 phosphate position (Demeshkina et al., 2012).

which the radio waves will then disturb briefly before returning to their aligned position. This will allow scientist to study the relative position of these atoms in a protein. The third method is Electron Microscopy (EM), where a beam of electrons is used to image the molecule directly. Typically, EM experiments are combined with information from X-ray crystallography or NMR spectroscopy for atomic details mainly due for present limitation of EM.

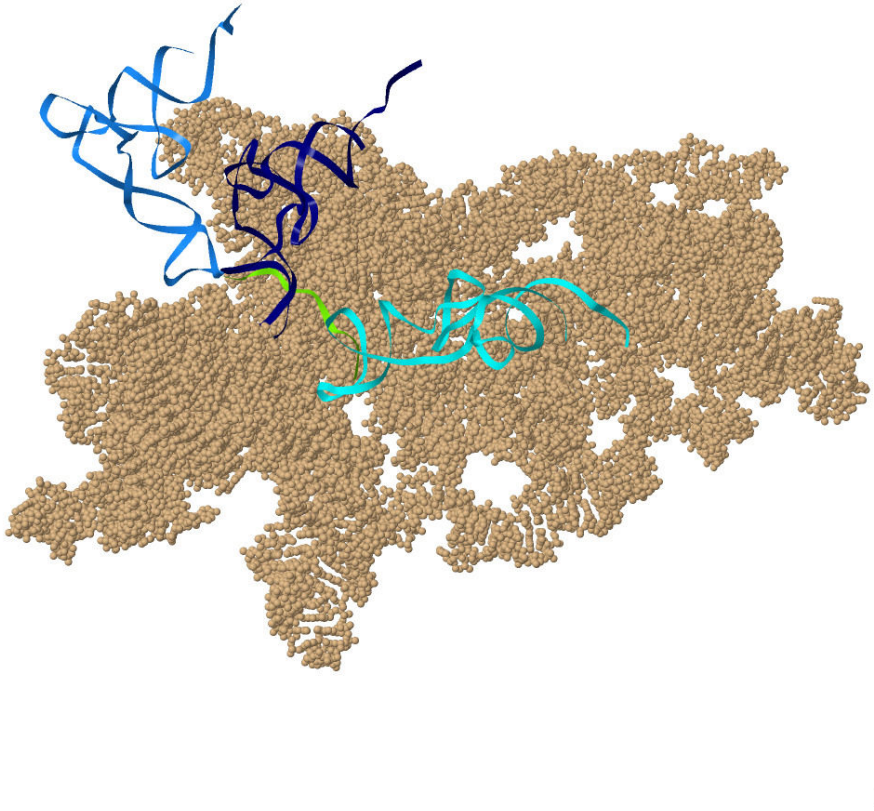

Figure 1.6: Crystallographic structure of the smaller subunit of *Escherichia coli* ribosome. The 3D positioning of mRNA in green, the three tRNAs in different shades of blue and the smaller ribosomal subunit (Generated using Jmol).

Determining the structure of a protein doesn't rely on these methods only, in most cases prior knowledge is necessary. Such as amino acid sequence and preferred geometry of atoms.

1.4 Biological databases

We will be presenting now the different biological databases used to retrieve the data on which the studies were conducted. For the first part of the study we used crystallographic data to examine if what we are searching for is in significant regions of the ribosome. In the second part of the study, we used large quantity of data of genomes to thoroughly examine the circular code in eukaryotic organisms.

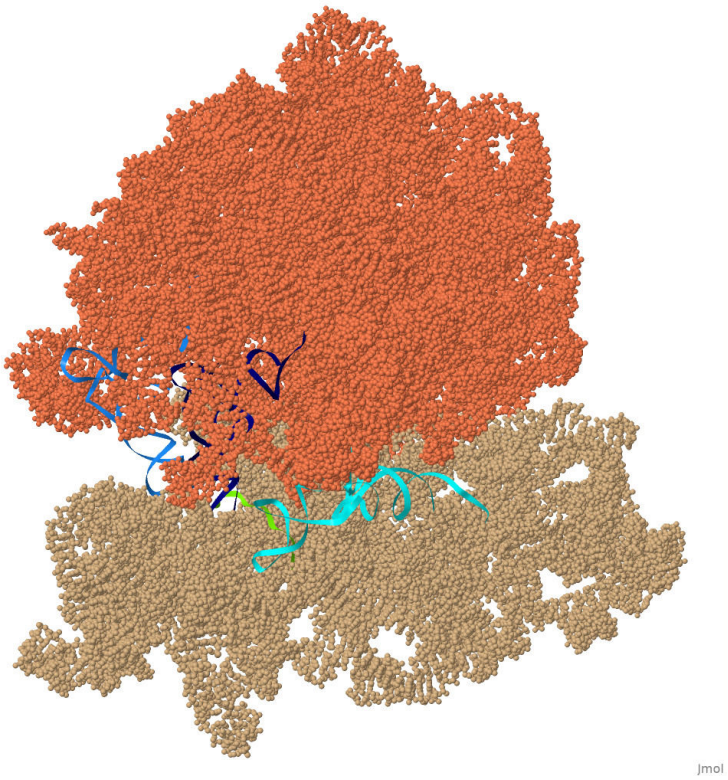

Figure 1.7: Crystallographic structure of the *Escherichia coli* ribosome. The 3D positioning of mRNA in green, the three tRNAs in different shades of blue, the smaller ribosomal subunit in light brown and the large subunit in orange (Generated using Jmol).

1.4.1 Crystallographic database

Crystallographic databases are created with the goal of collecting information about the structure of molecules and crystals. The protein data bank (PDB) archive is such a database (www.rcsb.org). Structural biologists use methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy to construct a 3D structure of proteins. This information is deposited in the PDB, which is then annotated and publicly released into the archive by the worldwide protein data bank (wwPDB), an organization that is responsible with maintaining the archive.

PDB holds structures for proteins and nucleic acids, such as ribosomes, oncogenes, drug targets and complete viruses. Multiple structures can exist for the same molecule depending on the test conducted or the scope of the study. The files found in the database consist in principle of, the atoms in each protein and their three-dimensional coordinates, a header that summarizes the input and the experiments in which this structure was acquired.

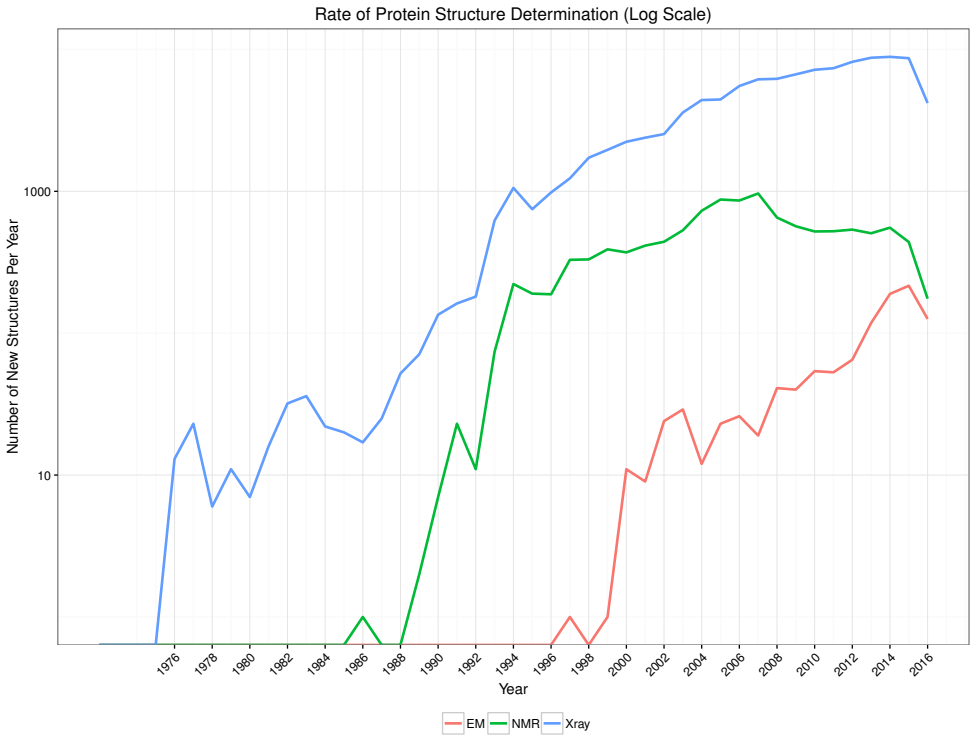

Figure 1.8: The increase of crystallographic structure submissions to the PDB according to methods. X-ray crystallography (blue) and Nuclear Magnetic Resonance (green) were the dominant methods while Electron Microscopy (red) is becoming more popular with the years (numbers provided by PDB).

Molecular graphics software are available to visualize these files in 3D, such as Jmol, RasMol, Swiss PDB viewer, ...etc, with additional features such as measuring distances and bond angles. These software allow us to carefully study and structure and identify interesting structural features.

1.4.2 Sequence database

The National Center for Biotechnology Information (NCBI) boasts an array of biological database (www.ncbi.nlm.nih.gov), bioinformatic tools and services. Particularly we are interested in Reference Sequence (RefSeq) and Genbank databases.

RefSeq (www.ncbi.nlm.nih.gov/refseq) is a large multi-species, non-redundant, curated sequence database consisting everything from transcripts and translation products to whole genomes. While the database is non-redundant, alternate assemblies of the same sequence can exist. RefSeq employs a strict curative process where a record may be an essentially unchanged, validated copy of the original submission, or include updated or additional information supplied

by collaborators or NCBI staff.

GenBank (www.ncbi.nlm.nih.gov/genbank), on the other hand, a redundant archival database is an annotated collection of all publicly available nucleotide sequences and their protein translations submitted directly by individual laboratories, as well as from bulk submissions from large-scale sequencing centres. GenBank continues to grow at an exponential rate, doubling every 10 months.

1.5 Summary

In this chapter we presented the nucleic acid sequences (DNA and RNA) and their composition. Then, we preceded to explain the translation process inside the ribosome and how a protein is synthesised inside a cell. Finally, we shed light on the bimolecular structure of a ribosome. This biological introduction is to help us better understand how we study circular code in a nucleic acid sequence. The 3D structure of a ribosome is vital for our first part of the study, where we focus on the decoding center of the ribosome and then move on to enlarge the scope of the study to examine the area around the decoding center and possible interactions with the tRNAs present in a ribosome at the time of translation.

We mentioned as well the nature of data we are working with. Crystallographic structures were an essential part when it came to figure out the working mechanics of a ribosome. As such, crystallographic databases were vital for us to understand and study circular code motifs presence in a ribosome. Finally, the sequence database RefSeq that houses a huge amount of complete chromosomes proved excellent for us to conduct a study on the entire set of eukaryotic genomes published at the time.

In the next chapter, we will start with a brief history of codes, and how the discovery of DNA structure ushered the race to crack the genetic code. Afterwards, we will present the circular code, how it was discovered and why it is an interesting study due to the many properties it has.

2

Code Theory

2.1 Introduction 18
2.2 History of codes 18
2.2.1 Diamond code 18
2.2.2 Comma-free code 19
2.2.3 Genetic code 21
2.3 Circular code 21
2.3.1 Circular code motifs 25
2.3.2 Unitary circular codes of dinucleotides, trinucleotides and tetranucleotides 26
2.4 Summary 29

2.1 Introduction

In this chapter we will be explaining what is circular code, how it came to be, and in order to do that we will begin with a brief history of how code theory began and its purpose. We will mention codes that were very important at the time of their inception and why they were dismissed eventually. Finally, we explain what are circular code motifs while shedding light a subclass called unitary circular codes that are interesting for us.

2.2 History of codes

The discovery of the double-helix structure of DNA (Watson and Crick, 1953) raised the question of how to translate a 4 letter alphabet into 20 words? The first deduction was it cannot be a one-to-one mapping, and a two letter word results in 16 words, still short of 20. Therefore, the representation could not be smaller than a three letter word (trinucleotide). But that would also give 64 words, which is an excess to the 20 amino acids. Scientist shortly rushed to crack the secret of genetic expression. This lead to the publication of several codes hoping to answer this question.

2.2.1 Diamond code

The first to propose a coding scheme following the Watson-Crick structure was George Gamow, better known for his work on the Big Bang theory.

Called the diamond code (Gamow, 1954), it suggested that double-stranded DNA acted directly as a template for assembling amino acids into proteins. Gamow envisioned the grooves in the double helix as holes that would fit the side chains of amino acids in a ”key-and-lock” fashion (Figure 2.1). Even though the diamond have four corners, only three of them are utilized because the paired bases on the horizontal diagonal are complementary. This is essence makes the diamond code a triplet code. Figure 2.1 show that there is 20 distinct holes.

Gamow reasoned that 12 holes are symmetrical, these diamonds could be flipped end-for-end or flopped side-to-side without changing their meaning. This allowed the possibility of several triplets coding for the same amino acid, thus dealing with the problem of having 64 trinucleotides and 20 amino acids.

This code had another feature that lead to its eventual dismissal. Each nucleotide was simultaneously present in three trinucleotides, for example, a

base sequence $ACGTAA$ would result in four overlapping trinucleotides: ACG, CGT, GTA, GTA. Which was proven to be wrong.

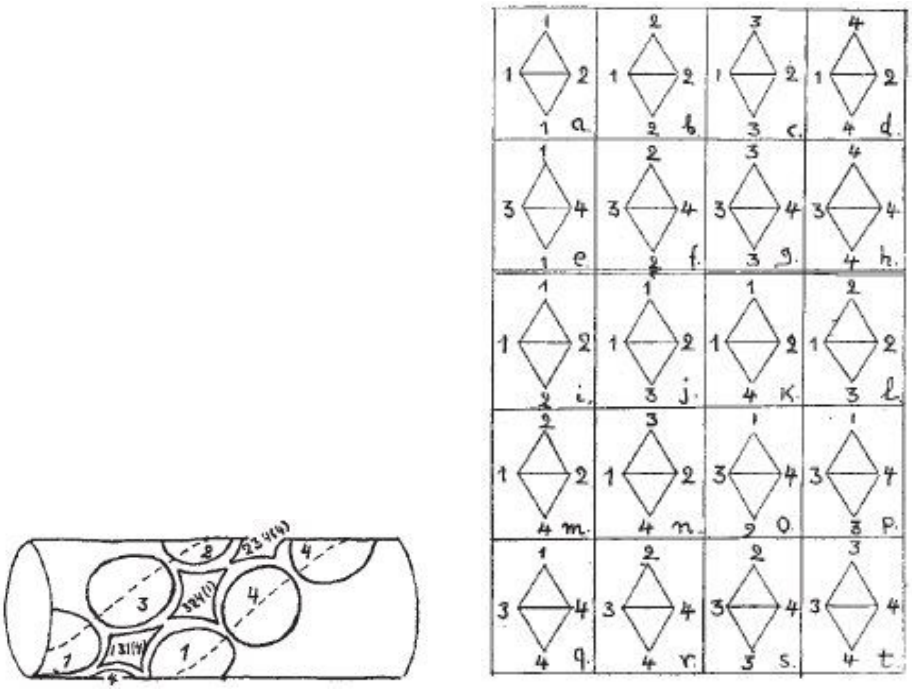

Figure 2.1: George Gamow's diamond code used the key-and-lock between the grooves found in DNA and the amino acids (Gamow, 1954).

2.2.2 Comma-free code

Proposed by Crick, Griffith, and Orgel, 1957, comma-free code was an effort to study the encoding of three nucleotides $\{AAA, AAC, \ldots, TTT\}$ into amino acids. By excluding the four periodic permuted trinucleotides $\{AAA, CCC, GGG, TTT\}$ and distributing the 60 remaining trinucleotides into 20 groups of three trinucleotides such that each group contains the set of trinucleotides that can be permuted from each other following the circular permutation map (Definition 2.1). Based on this we can deduce that a comma-free code can have only one trinucleotide from each group, therefore a set contains at most 20 trinucleotides.

Notation 2.1. The nucleotides define the genetic alphabet $B = \{A, C, G, T\}$. The set of non-empty words (words, respectively) over B is denoted by B^+ (B^*, respectively). The set of the 64 words of length 3 (trinucleotides or triletters) on B is denoted by $B^3 = \{AAA, \ldots, TTT\}$. Let $x_1 \cdots x_n$ be the concatenation of the words x_i for $i = 1, \ldots, n$, the symbol " $\cdot$ " being the concatenation operator.

Notation 2.2. In genes, there are three frames f. By convention here, the reading frame $f = 0$ is established by a start codon $\{ATG, GTG, TGT, TTG\}$ and the frames $f = 1$ and $f = 2$ are the reading frame $f = 0$ shifted by one and two nucleotides in the 5'-to-3' (left to right) direction, respectively.

Definition 2.1. The trinucleotide circular permutation map $\mathcal{P} : B \to B$ is defined by $\mathcal{P}(l_0 \cdot l_1 \cdot l_2) = l_1 \cdot l_2 \cdot l_0$ for all $l_0, l_1, l_2 \in B$, e.g. $\mathcal{P}(ATG) = TGA$. The second iterate of $\mathcal{P}$ is denoted as $\mathcal{P}^2$, e.g. $\mathcal{P}^2(ATG) = GAT$. By extension to a trinucleotide set S, the set circular permutation map $\mathcal{P} : \mathbb{P}(B^3) \to \mathbb{P}(B^3)$, $\mathbb{P}$ being the set of all subsets of B^3, is defined by $\mathcal{P}(S) = \left\{v | u, v \in B^3, u \in S, v = \mathcal{P}(u)\right\}$, i.e. a permuted trinucleotide set $\mathcal{P}(S)$ is obtained by applying the circular permutation map $\mathcal{P}$ to all its trinucleotides, e.g. $\mathcal{P}(\{ACG, AGT\}) = \{CGA, GTA\}$ and $\mathcal{P}^2(\{ACG, AGT\}) = \{GAC, TAG\}$.

Despite having an identical number of trinucleotides as amino acids, no trinucleotide comma-free code was identified in genes. Early in the sixties it was discovered that the trinucleotides TTT, an excluded trinucleotide in comma-free code, in fact codes phenylalanine (Nirenberg and Matthaei, 1961), this in turn would lead to the abandonment of comma-free code.

Definition 2.2. A set $S \subset B^+$ of words is a code if, for each $x_1, \ldots, x_n$, $y_1, \ldots, y_m \in S, n, m \geq 1$, the condition $x_1 \cdots x_n = y_1 \cdots y_m$ implies $n = m$ and $x_i = y_i$ for $i = 1, \ldots, n$, e.g. $B^3 = \{AAA, ..., TTT\}$ is a code, where as $X = \{A, GC, AGC\}$ is not a code as there are two decompositions $A{\cdot}GC = AGC$.

Definition 2.3. (Fimmel, Michel, and Strüngmann, 2016) Let $X \subseteq B^m$, $m \in \mathbb{N}$ with $m \geq 2$, be an m-nucleotide code. The directed graph $\mathcal{G}(X) = (V(X), E(X))$ associated with X has a set of vertices $V(X)$ and a set of edges $E(X)$ defined as follows:

$$\begin{cases} V(X) = \{N_1 \ldots N_i, N_{i+1} \ldots N_m : N_1 \ldots N_m \in X, 1 \leq i \leq m-1\} \\ E(X) = \{[N_1 \ldots N_i, N_{i+1} \ldots N_m] : N_1 \ldots N_m \in X, 1 \leq i \leq m-1\}. \end{cases} \tag{2.1}$$

Theorem 2.1. (Fimmel, Michel, and Strüngmann, 2016) Given an m-nucleotide code $X \subseteq B^m$, $m \geq 2$, the following statements are equivalent:

1. The code X is comma-free.
2. The maximal length of a path in $\mathcal{G}(X)$ is 2.

Theorem 2.2. (Fimmel, Michel, and Strüngmann, 2016) Given an m-nucleotide code $X \subseteq B^m$, $m \geq 2$, the code X is strong comma-free if the maximal length of a path in $\mathcal{G}(X)$ is 1.

2.2.3 Genetic code

In 1961, Marshall Nirenberg and Heinrich Matthaei managed to crack the first word of the genetic code. They performed an experiment which showed that a chain of the repeating bases U (Uracil) forced a protein chain made of one repeating amino acid, phenylalanine. This was a breakthrough experiment which proved that the code could be broken.

After the initial discovery the team grew in size to replicate the poly-U experiment model to other amino acids. Using 20 test tubes for each amino acid, respectively, they experimented with the different 64 combination of nucleotides to form three letter words.

By 1966 the genetic code was complete, all the mappings between the 64 codons and the 20 amino acids were established. As it turns out, there was no pattern in the code. Some amino acids were represented by one or two codons, some by more, ignoring all mathematical approaches to solving this coding mystery.

2.3 Circular code

In 1996, a statistical analysis of occurrence frequencies of the 64 trinucleotides was conducted in the three frames (Definition 2.2), of genes of both prokaryotes and eukaryotes. The study showed that the trinucleotides are not uniformly distributed in the three frames (Arquès and Michel, 1996). By convention here, the frame zero is the reading frame in a gene, and the frames 1 and 2 are the reading frame 0 shifted by 1 and 2 nucleotides in the 5'-to-3' direction, respectively. By excluding the four periodic permuted trinucleotides $\{AAA, CCC, GGG, TTT\}$ and by assigning each trinucleotide to a preferential frame (frame of its highest occurrence frequency), three subsets $X = X_0$, X_1, and X_2 of 20 trinucleotides each, were assigned to frame 0, 1 and 2, respectively (Table 2.1). The analysis was based on the large gene populations (protein coding regions) of eukaryotes (26,757 sequences, 11,397,678 trinucleotides) and prokaryotes (13,686 sequences, 4,709,758 trinucleotides) (Arquès and Michel, 1996). The following circular code X was observed in frame 0 (reading frame):

Table 2.1: Frequency of trinucleotides occurrences according to frame 0, 1 and 2 in genes sequences from eukaryotes and prokaryotes (Arquès and Michel, 1996). The table shows the occurrences of the first seven trinucleotides from B^3, with their preferred frame in bold.

	Frequency(%)		
Trinucleotide	Frame 0	Frame 1	Frame 2
AAA	**3.38**	2.75	2.44
AAC	**2.18**	1.59	1.38
AAG	1.98	**3.21**	0.81
AAT	**2.17**	1.37	1.69
ACA	1.22	**1.91**	1.11
ACC	**2.09**	1.60	0.79
ACG	1.30	**2.49**	0.68
...	...	...	...

$$X = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, \\ GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}. \tag{2.2}$$

The two sets X_1 and X_2 , of 20 trinucleotides each, in the shifted frames 1 and 2 of genes can be deduced from X by the circular permutation map (Definition 2.1), i.e. $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$.

Definition 2.4. The nucleotide complementarity map $\mathcal{C} : B \to B$ is defined by $\mathcal{C}(A) = T$, $\mathcal{C}(C) = G$, $\mathcal{C}(T) = A$ and $\mathcal{C}(G) = C$. According to the property of the complementary and anti-parallel double helix, the trinucleotide complementary map $\mathcal{C} : B^3 \to B^3_4$ is defined by $\mathcal{C}(l_0 \cdot l_1 \cdot l_2) = \mathcal{C}(l_2) \cdot \mathcal{C}(l_1) \cdot \mathcal{C}(l_0)$ for all $l_0, l_1, l_2 \in B$, e.g. $\mathcal{C}(ATG) = CAT$. By extension to a trinucleotide set S, the set complementarity map $\mathcal{C} : \mathbb{P}(B^3) \to \mathbb{P}(B^3)$, is defined by $\mathcal{C}(S) = \Big\{v \mid u, v \in B^3, u \in S, v = \mathcal{C}(u)\Big\}$, i.e. a complementary trinucleotide set $\mathcal{C}(S)$ is obtained by applying the complementarity map $\mathcal{C}$ to all its trinucleotides, e.g. $\mathcal{C}(\{ACG, AGT\}) = \{ACT, CGT\}$.

Definition 2.5. A trinucleotide code $X \subset B^3$ is circular code if, for each $x_1, \ldots, x_n, y_1, \ldots, y_m \in X, n, m \geq 1, r \in B^*, s \in B^+$, the conditions $sx_2 \cdots x_n r = y_1 \cdots y_m$ and $x_1 = rs$ imply $n = m, r = \varepsilon$ (empty word) and $x_i = y_i$ for $i = 1, \ldots, n$ (Figure 2.2 for a graphical representation of the circular code definition).

Theorem 2.3. (Fimmel, Michel, and Strüngmann, 2016) Given an m-nucleotide code $X \subseteq B^m$, $m \in \mathbb{N}$ with $m \geq 2$, the following statements are equivalent:

1. The code X is circular.
2. The graph $\mathcal{G}(X)$ is acyclic.

Definition 2.6. An m-nucleotide unitary circular code $X \subseteq B^m$ (UCC) is a code with a unique word.

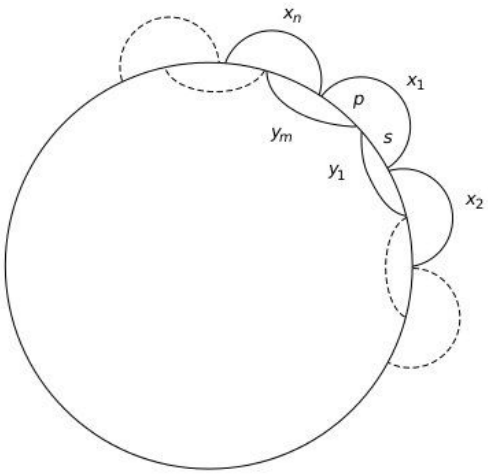

Figure 2.2: Graphical representation of the circular code definition (Arquès and Michel, 1996).

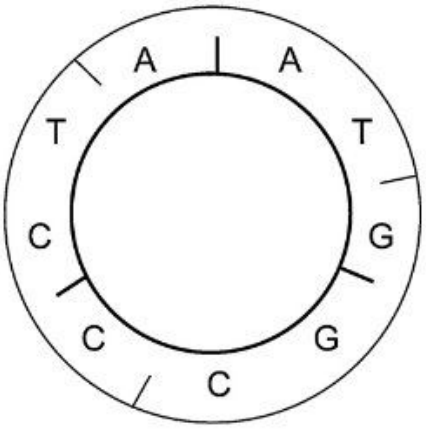

Figure 2.3: The following code $\{AAT, ATG, CCT, CTA, GCC, GGC\}$ is not circular, since (ATG, GCC, CTA) can be read differently if we shift frames showing us (AAT, GGC, CCT) (Michel, 2012).

Remark 2.1. A trinucleotide code C containing either one periodic permuted trinucleotide $PPT = \{AAA, CCC, GGG, TTT\}$ or two non-periodic permuted trinucleotides $NPPT = \{t, \mathcal{P}(t)\}$ for a trinucleotide $t \in B^3 \backslash PPT$ cannot be circular. Thus, the two trinucleotide codes B^3 and $B^3 \backslash PPT$ are not circular.

Remark 2.2. The fundamental property of a circular code is the ability to retrieve the reading (original or construction) frame of any sequence generated with this circular code. A circular code is a set of words over an alphabet such that any sequence written on a circle (the next letter after the last letter of the sequence being the first letter) has a unique decomposition (factorization) into words of the circular code (Michel, 2012) (Figure 2.3 for an example). The reading frame in a sequence (gene) is retrieved after the reading of a certain number of letters (nucleotides), called the window of the circular code. The length of this window

for retrieving the reading frame is the letter length of the longest ambiguous word, not necessarily unique, which can be read in at least two frames, plus one letter (Figure 2.4).

Figure 2.4: An example of how a window can determine in which frame the circular code motif is, if size is sufficient (13 nucleotides). Retrieval of the reading frame of the word $w = \ldots AGGTAATTACCAG \ldots$ of the trinucleotide circular code X (Equation 2.2). Among the three possible factorizations $\tilde{w}_0, \tilde{w}_1$ and $\tilde{w}_2$, only one factorization $\tilde{w}_1$ into trinucleotides of X is possible leading to $\ldots A \cdot GGT \cdot AAT \cdot TAC \cdot CAG \cdot \ldots$ Thus, the first letter A of w is the 3rd letter of a trinucleotide of X (Michel, 2012).

Remark 2.3. At window length $l > 12$ nucleotides there is no ambiguous word of the circular code X.

Definition 2.7. A trinucleotide set $X_1 = \mathcal{P}(X)$ of a trinucleotide circular code X is permuted if, for each $x \in X, \mathcal{P}(x) \in \mathcal{P}(X)$. The permuted trinucleotide set $X_2 = \mathcal{P}^2(X)$ is defined similarly.

Definition 2.8. A trinucleotide circular code $X \subset B^3$ is maximal if for each $x \in B^3, x \notin X, X \cup \{x\}$ is not a trinucleotide circular code.

Definition 2.9. A trinucleotide circular code X is self-complementary if, for each $x \in X, \mathcal{C}(x) \in X$.

Definition 2.10. An m-nucleotide circular code $X \subseteq B^m$ is C^n if the m permuted m-nucleotide codes $X_1 = \mathcal{P}(X), \ldots, X_m = \mathcal{P}^m(X)$ are circular. An m-nucleotide comma-free $X \subseteq B^m$ (strong comma-free, respectively) is CF^m (SCF^m, respectively) if the m permuted m-nucleotide codes $X_1 = \mathcal{P}(X), \ldots, X_m = \mathcal{P}^m(X)$ are comma-free (strong comma-free, respectively). A trinucleotide circular code $X \subset B^3$ is C^3 (m = 3) if the permuted trinucleotide sets $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$ are trinucleotide circular codes.

Definition 2.11. A trinucleotide circular code $X \subset B^3$ is C^3 self-complementary if X, $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$ are trinucleotide circular codes satisfying the following properties $X = \mathcal{C}(X)$, $\mathcal{C}(X_1) = X_2$ and $\mathcal{C}(X_2) = X_1$.

Table 2.2: The number of sets that can be circular for each class of trinucleotide (Arquès and Michel, 1996; Michel and Pirillo, 2010; Michel, Pirillo, and Pirillo, 2012).

Circular code (maximal)	Number
Potential	3 486 784 401
Identified	12 964 440
Self-complementary	528
C^3	221 328
C^3 Self-complementary	216

A circular code containing 20 codons is designated as maximal (Definition 2.8). The number possible maximal circular codes is 3^{20} out of 64^{20}, i.e. probability of 10^{-27}, only 12 964 440 are effectively circular codes. Which makes finding these three circular code sets in genes quite intriguing. These three trinucleotide sets posses several strong mathematical properties. They have the fundamental property to always retrieve the reading frame in any position of any sequence generated with the circular code. Initiation and stop trinucleotides, as well any frame signals are not necessary to define the reading frame. A window of 13 nucleotides allows to retrieve the reading frame for all the ambiguous words generated with X. Therefore, circular codes are less constrained than the comma-free codes. Moreover, the X_0 code is in particular very interesting since it is symmetric under the complementary transformation. In other words, we exchange every trinucleotide in the set with its complement (Definition 2.4), the set remains unchanged. While the X_1 and X_2 sets are complementary to each other.

2.3.1 Circular code motifs

A circular code motif is a phrase composed by words, in this case trinucleotides, from a circular code. Consequently, an X motif can be found in any DNA or RNA sequence, where we have successive trinucleotides from the X. As mentioned above, a window of 13 nucleotides is sufficient to distinguish a circular code motif in any given frame. Let us examine the following example:

Example 2.1.

Sequence: $AGTCAGTAGCTGAGGCAGCTCGAAATTCGT$

Comma separated: $AGT, CAG, TAG, CTG, AGG, CAG, CTC, GAA, ATT, CGT$

Consider the DNA sequence in example 2.1. If we read this sequence in frame 0, we can see the four following consecutive trinucleotides $\{CAG, CTC, GAA, ATT\}$ belonging to X (Equation 2.2). Therefore, $CAGCTCGAAATT$ is an X motif that starts at nucleotide 16 and ends at nucleotide 27 according to the sequence.

Definition 2.12. An m-nucleotide unitary circular code motif (UCC motif) generated by an m-nucleotide unitary circular code (Definition 2.6), is a concatenation of n words w of size m denoted w^n. The class of the motifs w^n for all n is denoted by m^+.

Definition 2.13. Two UCC motifs w_1^+ and w_2^+ are said equivalent if w_1^+ and w_2^+ are related by the circular permutation map $\mathcal{P}$ (Definition 2.1).

2.3.2 Unitary circular codes of dinucleotides, trinucleotides and tetranucleotides

The $4^2 - 4 = 12$ dinucleotide unitary codes $Di = \{l_1l_2\}$ with $l_1l_2 \in B$ and $l_1 \neq l_2$ are circular and C^2 . The $4^3 - 4 = 60$ trinucleotide unitary codes $Tri = \{l_1l_2l_3\}$ with $l_1, l_2, l_3 \in B$ and $l_1l_2 \neq l_2l_3$ are also circular and C^3. The $4^4-16 = 240$ tetranucleotide unitary codes $Tetra = \{l_1l_2l_3l_4\}$ with $l_1, l_2, l_3, l_4 \in B$ and $l_1l_2 \neq l_3l_4$ are also circular and C^4 (excluding $\{l_1l_2l_1l_2\}$ since it is not circular, $(l_1l_2)^2$ is a dinucleotide repeat of $\{l_1l_2\}$). We describe some additional stronger combinatorial properties for the codes Di, Tri and $Tetra$. From Theorem 2.2, an m-nucleotide unitary code $\{l_i \dots l_i\}$ with $l_i \in B$, i.e. starting and ending by the same letter, cannot be strong comma-free.

2.3.2.1 Unitary circular codes of dinucleotides

The 12 dinucleotide unitary codes $\{l_1l_2\}$ and $\{P(l_1l_2)\} = \{l_2l_1\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$, i.e. $\{AC\}$, $\{AG\}$, $\{AT\}$, $\{CA\}$, $\{CG\}$, $\{CT\}$, $\{GA\}$, $\{GC\}$, $\{GT\}$, $\{TA\}$, $\{TC\}$ and $\{TG\}$, are strong comma-free (SCF) by Theorem 2.2. Thus, a dinucleotide unitary strong comma-free code $\{l_1l_2\}$ has a permuted code $\{l_2l_1\}$ which is also strong comma-free (SCF^2 property, see Definition 2.10). Furthermore, the four dinucleotide unitary strong comma-free codes $\{AT\}, \{CG\}, \{GC\}$ and $\{TA\}$ are self- complementary.

2.3.2.2 Unitary circular codes of trinucleotides

The 24 trinucleotide unitary codes $\{l_1l_2l_3\}$, $\{\mathcal{P}(l_1l_2l_3)\} = \{l_2l_3l_1\}$ and $\{\mathcal{P}^2(l_1l_2l_3)\} = \{l_3l_1l_2\}$ with $l_1, l_2, l_3 \in B$ and $l_1 \neq l_2 \neq l_3$, i.e. $\{ACG\}$, $\{ACT\}$, $\{AGC\}$, $\{AGT\}$, $\{ATC\}$, $\{ATG\}$, $\{CAG\}$, $\{CAT\}$, $\{CGA\}$, $\{CGT\}$, $\{CTA\}$, $\{CTG\}$, $\{GAC\}$, $\{GAT\}$, $\{GCA\}$, $\{GCT\}$, $\{GTA\}$, $\{GTC\}$, $\{TAC\}$, $\{TAG\}$, $\{TCA\}$, $\{TCG\}$, $\{TGA\}$ and $\{TGC\}$, are strong comma-free (SCF) by Theorem 2.2. Thus, a trinucleotide unitary strong comma-free code $\{l_1l_2l_3\}$ has two permuted codes $\{l_2l_3l_1\}$ and $\{l_3l_1l_2\}$ which are also strong comma-free (SCF^3 property, see Definition 2.10).

The 24 trinucleotide unitary codes $\{l_1l_1l_2\}$ and $\{\mathcal{P}^2(l_1l_1l_2)\} = \{l_2l_1l_1\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$, i.e. $\{AAC\}$, $\{AAG\}$, $\{AAT\}$, $\{ACC\}$, $\{AGG\}$, $\{ATT\}$, $\{CAA\}$, $\{CCA\}$, $\{CCG\}$, $\{CCT\}$, $\{CGG\}$, $\{CTT\}$, $\{GAA\}$, $\{GCC\}$, $\{GGA\}$, $\{GGC\}$, $\{GGT\}$, $\{GTT\}$, $\{TAA\}$, $\{TCC\}$, $\{TGG\}$, $\{TTA\}$, $\{TTC\}$ and $\{TTG\}$, are strong comma-free (SCF) by Theorem 2.2. The 12 trinucleotide unitary codes $\{\mathcal{P}(l_1l_1l_2)\} = \{l_1l_2l_1\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$, i.e. $\{ACA\}$, $\{AGA\}$, $\{ATA\}$, $\{CAC\}$, $\{CGC\}$, $\{CTC\}$, $\{GAG\}$, $\{GCG\}$, $\{GTG\}$, $\{TAT\}$, $\{TCT\}$ and $\{TGT\}$, are comma-free (CF) by Theorem 2.1. Thus, a trinucleotide unitary strong comma-free code $\{l_1l_1l_2\}$ has one permuted code $\{l_2l_1l_1\}$ which is also strong comma-free and one permuted code $\{l_1l_2l_1\}$ which is comma-free code. Corollary, a trinucleotide unitary comma-free code $\{l_1l_2l_1\}$ has two permuted codes $\{l_1l_1l_2\}$ and $\{l_2l_1l_1\}$ which are strong comma-free.

2.3.2.3 Unitary circular codes of tetranucleotides

The 24 tetranucleotide unitary codes $\{l_1l_2l_3l_4\}$, $\{\mathcal{P}(l_1l_2l_3l_4)\} = \{l_2l_3l_4l_1\}$, $\{\mathcal{P}^2(l_1l_2l_3l_4)\} = \{l_3l_4l_1l_2\}$ and $\{\mathcal{P}^3(l_1l_2l_3l_4)\} = \{l_4l_1l_2l_3\}$ with $l_1, l_2, l_3, l_4 \in B$ and $l_1 \neq l_2 \neq l_3 \neq l_4$ are strong comma-free (SCF) by Theorem 2.2. Thus, a tetranucleotide unitary strong comma-free code $\{l_1l_2l_3l_4\}$ has three permuted codes $\{l_2l_3l_4l_1\}$, $\{l_3l_4l_1l_2\}$ and $\{l_4l_1l_2l_3\}$ which are also strong comma-free (SCF^4 property, see Definition 2.1). The 48 tetranucleotide unitary codes $\{l_1l_2l_1l_3\}$, $\{\mathcal{P}(l_1l_2l_1l_3)\} = \{l_2l_1l_3l_1\}$, $\{\mathcal{P}^2(l_1l_2l_1l_3)\} = \{l_1l_3l_1l_2\}$ and $\{\mathcal{P}^3(l_1l_2l_1l_3)\} = \{l_3l_1l_2l_1\}$ with $l_1, l_2, l_3 \in B$ and $l_1 \neq l_2 \neq l_3$ are strong comma-free (SCF) by Theorem 2.2. Thus, a tetranucleotide unitary strong comma-free code $\{l_1l_2l_1l_3\}$ has three permuted codes $\{l_2l_1l_3l_1\}$, $\{l_1l_3l_1l_2\}$ and $\{l_3l_1l_2l_1\}$ which are also strong comma-free (SCF^4 property).

The 72 tetranucleotide unitary codes $\{l_1l_1l_2l_3\}$, $\{\mathcal{P}^2(l_1l_1l_2l_3)\} = \{l_2l_3l_1l_1\}$ and

$\{\mathcal{P}^3(l_1l_1l_2l_3)\} = \{l_3l_1l_1l_2\}$ with $l_1, l_2, l_3 \in B$ and $l_1 \neq l_2 \neq l_3$ are strong comma-free (SCF) by Theorem 2.2. The 24 tetranucleotide unitary codes $\{\mathcal{P}(l_1l_1l_2l_3)\}$ $= \{l_1l_2l_3l_1\}$ with $l_1, l_2, l_3 \in B$ and $l_1 \neq l_2 \neq l_3$ are comma-free (CF) by Theorem 2.1. Thus, a tetranucleotide unitary strong comma-free code $\{l_1l_1l_2l_3\}$ has two permuted codes $\{l_2l_3l_1l_1\}$ and $\{l_3l_1l_1l_2\}$ which are also strong comma-free and one permuted code $\{l_1l_2l_3l_1\}$ which is comma-free. Corollary, a tetranucleotide unitary comma-free code $\{l_1l_2l_3l_1\}$ has three permuted codes $\{l_1l_1l_2l_3\}$, $\{l_2l_3l_1l_1\}$ and $\{l_3l_1l_1l_2\}$ which are strong comma-free.

The 24 tetranucleotide unitary codes $\{l_1l_1l_1l_2\}$ and $\{\mathcal{P}^3(l_1l_1l_1l_2)\} = \{l_2l_1l_1l_1\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$ are strong comma-free (SCF) by Theorem 2.2. The 24 tetranucleotide unitary codes $\{\mathcal{P}(l_1l_1l_1l_2)\} = \{l_1l_1l_2l_1\}$ and $\{\mathcal{P}^2(l_1l_1l_1l_2)\} =$ $\{l_1l_2l_1l_1\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$ are comma-free (CF) by Theorem 2.1. Thus, a tetranucleotide unitary strong comma-free code $\{l_1l_1l_1l_2\}$ has one permuted code $\{l_2l_1l_1l_1\}$ which is also strong comma-free and two permuted codes $\{l_1l_1l_2l_1\}$ and $\{l_1l_2l_1l_1\}$ which are comma-free. Corollary, a tetranucleotide unitary comma-free code $\{l_1l_1l_2l_1\}$ has one permuted code $\{l_1l_2l_1l_1\}$ which is also comma-free and two permuted codes $\{l_1l_1l_1l_2\}$ and $\{l_2l_1l_1l_1\}$ which are strong comma-free. The 12 tetranucleotide unitary codes $\{l_1l_1l_2l_2\}$ and $\{\mathcal{P}^2(l_1l_1l_2l_2)\} = \{l_2l_2l_1l_1\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$ are strong comma-free (SCF) by Theorem 2.2.

The 12 tetranucleotide unitary codes $\{\mathcal{P}(l_1l_1l_2l_2)\} = \{l_1l_2l_2l_1\}$ and $\{\mathcal{P}^3(l_1l_1l_2l_2)\} = \{l_2l_1l_1l_2\}$ with $l_1, l_2 \in B$ and $l_1 \neq l_2$ are comma-free (CF) by Theorem 2.1. Thus, a tetranucleotide unitary strong comma-free code $\{l_1l_1l_2l_2\}$ has one permuted code $\{l_2l_2l_1l_1\}$ which is also strong comma-free and two permuted codes $\{l_1l_2l_2l_1\}$ and $\{l_2l_1l_1l_2\}$ which are comma-free. Corollary, a tetranucleotide unitary comma-free code $\{l_1l_2l_2l_1\}$ has one permuted code $\{l_2l_1l_1l_2\}$ which is also comma-free and two permuted codes $\{l_1l_1l_2l_2\}$ and $\{l_2l_2l_1l_1\}$ which are strong comma-free.

Furthermore, the 12 tetranucleotide unitary strong comma-free codes $\{AATT\}$, $\{ACGT\}$, $\{AGCT\}$, $\{CATG\}$, $\{CCGG\}$, $\{CTAG\}$, $\{GATC\}$, $\{GGCC\}$, $\{GTAC\}$, $\{TCGA\}$, $\{TGCA\}$ and $\{TTAA\}$ are self-complementary. There is no tetranucleotide unitary comma-free code which is self-complementary.

2.4 Summary

We briefly gave the history of the genetic code, starting from its root with the discovery of the structure of DNA, through the various code theories, such as diamond code and comma-free code, that were proposed to help solve the mystery of nucleotides coding for amino acids. This shows us how circular code was discovered in 1996, why it is still relevant and an interesting study subject even though the genetic code has been already discovered and proven. We presented the circular code with its features and properties. We defined the unitary circular code, explaining how simple repeats are circular code. This allows us to study the large amount of sequences that has low-complexity DNA.

In the next chapter we will present the data that was retrieved for ribosomal RNA and eukaryotic genomes. For each environment we developed different algorithms to extract the results we need. Furthermore, several methods were adopted to help use analyse the results obtained. For this we divided our context into: (i) X circular code motifs in ribosomal RNA, (ii) X circular code motifs in eukaryotic genomes (iii) unitary circular codes (simple repeats) in eukaryotic genomes and (iv) trinucleotide pairs in gene sequences.

3

Data and Methods

3.1 Introduction . . . 33
3.2 Data acquisition . . . 33
3.2.1 Ribosomal data . . . 33
3.2.2 Genomic database . . . 34
3.3 Circular code motifs . . . 36
3.3.1 Search algorithms . . . 36
3.3.1.1 Biinfinite word . . . 36
3.3.1.2 Sets algorithm . . . 37
3.3.1.3 Frames algorithm . . . 41
3.3.2 Definition of random code motifs . . . 42
3.3.3 Definition of the 23 bijective circular codes motifs . . . 43
3.3.3.1 Bijective transformation circular codes . . . 43
3.3.3.2 Main properties of the 23 bijective transformation circular codes . . . 47
3.3.4 Statistical analysis of X circular code motifs . . . 47
3.3.4.1 Coverage of X motifs in tRNA . . . 47
3.3.4.2 Expectation of the occurrence number of a motif . . . 48
3.3.4.3 Ratio of X motifs in coding and non-coding regions . . . 48
3.4 Ribosome study tools . . . 50
3.4.1 Multiple sequence alignment . . . 50
3.4.2 Molecule viewer . . . 50
3.5 Repeated motifs . . . 52

3.5.1 Dinucleotide unitary circular code motifs 52
3.5.2 Trinucleotide unitary circular code motifs 52
3.5.3 Tetranucleotide unitary circular code motifs 53
3.5.4 Statistical analysis of repeated motifs 53
3.5.4.1 Occurrence number of unitary circular code motifs . 53
3.5.4.2 Base number of unitary circular code motifs 55
3.5.4.3 Total base number of unitary circular code motifs . . 55
3.6 Occurrence number of trinucleotide pairs 56
3.7 Summary 59

3.1 Introduction

Previously, we explained the biological context in which we are working and the type of data needed for our study. Now we will show the actual data, how it was obtained, processed and dealt with it.

An extensive collection of classes and algorithms were written to handle our data, from reading PDB, alignments and huge FastA files, to searching sequences for motifs, all the way to the statistical tests applied to the found motifs. Our code has been refined and optimized to handle huge flat files with fast processing time.

We grouped together the search algorithms and statistical analysis tools depending on topic. We have a group that handles X circular code motifs and its comparison with various other codes. While another group aim at the study of unitary circular codes, otherwise known as simple repeats. Two sections handle the tools used for ribosome examination and another that studies the trinucleotide pairing in gene sequences.

3.2 Data acquisition

The used in this work can be divided along two different studies. The first focused on spatial interaction of circular code motifs, for which it required structural data of ribosomes from PDB archive. The second study concerned the search of circular code motifs on a genomic scale, which were retrieved from RefSeq database.

3.2.1 Ribosomal data

To study structural significance of circular code motifs, we collected ribosomal data from the before mentioned PDB archive in 2014 (Section 1.4.1, `www.rcsb.org/pdb`). The selected PDB entries have necessarily a bacterial 16S rRNA or a eukaryotic 18S rRNA. An entry from each organism having this criteria was chosen, with preference going towards entries having an mRNA and all or any tRNAs. This emphasize allows us to better study the spatial interaction of X motifs from rRNA, mRNA and tRNA (Section 4.2).

The studied PDB crystallographic structures are two bacterial entries: *Escherichia coli* (Brilot et al., 2013) and *Thermus thermophilus* (Jenner et al., 2010); one archaea: *Pyrococcus furiosus* (Armache et al., 2013); three nuclear eukaryotes: *Saccharomyces cerevisiae* (Armache et al., 2010a), *Triticum aes-*

Table 3.1: X circular code motifs studied in seven crystallographic structures of the Protein Data Bank PDB (www.rcsb.org/pdb). The main features of the studied crystallographic structures are given: PDB identification, kingdom, organism, type (16S for prokaryotes, 18S for eukaryotes) and base length (b) of rRNA, mRNA (Yes for available, No for unavailable), location of tRNA for the A, P, and E sites (No for unavailable).

					tRNAs		
PDB ID	Kingdom	*Organism*	rRNA	mRNA	A	P	E
3J5T	Bacteria	*Escherichia coli*	16S (1542 b)	Yes	Phe	Phe	No
3I8G	Bacteria	*Thermus thermophilus*	16S (1516 b)	Yes	Phe	Phe	Phe
3J20	Archaea	*Pyrococcus furiosus*	16S (1495 b)	No	No	Phe	Phe
3IZE	Eukaryote	*Saccharomyces cerevisiaie*	18S (1800 b)	Yes	No	Asp	No
3J5Z	Eukaryote	*Triticum aestivum*	18S (1810 b)	Yes	No	Asp	No
3J3D	Eukaryote	*Homo sapiens*	18S (1869 b)	Yes	No	No	Met
3BBN	Eukaryote	*Spinacia oleracea*	16S (1491 b)	No	No	No	No

tivum (Armache et al., 2010a,b; Gogala et al., 2014) and *Homo sapiens* (Anger et al., 2013); finally one chloroplast (eukaryotic organelle): *Spinacia oleracea* (Sharma et al., 2007).

3.2.2 Genomic database

Using BioPerl, we were able to retrieve all the eukaryotic chromosome sequences from the RefSeq database (Reference Sequence). The RefSeq is a curated non-redundant sequence database of genomes. We took one species from each genus and only complete genomic molecules (designated as NC), excluding alternate assembly. One strain from each species is considered.

Genomes with total size less than 400000 bases were filtered out (to avoid data results with null value). The following six genomes were dropped due to this: *Cryptomonas paramecium* (size = 82348 bases), *Encephalitozoon cuniculi* (size = 357485 bases), *Encephalitozoon hellem* (size = 245811 bases), *Encephalitozoon intestinalis* (size = 230782 bases), *Encephalitozoon romaleae* (size = 215619 bases) and *Nitzschia* (size = 14661 bases).

After this we retrieved the Genbank file associated with each chromosome which allowed us to extract the coordinates of the coding regions (CDS) mapped on that chromosome.

The outcome of this is 138 eukaryotic genomes (Sections 4.3, 4.4 and 4.5), displayed in Table 3.2. This genome information represents a total of 91,421,182,030 bases with 3,133,622,680 bases in coding regions (3.4%) and 88,287,559,350 bases for the non-coding regions (96.6%).

Table 3.2: Genomes characteristics.

Organism	Size (base)	CDS	non-CDS	*Organism*	Size	CDS	non-CDS
Anolis carolinensis	1081644591	16670366	1064974225	*Microtus ochrogaster*	1655383507	23979075	1631404432
Anopheles gambiae	24393108	1935976	22457132	*Monodelphis domestica*	2754317877	26405867	2727912010
Apis mellifera	219629612	16592730	203036882	*Mus musculus*	1205572488	14915903	1190656585
Arabidopsis thaliana	119146348	33175579	85970769	*Myceliophthora thermophila*	16385300	5976840	10408460
Aspergillus fumigatus	29384958	14214225	15170733	*Nasonia vitripennis*	116029644	13849880	102179764
Babesia bigemina	10271324	6801553	3469771	*Naumovozyma castellii*	11219539	8316040	2903499
Babesia bovis	4322739	2942868	1379871	*Naumovozyma dairenensis*	13527580	8618630	4908950
Babesia microti	6381289	4627831	1753458	*Callithrix jacchus*	2770219215	32941100	2737278115
Beta vulgaris	376583697	24577029	352006668	*Cyanidioschyzon merolae*	16546747	7429255	9117492
Bombus terrestris	216849342	16328176	200521166	*Esox lucius*	701024151	36362423	664661728
Bos taurus	2715765904	34037257	2681728647	*Leishmania mexicana*	30937689	15190689	15747000
Brachypodium distachyon	271776478	33348761	238427717	*Neospora caninum*	57547420	17793122	39754298
Brassica napus	775113993	94788198	680325795	*Phaeodactylum tricornutum*	26138756	13966979	12171777
Brassica oleracea	446885882	51062958	395822924	*Solanum lycopersicum*	802138220	33641774	768496446
Brassica rapa	256423463	48059517	208363946	*Neurospora crassa*	40463072	14868399	25594673
Caenorhabditis briggsae	91234787	20457533	70777254	*Nomascus leucogenys*	2795260045	32421857	2762838188
Caenorhabditis elegans	100272607	26613936	73658671	*Ogataea parapolymorpha*	8874589	7499949	1374640
Camelina sativa	578444267	95232658	483211609	*Oreochromis niloticus*	657350972	35450942	621900030
Candida dubliniensis	14618422	8917936	5700486	*Ornithorhynchus anatinus*	437080024	4691381	432388643
Candida glabrata	12318245	7914961	4403284	*Oryctolagus cuniculus*	2247752104	24420043	2223332061
Candida orthopsilosis	12659401	8468943	4190458	*Oryza brachyantha*	250923338	28480058	222443280
Canis lupus	2327633984	34021609	2293612375	*Oryza sativa*	382150945	30547069	351603876
Capra hircus	2524662720	30265609	2494397111	*Oryzias latipes*	723441489	33534282	689907207
Chlorocebus sabaeus	2744115311	35692308	2708423003	*Ostreococcus lucimarinus*	13204888	9216998	3987890
Chrysemys picta	461747357	5803852	455943505	*Ostreococcus tauri*	12456351	10138133	2318218
Cicer arietinum	347247377	28623811	318623566	*Ovis aries*	2584815894	33794145	2551021749
Ciona intestinalis	78296155	15550846	62745309	*Pan paniscus*	3151907227	33289497	3118617730
Citrus sinensis	238999708	27421823	211577885	*Pan troglodytes*	3091112213	34403316	3056708897
Cryptococcus gattii	18374760	10193549	8181211	*Papio anubis*	2724327674	34126862	2690200812
Cryptococcus neoformans	19699782	10546316	9153466	*Phaseolus vulgaris*	514820528	34393133	480427395
Cryptosporidium parvum	9102324	6820333	2281991	*Plasmodium cynomolgi*	22728335	9350600	13377735
Cucumis sativus	191859024	25366500	166492524	*Plasmodium falciparum*	23264338	12245290	11019048
Cynoglossus semilaevis	445139357	36786472	408352885	*Plasmodium knowlesi*	23462187	11118740	12343447
Danio rerio	1340430591	44231259	1296199332	*Plasmodium vivax*	22621071	10906305	11714766
Debaryomyces hansenii	12152486	9022180	3130306	*Poecilia reticulata*	696700953	38655401	658045552
Dictyostelium discoideum	33943072	20979100	12963972	*Pongo abelii*	3029491029	32110815	2997380214
Drosophila melanogaster	28557754	4239527	24318227	*Populus trichocarpa*	378545565	44068914	334476651
Drosophila pseudoobscura	50607275	9775205	40832070	*Prunus mume*	198852406	29298313	169554093
Drosophila simulans	17992287	2308887	15683400	*Rattus norvegicus*	2782012602	37109116	2744903486
Drosophila yakuba	23145337	3900892	19244445	*Saccharomyces cerevisiae*	12071326	8691722	3379604
Elaeis guineensis	657968836	27281976	630686860	*Salmo salar*	2240204991	71170520	2169034471
Equus caballus	2367053447	32994722	2334058725	*Scheffersomyces stipitis*	15441179	8587907	6853272
Eremothecium cymbalariae	9669424	6473618	3195806	*Schizosaccharomyces pombe*	12571820	7138394	5433426
Eremothecium gossypii	9095748	7007631	2088117	*Sesamum indicum*	233222381	30558151	202664230
Felis catus	2419212910	32750934	2386461976	*Setaria italica*	401296418	35711158	365585260
Ficedula albicollis	1044065291	25797019	1018268272	*Solanum pennellii*	926426464	34535865	891890599
Fragaria vesca	198117109	30904815	167212294	*Sorghum bicolor*	659229367	37431478	621797889
Gallus gallus	1021439028	28351491	993087537	*Sus scrofa*	2596639456	32005814	2564633642
Glycine max	949176042	60862926	888313116	*Taeniopygia guttata*	1021462940	23660215	997802725
Gorilla gorilla	2917687013	33405815	2884281198	*Takifugu rubripes*	281572362	28676957	252895405
Gossypium raimondii	749228090	44683789	704544301	*Tetrapisispora blattae*	14048593	8755400	5293193
Homo sapiens	3088269832	35915410	3052354422	*Tetrapisispora phaffii*	12100190	8034104	4066086
Kazachstania africana	11130140	7848851	3281289	*Thalassiosira pseudonana*	28733535	15615332	13118203

Organism	Size	CDS	non-CDS	*Organism*	Size (base)	CDS	non-CDS
Kluyveromyces lactis	10689156	7388969	3300187	*Theileria annulata*	8352520	6074113	2278407
Komagataella phaffii	9216378	7207175	2009203	*Theileria equi*	6015803	4155315	1860488
Lachancea thermotolerans	10392862	7509690	2883172	*Theileria orientalis*	8983596	6141721	2841875
Leishmania braziliensis	31238104	15200552	16037552	*Theileria parva*	4511914	3080757	1431157
Leishmania donovani	32444968	14635818	17809150	*Theobroma cacao*	330456197	34445261	296010936
Leishmania infantum	31924853	15556104	16368749	*Thielavia terrestris*	36912256	13614268	23297988
Leishmania major	32855089	15710352	17144737	*Torulaspora delbrueckii*	9220678	7248844	1971834
Leishmania panamensis	30688794	14542185	16146609	*Tribolium castaneum*	187494969	19166507	168328462
Lepisosteus oculatus	891144077	31820297	859323780	*Trypanosoma brucei*	22148088	13292454	8855634
Macaca fascicularis	2871826009	34465017	2837360992	*Ustilago maydis*	19643891	11979357	7664534
Macaca mulatta	2835963390	34674220	2801289170	*Vigna radiata*	333308464	29662798	303645666
Magnaporthe oryzae	40491973	16673311	23818662	*Vitis vinifera*	426176009	31432213	394743796
Malus domestica	526197889	36138315	490059574	*Yarrowia lipolytica*	20502981	9430576	11072405
Medicago truncatula	384466993	47725325	336741668	*Zea mays*	2059701728	44366789	2015334939
Meleagris gallopavo	972203167	25172179	947030988	*Zygosaccharomyces rouxii*	9764635	7436797	2327838
Micromonas sp.	20989326	14597320	6392006	*Zymoseptoria tritici*	39686251	14379863	25306388

3.3 Circular code motifs

3.3.1 Search algorithms

We present here the search algorithms of circular code motifs in a nucleic acid sequence. First, we will present the biinfinite motif, a special case of the circular code motif that is not restricted to complete trinucleotides from a circular code set at the start and end of the motif, i.e. it can include a nucleotide or dinucleotide (prefix or suffix) belonging to a circular code trinucleotide. Second, we will present the approach used in previous works which focused on the usage of sets to determine circular code motifs. Finally, we will present the algorithm used to search for circular code motifs in the studies done in this thesis which approached the search from frames point of view.

3.3.1.1 Biinfinite word

We consider the classical case of the window length $\tilde{l}$ for a biinfinite word $\tilde{w} = \ldots l_{-1} l_0 l_1 \ldots$ for a circular code X. We will take the example from Figure 2.4 for a biinfinite word $\tilde{w} = \ldots AGGTAATTACCAG \ldots$ of the common circular code X. The first nucleotide of $\tilde{w}$, A, is possibly the 1st, 2nd or 3rd nucleotide of trinucleotide of X. By trying the three possible factorizations (frames) $\tilde{w}_0$, $\tilde{w}_1$ and $\tilde{w}_2$ into trinucleotides of X, only one factorization, $\tilde{w}_1$, is possible ($\tilde{w}_1$ $\tilde{w}_2$ being $\tilde{w}_0$ shifted by one and two nucleotides receptively). Thus, the first nucleotide A of $\tilde{w}$ is the 3rd nucleotide of a trinucleotide of X. The factorization of $\tilde{w}_1$ leads to having the following trinucleotides NNA, GGT, AAT, TAC and CAG that belong to X (N designates any appropriate letter of X). The factorization of $\tilde{w}_0$ and $\tilde{w}_2$ are not a viable options due to the fact that no trinucleotide

of X starts with the prefix AG. This case occurs immediately for $\tilde{w}_0$ and after 11 nucleotides for $\tilde{w}_2$.

Thus, the unique factorization of $\tilde{w}$ is $\tilde{w}_1 = \ldots A, GGT, AAT, TAC, CAG, \ldots$ $\tilde{w}$ can be located anywhere in a sequence of X, i.e. the sequence of X does not require an initiator codon (or a stop codon) to retrieve the reading frame. The finite word $\tilde{w}_\alpha = AGGTAATTACCA$ ($\tilde{w}$ without the last nucleotide G), with a length of 12 nucleotides, is ambiguous as it has two factorizations $\tilde{w}_1$ and $\tilde{w}_2$ into trinucleotides of X. The word $\tilde{w}_\alpha$ is called ambiguous word of X. By definition of a circular code, all the ambiguous words are finite words. We will prove that $\tilde{w}_\alpha$, taken as an illustration example here, is one of the four longest of X. Then, the window length $\tilde{l}$ to retrieve the construction frame of any biinfinite word of a circular code Y is the letter length of the longest ambiguous word $\tilde{w}_\alpha$, plus one letter. Thus, with the common circular code X, $\tilde{l} = 12+1 = 13$ nucleotides. The window lengths $\tilde{l}$ for the trinucleotide circular codes $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$ are also equal to $\tilde{l} = 13$ nucleotides. In conclusion, the retrieval of the reading frame of the common circular code X needs a window length $\tilde{l}$ of 13 nucleotides in each frame.

3.3.1.2 Sets algorithm

A set is a collection of distinct elements without repetition and without order. It is written here with sans serif font, e.g. S. A multiset is a generalization of a set. It is an unordered collection of elements with multiple but finite occurrences of any element. It is written here with calligraphy font, e.g. $\mathcal{S}$. The multiplicity $m_{\mathcal{S}}(e)$ of an element e in a multiset $\mathcal{S}$ is its occurrence number. In our context, and for readability reason, a multiset is represented as follows, e.g. $\mathcal{S} = \{A : m_{\mathcal{S}}(A), C : m_{\mathcal{S}}(C), G : m_{\mathcal{S}}(G), T : m_{\mathcal{S}}(T)\}$.

We briefly recall the definitions of intersection and union of multisets. Let $\mathcal{S}$ and $\mathcal{T}$ be two multisets . The union $\mathcal{S} \cup \mathcal{T}$ of $\mathcal{S}$ and $\mathcal{T}$ is the multiset defined by $m_{\mathcal{S}\cup\mathcal{T}}(e) = max(m_{\mathcal{S}}(e), m_{\mathcal{T}}(e))$, i.e. the multiplicity of an element in $\mathcal{S} \cup \mathcal{T}$ is equal to the maximum of the multiplicities of the element in $\mathcal{S}$ and $\mathcal{T}$. For example, if $\mathcal{S} = \{A : 3, G : 1, T : 2\}$ and $\mathcal{T} = \{A : 2, C : 1, G : 2\}$ then $\mathcal{S} \cup \mathcal{T} = \{A : 3, C : 1, G : 2, T : 2\}$. The intersection $\mathcal{S} \cap \mathcal{T}$ of $\mathcal{S}$ and $\mathcal{T}$ is the multiset defined by $m_{\mathcal{S}\cap\mathcal{T}}(e) = min(m_{\mathcal{S}}(e), m_{\mathcal{T}}(e))$, i.e. the multiplicity of an element in $\mathcal{S} \cap \mathcal{T}$ is equal to the minimum of the multiplicities of the element in $\mathcal{S}$ and $\mathcal{T}$. Using the previous example, $\mathcal{S} \cap \mathcal{T} = \{A : 2, G : 1\}$. Finally, a subset S of a multiset $\mathcal{S}$ is called the support of $\mathcal{S}$ if for every element e such that

Algorithm 3.1 The algorithm AmbiguousWordsX gives the ambiguous words of the common circular code X in the shifted frames (f) 1 and 2 with a length varying from 1 to 11 nucleotides. Remember that there is no ambiguous word of X with a length $l > 12$ nucleotides.

```
Algorithm_AmbiguousWordsX
// Ambiguous words of X in frames 1 and 2
1   for l ← 1 to 11 step +1 do // Nucleotide length l

        // Set W^l of words of X at length l
2       W^l ← wordsX(l)

        // Multiset W_1^l of words of X in frame 1 at length l
3       W_1^l ←wordsXFrame1(l)
        // Multiset W_2^l of words of X in frame 2 at length l
4       W_2^l ←wordsXFrame2(l)

5       for f ← 1 to 2 step +1 do // Frame f
            // Set M_f^l of ambiguous words of X in frame f at length l
6           M_f^l ← W^l ∩ W_f^l
            // Multiset M_f^l of ambiguous words of X in frame f at length l
7           M_f^l ← M_f^l ∪ W_f^l
```

```
wordsX(l)
// Determination of the set W of words of X
1   W ← { }
2   if l = 1[3] then W ← X^(⌊l/3⌋) · S_1
3   else if l = 2[3] then W ← X^(⌊l/3⌋) · S_12
4       else W ← X^(⌊l/3⌋)
5   return W
```

```
wordsXFrame1(l)
// Determination of the multiset W of words of X in frame 1
1   W ← { }
2   if l = 1 then W ← S_2
3   else if l = 2[3] then W ← S_23 · X^(⌊l/3⌋)
4       else if l = 0[3] then W ← S_23 · X^(⌊l/3⌋−1) · S_1
5           else W ← S_23 · X^(⌊l/3⌋−1) · S_12
6   return W
```

```
wordsXFrame2(l)
// Determination of the multiset W of words of X in frame 2
1   W ← { }
2   if l = 1[3] then W ← S_3 · X^(⌊l/3⌋)
3   else if l = 2[3] then W ← S_3 · X^(⌊l/3⌋) · S_1
4       else W ← S_3 · X^(⌊l/3⌋) · S_12
5   return W
```

$m_{\mathcal{S}}(e) > 0$ this implies that $e \in \mathsf{S}$, and for every element e such that $m_{\mathcal{S}}(e) = 0$ this implies that $e \notin \mathsf{S}$. For example, the set $\mathsf{S} = \{A, G, T\}$ is the support of $\mathcal{S} = \{A : 3, G : 1, T : 2\}$. For simplification in the writing of the algorithm, the

same operators of intersection and union are used for sets and multisets. Thus, the intersection $\mathcal{S} \cap \mathcal{T}$ of a set S and a multiset $\mathcal{T}$ leads to a set (the support T of $\mathcal{T}$ replacing $\mathcal{T}$). The union $\mathcal{S} \cup \mathcal{T}$ of a set S and a multiset S_2 leads to a set (the multiset $\mathcal{S}$ of multiplicity 1 replacing S). We recall the circular code X:

$$X = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, \\ GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}.$$

Let a word of circular code set X be the three letters $l_1 l_2 l_3$. Let S_1 be the set containing the letters l_1 of X, S_2 and S_3 for l_2 and l_3 respectively. Then, $(\mathsf{S}_1 = \mathsf{S}_2 = \mathsf{S}_3 = B = \{A, C, G, T\}$.

Let $\mathcal{S}_1$ be the multiset containing letters l_1 of X, $\mathcal{S}_2$ and $\mathcal{S}_3$ for l_2 and l_3 respectively. Then,

$$\begin{aligned} \mathcal{S}_1 &= \{A:5, C:3, G:10, T:2\} \\ \mathcal{S}_2 &= \{A:8, C:2, G:2, T:8\} \\ \mathcal{S}_3 &= \{A:2, C:10, G:3, T:5\}. \end{aligned} \tag{3.1}$$

Remark 3.1. As X is self-complementary, $\mathcal{C}(\mathcal{S}_1) = \{\mathcal{C}(A) : 5, \mathcal{C}(C) : 3, \mathcal{C}(G) : 10, \mathcal{C}(T) : 2\} = \{T : 5, G : 3, C : 10, A : 2\} = \mathcal{S}_3$. Similarly, $\mathcal{C}(\mathcal{S}_3) = \{\mathcal{C}(A) : 8, \mathcal{C}(C) : 2, \mathcal{C}(G) : 2, \mathcal{C}(T) : 8\} = \{T : 8, G : 2, C : 2, A : 8\} = \mathcal{S}_2$.

Let S_{12} be the set containing the prefix $l_1 l_2$ of X, $\mathcal{S}_{12}$ the respective multiset. Then,

$$\begin{aligned} \mathsf{S}_{12} &= \{AA, AC, AT, CA, CT, GA, GC, GG, GT, TA, TT\} \\ \mathcal{S}_{12} &= \{AA:2, AC:1, AT:2, CA:1, CT:2, \\ &\qquad GA:4, GC:1, GG:2, GT:3, TA:1, TT:1\}. \end{aligned} \tag{3.2}$$

Let S_{23} be the set containing the suffix $l_2 l_3$ of X, $\mathcal{S}_{23}$ the respective multiset. Then,

$$\begin{aligned} \mathsf{S}_{23} &= \{AA, AC, AG, AT, CC, GC, GT, TA, TC, TG, TT\} \\ \mathcal{S}_{23} &= \{AA:1, AC:3, AG:2, AT:2, CC:2, \\ &\qquad GC:1, GT:1, TA:1, TC:4, TG:1, TT:2\}. \end{aligned} \tag{3.3}$$

Remark 3.2. $\mathrm{card}(\mathsf{S}_{12}) = \mathrm{card}(\mathsf{S}_{23}) = 11$ (among 16 dinucleotides). $\mathsf{S}_{12} \neq \mathsf{S}_{23}$

and $\mathtt{S}_{12} \cap \mathtt{S}_{23} = \{AA, AC, AT, GC, GT, TA, TT\}$. $\mathcal{C}(\mathtt{S}_{12}) = \mathtt{S}_{23}$, $\mathcal{C}(\mathtt{S}_{23}) = \mathtt{S}_{12}$ and $\mathcal{C}(\mathtt{S}_{12} \cap \mathtt{S}_{23}) = \mathtt{S}_{23} \cap \mathtt{S}_{23}$.

Let A and B be two sets of words. $A \cdot B$ is the set of words which are the product (concatenation) of one word of A and one word of B, i.e. $A \cdot B = \{a_i \cdot b_j \big| a_i \in A, b_j \in B\}$. Thus, $A^n = \underbrace{A \cdot A \cdot \ldots \cdot A}_{n}$ is the set of words which are the product of $n, n \leq 0$, words of A, i.e. $A^n = \{a_1 \cdot a_2 \cdot \ldots \cdot a_n \big| a_i \in A\}$ A^0 being the empty set. For example, X^2 is the set of all concatenations of two words of X, i.e. $\{AACAAC, AACAAT, \ldots, TTCTTC\}$. All these definitions on sets are naturally extended on multisets.

Algorithm 3.2 The frames algorithm can retrieve all circular code motifs from all the frames of a sequence. It is also able to retrieve motifs from other codes, such as Random codes (Section 4.3.1).

```
1.  Read sequence
2.  INIT X AS the set of circular code trinucleotides
3.  INIT minsize AS the minimum size of motifs
4.  INIT shift
5.  FOR EACH frame
6.      CASE frame OF
7.          0 : set shift to 0
8.          1 : set shift to 2
9.          2 : set shift to 1
10.     ENDCASE
11.     INIT motif AS empty
12.     FOR EACH trinucleotide in the sequence beginning from
    shift AS tri
13.         IF X contains tri THEN
14.             IF motif is empty THEN
15.                 Set motif to tri
16.             ELSE
17.                 Concatenate tri to motif
18.             ENDIF
19.         ELSEIF motif length is larger than minsize THEN
20.             Add motif to list of motifs
21.             Set motif to empty
22.         ELSE
23.             Set motif to empty
24.         ENDIF
25.     ENDFOR
26. ENDFOR
```

3.3.1.3 Frames algorithm

Let S be a sequence of nucleotides over $B = \{A, C, G, T\}$. Let $|\mathsf{S}|$ be the number of nucleotides in sequence S. Let Tri be a trinucleotide of S such that $Tri = l_{i-2}l_{i-1}l_i \in A_4^3, i \leq |\mathsf{S}|$. We can express S as a concatenation of trinucleotides.

With respect to the start of the sequence, the concatenation for frame 0 is $\mathsf{S}_0 = Tri_1 \cdot Tri_2 \cdot \ldots \cdot Tri_n, n = \frac{|\mathsf{S}|}{3}$ where Tri_1 being the first trinucleotide containing the first three nucleotides $l_1l_2l_3$ of S, Tri_2 containing the following nucleotides $l_4l_5l_6$ of S and so on and so forth. Respectively, $\mathsf{S}_1 = l_1 \cdot Tri_1 \cdot Tri_2 \cdot \ldots \cdot Tri_n, n = \frac{|\mathsf{S}|-1}{3}$ and $\mathsf{S}_2 = l_1 \cdot l_2 \cdot Tri_1 \cdot Tri_2 \cdot \ldots \cdot Tri_n, n = \frac{|\mathsf{S}|-2}{3}$ are the concatenation of trinucleotides of S in frames 1 and 2.

Let $\mathcal{S}_0$ be the multiset containing the trinucleotides Tri_i from S_0 with their order respecting their position in the sequence. The frames algorithm will iterate through this multiset matching trinucleotides to those from X, e.g. if $Tri_i \in \mathcal{S}_0$, Tri_i will be added to motif m_X (motif will start as empty), in case of biinfinite motifs then Tri_{i-1} will be examined if l_2l_3 could possibly be part of X motif (as previous algorithm), the algorithm will continue iterating through $\mathcal{S}_0$ until it reaches a trinucleotide which does not belong to X which will be inspected if its l_1l_2 belong to X trinucleotides. This process will continue over the whole of $\mathcal{S}_0$ till the last trinucleotide, at that point we have collected all possible X circular code motifs in frame 0, the same will be done for $\mathcal{S}_1$ and $\mathcal{S}_2$ for frames 1 and 2 respectively.

This approach is faster and allows the maximal retrieval of X circular code motifs in any given sequence regardless of overlapping motifs (for lengths less than 12). The algorithm has four parameters, the sequence to be searched for the motifs, the set used to construct the motifs (which makes the algorithm generic and not restricted to X circular code), the minimum length of a motif (in nucleotides) and minimum number of unique trinucleotides from the given set that a motif should be composed of (used for the retrieval of motifs in sections 4.2 and 4.3).

3.3.1.3.1 Definition of a motif

The output of the Frames algorithm is a list of motifs found in all the frames of an examined sequence S using a set Y. Each motif m has the following set of attributes: the motif (as a sequence or displayed as trinucleotides with comma delimiter), the index of the first nucleotide in the motif with respect to the

sequence S, the index of the last nucleotide in the motif with respect to the sequence S, the length of the motif in nucleotides, the frame in which the motif was found with respect to the sequence S, the number of unique trinucleotides present in the motif m according to the set Y and finally the number of unique trinucleotides (W) present in the motif according the X_0 circular code (Equation 2.2) which will be called trinucleotide cardinality (composition).

In particular, we are interested in the length and cardinality of a motif.

$$\begin{cases} n = l(m(Y)) \\ Card(\{W_1\} \cup \{W_2\} \cup \ldots \cup \{W_n\}) = Card(\{W(m(Y))\}). \end{cases}$$

The motif class studied in this paper are called large motifs, they are defined by two conditions on their length and cardinality

$$\begin{cases} n = l(m(Y)) \geq 15 \text{ trinucleotides} \\ Card(\{W(m(Y))\}) \geq 10 \text{ trinucleotides} \end{cases} \tag{3.4}$$

These attributes, coupled with the knowledge of what organism we are searching in, allows us to conduct very specific and detailed studies along different angles of a motif. The algorithm in its ability to construct motifs from any given set, which allows to compare the results of X motifs against other sets.

3.3.2 Definition of random code motifs

The motifs $m(X)$, $m(\Pi(X))$, $m(X_1)$ and $m(X_2)$ are generated from the maximal circular codes X, $\Pi(X)$, X_1 and X_2, respectively. All these circular codes have 20 trinucleotides with the same total numbers of nucleotides, i.e. 15 A, 15 C, 15 G, 15 T. Furthermore, by definition of a circular code, they have neither a periodic trinucleotide $P^3 = \{AAA, CCC, GGG, TTT\}$ nor two non-periodic permuted trinucleotides $\{t, \mathcal{P}(t)\}$ (Remark 2.1). In order to have an evaluation of the statistical significance of occurrence numbers of the large motifs $m(X)$, $m(\Pi(X))$, $m(X_1)$ and $m(X_2)$, 30 random codes R are generated with respect to the four necessary conditions of maximal circular codes: (i) a random code R with a number of trinucleotides equal to 20; (ii) a random code R without a periodic trinucleotide P^3; (iii) a random code R without two non-periodic

permuted trinucleotide and (iv) a random code R containing the same total numbers of nucleotides (15 A, 15 C, 15 G, 15 T). Then, a random code R of trinucleotides randomly chosen in B^3 is generated satisfying the four conditions (i), (ii), (iii) and (iv).

The complete list of generated random sets are displayed below for reference for future comparison and reproducibility of results:

$$
\begin{aligned}
R_{1} &= \{AAC, AAT, ACA, ACG, ACT, AGA, CCT, CGA, CTG, GAC, GCA, GCG, GGC, GTC, TAG, TCT, TGA, TGT, TTC, TTG\}\\
R_{2} &= \{ACC, ACT, AGC, ATG, ATT, CCA, CTA, GAA, GAG, GCA, GCG, GGC, GTT, TAC, TAG, TAT, TCC, TCG, TGA, TGC\}\\
R_{3} &= \{AAG, AAT, ACA, ACT, AGT, ATG, ATT, CAG, CCT, CGG, CTT, GAC, GAT, GCA, GCG, GTC, TCC, TCG, TGA, TGC\}\\
R_{4} &= \{AAT, AGA, AGT, ATC, ATG, CAA, CAG, CCG, CGC, CGG, CTA, GAT, GCC, GCT, GTC, TAA, TCT, TGC, TGG, TTA\}\\
R_{5} &= \{AAC, ACA, ACC, AGA, AGT, CAA, CCG, CGG, CGT, GCG, GCT, GGT, GTC, GTT, TAA, TAG, TAT, TCA, TCG, TTC\}\\
R_{6} &= \{AAC, ACA, ACG, AGG, ATG, CAA, CAT, CCG, CGC, CTT, GAA, GCT, GGC, GTC, GTT, TAG, TCA, TGC, TTA, TTG\}\\
R_{7} &= \{AAC, AAT, ACA, AGG, ATA, ATG, CAC, CCA, CGC, GAT, GCC, GGA, GGT, GTC, GTT, TAG, TCC, TCT, TGG, TTC\}\\
R_{8} &= \{AAG, ACA, ACC, ATC, ATG, ATT, CAA, CCT, CGG, CTA, CTG, GAT, GCA, GGA, GGT, GTG, TCA, TCC, TGG, TTC\}\\
R_{9} &= \{AAC, ACT, AGC, AGT, ATA, ATG, CCG, CGA, CGC, CTT, GAA, GCG, GCT, TAC, TAG, TCA, TCC, TGA, TGG, TGT\}\\
R_{10} &= \{ACA, ACC, AGG, AGT, ATT, CAA, CCA, CCT, CGA, CGG, CTA, GAG, GCG, GCT, GTT, TAC, TAG, TCG, TTA, TTG\}\\
R_{11} &= \{AAC, ACG, ATA, CAA, CCT, CGC, CTA, CTG, GAC, GCA, GCT, GGA, GTG, GTT, TAA, TAC, TAG, TCT, TGC, TGG\}\\
R_{12} &= \{AAC, ACC, ACG, AGC, ATA, CAA, CAC, CGC, CTT, GAT, GCT, GGC, GGT, GTG, TAA, TCA, TGA, TGC, TGT, TTG\}\\
R_{13} &= \{ACC, AGC, ATA, ATC, ATG, CAA, CCT, CGG, CTC, GAC, GAT, GCT, GGA, GTA, GTC, GTG, TAA, TCA, TCT, TGG\}\\
R_{14} &= \{AAT, ACA, ACG, ACT, AGT, ATA, CAA, CCT, CGC, CTG, GAC, GAT, GCT, GGA, GGT, GTC, TAG, TCC, TGC, TGT\}\\
R_{15} &= \{AAC, ACA, ACT, AGT, ATC, ATG, CAC, CAG, CCG, CGC, CTT, GAA, GCG, GGT, GTG, TCA, TCG, TGA, TGT, TTA\}\\
R_{16} &= \{AAG, ACA, AGG, ATC, CAG, CAT, CCA, CGA, CGG, CGT, CTA, CTT, GCG, GTA, TAA, TAC, TCG, TGG, TTC, TTG\}\\
R_{17} &= \{AAC, AAT, ACA, ACC, AGG, CAT, CCT, CGT, GAA, GAC, GCA, GCG, GGA, GGT, GTT, TCC, TCT, TGC, TTA, TTG\}\\
R_{18} &= \{AAC, AAG, ACC, ACG, AGC, ATA, ATT, CAT, CGA, CGG, CTG, GAC, GCA, GCT, GGC, GTT, TCG, TCT, TGT, TTA\}\\
R_{19} &= \{AAC, ACC, ACT, AGG, ATA, ATC, CAG, CGA, CGT, CTC, CTT, GAG, GCT, GGA, GGC, TAC, TAT, TCT, TGA, TGG\}\\
R_{20} &= \{AAT, ACC, ACT, AGT, ATT, CAC, CCG, CGG, CGT, CTG, GAA, GAT, GCC, GGC, GTA, GTC, TAA, TAT, TCG, TGA\}\\
R_{21} &= \{ACG, CAC, CCA, CCG, CTA, CTC, GAA, GAG, GAT, GCA, GCG, GTA, GTG, TAA, TCA, TGA, TGC, TGT, TTA, TTC\}\\
R_{22} &= \{AAG, AAT, ACA, ACT, AGC, ATC, ATT, CCG, CCT, CGC, CGG, GAC, GAG, GAT, GGC, GGT, TAT, TGC, TTA, TTC\}\\
R_{23} &= \{AAC, AGT, ATG, CAA, CCA, CCT, CGA, CGG, CTA, CTC, CTG, GAT, GCA, GGA, GGT, TAG, TCA, TCT, TGG, TTA\}\\
R_{24} &= \{AAG, ACG, ATA, ATC, CAA, CAG, CAT, CCG, CGT, CTC, CTG, GAT, GGC, GGT, GTA, GTG, TAC, TCA, TCG, TTA\}\\
R_{25} &= \{AAC, ACC, ATA, ATG, CAG, CGA, CGT, CTA, CTT, GAC, GCA, GCG, GGA, GGT, GTT, TAA, TAC, TCC, TGC, TTG\}\\
R_{26} &= \{ACA, AGA, AGG, CAC, CCA, CCG, CTA, CTC, CTG, GAA, GAT, GCT, GGA, GTG, TAG, TCA, TCT, TGG, TTA, TTC\}\\
R_{27} &= \{AAT, ATG, CAA, CAC, CAG, CCA, CGG, CTA, CTT, GAC, GCC, GGA, GGC, GGT, TAG, TAT, TCT, TGA, TGC, TTA\}\\
R_{28} &= \{AAG, AAT, ACA, ACC, AGA, ATG, CAT, CTG, GAC, GCC, GCT, GGA, GGT, GTT, TAC, TCC, TCG, TGA, TGC, TTC\}\\
R_{29} &= \{AAT, ACA, ACG, ATG, ATT, CAC, CCG, CCT, CGA, CTC, GAA, GCC, GGA, GGT, GTA, GTC, GTT, TAG, TAT, TCG\}\\
R_{30} &= \{AAG, ACA, ACT, AGA, AGT, ATA, ATT, CCG, CGA, CGC, CTC, GAC, GCC, GGC, GTG, TAA, TGC, TGT, TTC, TTG\}
\end{aligned}
$$

3.3.3 Definition of the 23 bijective circular codes motifs

3.3.3.1 Bijective transformation circular codes

There are 23 bijective transformation circular codes $\Pi(X) = \{\pi_1(X), ..., \pi_{23}(X)\}$ of the maximal C^3 self-complementary trinucleotide circular code $X = \pi_0(X)$. The notation of bijective transformations used here is based on the notation of (Michel and Seligmann, 2014) which relies on (i) the transcript data identified

from the human mitochondrial genomes by (Seligmann, 2013a,b); (ii) the biological function of the polymerase. These biological observations suggest that bijective transformations of RNA transcripts using only two bases are simpler than bijective transformations of three bases which are also simpler than bijective transformations of four bases. Another notation of bijective transformations of circular codes is also proposed by Fimmel, Danielli, and Strüngmann, 2013 in a study of circular codes based on group theory.

3.3.3.1.1 Symmetric and asymmetric transformation

The 23 bijective transformation circular codes $\Pi(X)$ of X can be partitioned into nine symmetric bijective transformation circular codes $\Pi_{\mathcal{S}}(X) = \{\pi_1(X), \ldots, \pi_9(X)\}$ and 14 asymmetric bijective transformation circular codes $\Pi_{\mathcal{A}}(X) = \{\pi_{10}(X), \ldots, \pi_{23}(X)\}$ (Table 3.3). The number $N(n,p)$ of bijective transformation circular codes at p letters among n letters is equal to $N(n,p) = \frac{n!}{(n-p)!p}$.

The nine symmetric bijective transformation circular codes $\Pi_{\mathcal{S}}(X)$ can be partitioned into:

1. $N(4,2) = 6$ symmetric bijective transformation circular codes $\Pi_{\mathcal{S},2}(X)$ at 2 letters.

$$\Pi_{\mathcal{S},2}(X) = \{\pi_1(X) : (A,C), \pi_2(X) : (A,G), \pi_3(X) : (A,T), \\ \pi_4(X) : (C,G), \pi_5(X) : (C,T), \pi_6(X) : (G,T)\}$$

   where $\pi_i(X) : (l_1, l_2)$ is the ith bijective transformation in the lexicographical order of the letter $l_1 \in B$ into the letter $l_2 \in B$, $l_2 \neq l_1$, and reciprocally.

2. $\frac{N(4,2)}{2} = 3$ symmetric bijective transformation circular codes $\Pi_{\mathcal{S},2,2(X)}$ of two disjoint transformations at 2 letters.

$$\Pi_{\mathcal{S},2,2}(X) = \{\pi_7(X), \pi_8(X), \pi_9(X)\}$$

   where $\pi_i(X) : (l_1, l_2)(l_3, l_4)$ is the ith bijective transformation in the lexicographical order of the letter $l_1 \in B$ into the letter $l_2 \in B$, $l_2 \neq l_1$, and reciprocally, and of the letter $l_3 \in B$, $l_3 \neq l_2 \neq l_1$, into the letter $l_4 \in B$, $l_4 \neq l_3 \neq l_2 \neq l_1$, and reciprocally.

The 14 asymmetric bijective transformation circular codes $\Pi_{\mathcal{A}}$ can also be partitioned into:

1. $N(4,3) = 8$ symmetric bijective transformation circular codes $\Pi_{\mathcal{A},3}(X)$ at 3 letters.

$$\begin{aligned}\Pi_{\mathcal{A},3}(X) =&\{\pi_{10}(X) : (A,C,G), \pi_{11}(X) : (A,C,T), \pi_{12}(X) : (A,G,C),\\ &\pi_{13}(X) : (A,G,T), \pi_{14}(X) : (A,T,C), \pi_{15}(X) : (A,T,G),\\ &\pi_{16}(X) : (C,G,T), \pi_{17}(X) : (C,T,G)\}\end{aligned}$$

   where $\pi_i(X) : (l_1, l_2, l_3)$ is the ith bijective transformation in the lexicographical order of the letter $l_1 \in B$ into the letter $l_2 \in B$, $l_2 \neq l_1$, the letter l_2 into the letter $l_3 \in B$, $l_3 \neq l_2 \neq l_1$, and the letter l_3 into the letter l_1.

2. $N(4,4) = 6$ asymmetric bijective transformation circular codes $\Pi_{\mathcal{A},4}(X)$ at 4 letters

$$\begin{aligned}\Pi_{\mathcal{A},4}(X) =&\{\pi_{18}(X) : (A,C,G,T), \pi_{19}(X) : (A,C,T,G),\\ &\pi_{20}(X) : (A,G,C,T), \pi_{21}(X) : (A,G,T,C),\\ &\pi_{22}(X) : (A,T,C,G), \pi_{23}(X) : (A,T,G,C)\}\end{aligned}$$

   where $\pi_i(X) : (l_1, l_2, l_3, l_4)$ is the ith bijective transformation in the lexicographical order of the letter $l_1 \in B$ into the letter $l_2 \in B$, $l_2 \neq l_1$, the letter l_2 into the letter $l_3 \in B$, $l_3 \neq l_2 \neq l_1$, and the letter l_3 into the letter $l_4 \in B, l_4 \neq l_3 \neq l_2 \neq l_1$, and the letter l_4 into the letter l_1.

Note that the transformation at 1 ($X = \pi_0(X)$), 2, 3 and 4 letters are the transformations or order 1, 2, 3 and 4, respectively, according to the notation in (Fimmel, Danielli, and Strüngmann, 2013).

3.3.3.1.2 Complementary and non-complementary transformation

The 23 bijective complementary transformation circular codes $\Pi(X)$ of X can also be partitioned into seven selt-complementary bijective transformation circular codes $\Pi_{\mathcal{C}}(X) = \{\pi_3(X), \pi_4(X), \pi_7(X), \pi_8(X), \pi_9(X), \pi_19(X), \pi_21(X)\}$ and 16 non selft-complementary bijective transformation circular codes $\Pi_{\overline{\mathcal{C}}}(X) = \Pi(X) \backslash \Pi_{\mathcal{C}}(X)$ of X (Table 3.3).

Table 3.3: The maximal C^3 self-complementary trinucleotide circular code $X = \pi_0(X)$ and its 23 bijective transformation circular codes $\Pi(X) = \{\pi_1(X), \ldots, \pi_{23}(X)\}$; the six symmetric bijective transformation circular codes $\Pi_{\mathcal{S},2}(X) = \{\pi_1(X), \ldots, \pi_6(X)\}$ at 2 letters, the three symmetric bijective transformation circular codes $\Pi_{\mathcal{S},2,2}(X) = \{\pi_7(X), \pi_8(X), \pi_9(X)\}$ of two disjoint transformations at 2 letters, the eight asymmetric bijective transformation circular codes $\Pi_{\mathcal{A},3}(X) = \{\pi_{10}(X), \ldots, \pi_{17}(X)\}$ at 3 letters and the six asymmetric bijective transformation circular codes $\Pi_{\mathcal{S},4}(X) = \{\pi_{18}(X), \ldots, \pi_{23}(X)\}$ at 4 letters. The seven bijective transformations $\{\pi_3(X), \pi_4(X), \pi_7(X), \pi_8(X), \pi_9(X), \pi_{19}(X), \pi_{21}(X)\}$, in bold, are maximal C^3 self-complementary trinucleotide circular codes.

$X = \pi_0(X)$	AAC	AAT	ACC	ATC	ATT	CAG	CTC	CTG	GAA	GAC	GAG	GAT	GCC	GGC	GGT	GTA	GTC	GTT	TAC	TTC
$\pi_1(X) : (A, C)$	CCA	CCT	CAA	CTA	CTT	ACG	ATA	ATG	GCC	GCA	GCG	GCT	GAA	GGA	GGT	GTC	GTA	GTT	TCA	TTA
$\pi_2(X) : (A, G)$	GGC	GGT	GCC	GTC	GTT	CGA	CTC	CTA	AGG	AGC	AGA	AGT	ACC	AAC	AAT	ATG	ATC	ATT	TGC	TTC
$\boldsymbol{\pi_3(X) : (A, T)}$	TTC	TTA	TCC	TAC	TAA	CTG	CAC	CAG	GTT	GTC	GTG	GTA	GCC	GGC	GGA	GAT	GAC	GAA	ATC	AAC
$\boldsymbol{\pi_4(X) : (C, G)}$	AAG	AAT	AGG	ATG	ATT	GAC	GTG	GTC	CAA	CAG	CAC	CAT	CGG	CCG	CCT	CTA	CTG	CTT	TAG	TTG
$\pi_5(X) : (C, T)$	AAT	AAC	ATT	ACT	ACC	TAG	TCT	TCG	GAA	GAT	GAG	GAC	GTT	GGT	GGC	GCA	GCT	GCC	CAT	CCT
$\pi_6(X) : (G, T)$	AAC	AAG	ACC	AGC	AGG	CAT	CGC	CGT	TAA	TAC	TAT	TAG	TCC	TTC	TTG	TGA	TGC	TGG	GAC	GGC
$\boldsymbol{\pi_7(X) : (A, C)(G, T)}$	CCA	CCG	CAA	CGA	CGG	ACT	AGA	AGT	TCC	TCA	TCT	TCG	TAA	TTA	TTG	TGC	TGA	TGG	GCA	GGA
$\boldsymbol{\pi_8(X) : (A, G)(C, T)}$	GGT	GGC	GTT	GCT	GCC	TGA	TCT	TCA	AGG	AGT	AGA	AGC	ATT	AAT	AAC	ACG	ACT	ACC	CGT	CCT
$\boldsymbol{\pi_9(X) : (A, T)(C, G)}$	TTG	TTA	TGG	TAG	TAA	GTC	GAG	GAC	CTT	CTG	CTC	CTA	CGG	CCG	CCA	CAT	CAG	CAA	ATG	AAG
$\pi_{10}(X) : (A, C, G)$	CCG	CCT	CGG	CTG	CTT	GCA	GTG	GTA	ACC	ACG	ACA	ACT	AGG	AAG	AAT	ATC	ATG	ATT	TCG	TTG
$\pi_{11}(X) : (A, C, T)$	CCT	CCA	CTT	CAT	CAA	TCG	TAT	TAG	GCC	GCT	GCG	GCA	GTT	GGT	GGA	GAC	GAT	GAA	ACT	AAT
$\pi_{12}(X) : (A, G, C)$	GGA	GGT	GAA	GTA	GTT	AGC	ATA	ATC	CGG	CGA	CGC	CGT	CAA	CCA	CCT	CTG	CTA	CTT	TGA	TTA
$\pi_{13}(X) : (A, G, T)$	GGC	GGA	GCC	GAC	GAA	CGT	CAC	CAT	TGG	TGC	TGT	TGA	TCC	TTC	TTA	TAG	TAC	TAA	AGC	AAC
$\pi_{14}(X) : (A, T, C)$	TTA	TTC	TAA	TCA	TCC	ATG	ACA	ACG	GTT	GTA	GTG	GTC	GAA	GGA	GGC	GCT	GCA	GCC	CTA	CCA
$\pi_{15}(X) : (A, T, G)$	TTC	TTG	TCC	TGC	TGG	CTA	CGC	CGA	ATT	ATC	ATA	ATG	ACC	AAC	AAG	AGT	AGC	AGG	GTC	GGC
$\pi_{16}(X) : (C, G, T)$	AAG	AAC	AGG	ACG	ACC	GAT	GCG	GCT	TAA	TAG	TAT	TAC	TGG	TTG	TTC	TCA	TCG	TCC	CAG	CCG
$\pi_{17}(X) : (C, T, G)$	AAT	AAG	ATT	AGT	AGG	TAC	TGT	TGC	CAA	CAT	CAC	CAG	CTT	CCT	CCG	CGA	CGT	CGG	GAT	GGT
$\pi_{18}(X) : (A, C, G, T)$	CCG	CCA	CGG	CAG	CAA	GCT	GAG	GAT	TCC	TCG	TCT	TCA	TGG	TTG	TTA	TAC	TAG	TAA	ACG	AAG
$\boldsymbol{\pi_{19}(X) : (A, C, T, G)}$	CCT	CCG	CTT	CGT	CGG	TCA	TGT	TGA	ACC	ACT	ACA	ACG	ATT	AAT	AAG	AGC	AGT	AGG	GCT	GGT
$\pi_{20}(X) : (A, G, C, T)$	GGT	GGA	GTT	GAT	GAA	TGC	TAT	TAC	CGG	CGT	CGC	CGA	CTT	CCT	CCA	CAG	CAT	CAA	AGT	AAT
$\boldsymbol{\pi_{21}(X) : (A, G, T, C)}$	GGA	GGC	GAA	GCA	GCC	AGT	ACA	ACT	TGG	TGA	TGT	TGC	TAA	TTA	TTC	TCG	TCA	TCC	CGA	CCA
$\pi_{22}(X) : (A, T, C, G)$	TTG	TTC	TGG	TCG	TCC	GTA	GCG	GCA	ATT	ATG	ATA	ATC	AGG	AAG	AAC	ACT	ACG	ACC	CTG	CCG
$\pi_{23}(X) : (A, T, G, C)$	TTA	TTG	TAA	TGA	TGG	ATC	AGA	AGC	CTT	CTA	CTC	CTG	CAA	CCA	CCG	CGT	CGA	CGG	GTA	GGA

3.3.3.2 Main properties of the 23 bijective transformation circular codes

Proposition 3.1. The 23 bijective transformation circular codes $\Pi(X)$ of X are C^3.

Proof. By letter invariance, $\Pi(X)$ belongs to the set of the 221,328 C^3 trinucleotides circular codes or by Proposition 3 in Michel and Pirillo, 2010 or by Theorem 1 in Fimmel et al., 2014.

Proposition 3.2. The seven bijective transformation circular codes $\Pi_{\mathcal{C}}(X) = \{\pi_3(X), \pi_4(X), \pi_7(X), \pi_8(X), \pi_3(X), \pi_9(X), \pi_1 9(X), \pi_2 1(X)\}$ are C^3 self-complementary.

Proof. By letter invariance for the complementarity map $\mathcal{C}$, $\Pi_{\mathcal{C}}(X)$ belongs to the set of the 216 C^3 self-complementary trinucleotides circular codes (Arquès and Michel, 1996) or by Proposition 3 in Michel and Pirillo, 2010 or by Theorem 2 in Fimmel et al., 2014.

Proposition 3.3. The probability PrRFC (Definition 2.21 in Michel, 2014) of reading frame coding (RFC) of the 23 bijective transformation circular codes $\Pi(X)$ of X are obviously all equal to the probability $PrRFC = 81.3\%$ of X (Section 2.2.2.(vi) in Michel, 2014).

3.3.4 Statistical analysis of X circular code motifs

Several minor methods were developed to aid us interpret and compare our data to uncover significance that would be otherwise difficult to notice to naked eye.

3.3.4.1 Coverage of X motifs in tRNA

The following method was used to estimate the probability of a base belonging to a circular code motif across several sequences. This is ideally used for sequences of similar lengths, i.e. a set of tRNA or 16S/18S rRNA sequences. This approach will gather the motifs from each sequence and calculate the probability for a base in a sequence to belong to a circular code motif.

Let m be a set of X motifs, and their respective start and end positions be the set $P = \{[b_1..e_1], \ldots, [b_m..e_m]\}$ in a nucleic acid region designated $R = [a..b]$. Then the $Interval(P) = \{[min_1..max_1], \ldots, [min_n..max_n]\}$ is the union of the ranges b_1 to $e_1, \ldots, b_m$ to e_m.

Therefore, the $Coverage(X, R)$ giving the probability of sites in nucleic acid region R occupied by X motifs is:

$$Coverage(X, R) = \frac{1}{b - a + 1} \sum_{i=1}^{n} (max_i - min_i + 1). \tag{3.5}$$

3.3.4.2 Expectation of the occurrence number of a motif

The expectation $E\big[N\big(m_{chr}(X)\big)\big]$ of the occurrence number $N\big(m_{chr}(X)\big)$ of an X motif m_X in chromosome chr of a genome can easily be calculated with Bernoulli model:

$$E\big[N\big(m_{chr}(X)\big)\big] = (N_{chr} - 3l + 1) \left(\frac{20}{64}\right)^{l}. \tag{3.6}$$

where N_{chr} is the total number of bases (size) of a chromosome chr, $l = l(m_{chr}(X))$ is the length of the motif in trinucleotides and the term $\frac{20}{64}$ is the probability of occurrence of a trinucleotide X (we recall X is maximal, thus it has 20 trinucleotides). Given that any X motif of length greater than four trinucleotides cannot overlap by definition of circular code.

3.3.4.3 Ratio of X motifs in coding and non-coding regions

The statistical analysis of X motifs $m(X)$ in a genome is based on two simple ratios: a base ratio of coding/non-coding for characterizing the base proportion of coding regions in a genome and a base ratio of X motifs in coding/non-coding for analysing the base proportion of X motifs $m(X)$ in coding and non-coding regions of a genome.

The base ratio $r(\mathcal{G})$ of coding/non-coding in a genome $\mathcal{G}$ is defined as follows:

$$r(\mathcal{G}) = \frac{N(\mathcal{G}_C)}{N(\mathcal{G}_{\overline{C}})} \tag{3.7}$$

where $N(\mathcal{G}_C)$ is the total base number in coding regions in a given genome $\mathcal{G}$ and $N(\mathcal{G}_{\overline{C}})$ is the total base number in non-coding regions in $\mathcal{G}$.

Remark 3.3. $N(\mathcal{G}_C) + N(\mathcal{G}_{\overline{C}}) = N(\mathcal{G})$ where $N(\mathcal{G})$ is the total base number (size) of a genome $\mathcal{G}$.

Remark 3.4. When $r(\mathcal{G}) < 1$, the total base number $N(\mathcal{G})$ of all coding regions

$\mathcal{G}_C$ in a genome $\mathcal{G}$ is less than the total base number $N(\mathcal{G}_{\overline{C}})$ of all non-coding regions $\mathcal{G}_{\overline{C}}$ in $\mathcal{G}$, and conversly when $r(\mathcal{G}) > 1$.

Example 3.1. With the genome $\mathcal{G}$ = *Anolis carolinensis*, $N(\mathcal{G}_C) = 16670366$ and $N(\mathcal{G}_{\overline{C}}) = 1064974225$ (see Appendix A), then $r(\mathcal{G}) = 1.6\%$.

In order to study a greater variety of X motifs, i.e. not necessarily large, the two constraints on length and cardinality (composition) defined in Equation 3.4 are relaxed. Thus, the X motifs studied in this section are based on the following conditions

$$\begin{cases} n = l(m(Y)) \geq 10 \text{ trinucleotides} \\ Card(\{W(m(Y))\}) \geq 5 \text{ trinucleotides.} \end{cases} \tag{3.8}$$

The base ratio $r_{m(X)}(\mathcal{G})$ of X motifs in coding/non-coding regions in a genome $\mathcal{G}$ is defined as follows:

$$r_{m(X)}(\mathcal{G}) = \frac{P(m_{\mathcal{G}_C}(X))}{P(m_{\mathcal{G}_{\overline{C}}}(X))} \tag{3.9}$$

where the probability $P(m_{\mathcal{G}_C}(X)) = \frac{N(m_{\mathcal{G}_C}(X))}{N(\mathcal{G}_C)}$ is the total base number $N(m_{\mathcal{G}_C}(X))$ of X motifs in the coding regions of a genome $\mathcal{G}$ divided by the total base number $N(\mathcal{G}_C)$ of coding regions in $m\mathcal{G}$ (see Equation 3.7), and the probability $P(m_{\mathcal{G}_{\overline{C}}}(X)) = \frac{N(m_{\mathcal{G}_{\overline{C}}}(X))}{N(\mathcal{G}_{\overline{C}})}$ is the total base number $N(m_{\mathcal{G}_{\overline{C}}}(X))$ of X motifs in the non-coding regions of $\mathcal{G}$ divided by the total base number $N(\mathcal{G}_{\overline{C}})$ of non-coding regions in $\mathcal{G}$ (see Equation 3.7).

Remark 3.5. A ratio $r_{m(X)}(\mathcal{G}) = 1$ means that the proportion of X motifs in coding and non coding regions is identical in the genome. A ratio $r_{m(X)}(\mathcal{G}) < 1$ means that there is preferential occurrence of X motifs in non-coding regions of $\mathcal{G}$. Conversely, a $r_{m(X)}(\mathcal{G}) > 1$ means that there is a preferential occurrence of X motifs in coding regions of $\mathcal{G}$.

The following algorithm computes the numbers $N(m_{\mathcal{G}_C}(X))$ and $N(m_{\mathcal{G}_{\overline{C}}}(X))$ of X motifs in coding and non-coding regions, respectively, of a genome $\mathcal{G}$. The sequence from genome $\mathcal{G}$ is laid with two markers, the first marker labels the nucleotides that belong to an X motif and the second marker labels the nucleotides that are in a coding region. The number $N(m_{\mathcal{G}_{\overline{C}}}(X))$ of X motifs in non-coding regions is the number of nucleotides that have only the first marker. Note that

if an X motif overlaps a coding and non-coding region, its nucleotides are split accordingly.

3.4 Ribosome study tools

3.4.1 Multiple sequence alignment

The rRNA sequences from the collected PDB entries were aligned using a multiple sequence alignment (MSA) software, CLUSTAL X 2.0.12 (Chenna et al., 2003; Higgins, Thompson, and Gibson, 1996; Jeanmougin et al., 1998; Larkin et al., 2007; Thompson et al., 1997), to help use better visualize similarities between the sequences of the various organisms at our disposal. The alignments were done on four groups. Group one was the bacterial group, which aligned the rRNA of *E. coli* and *T. thermophilus*. The second group was prokaryotic which included the rRNA of *E. coli*, *T. thermophilus* and *P. furiosus*. Third group was the eukaryotic, which included *S. cerevisiae*, *T. aestivum* and *H. sapiens* (*S. aleracea*, chloroplast, was left out due to its size difference with other eukaryotic rRNAs).

```
3J5T:A          CGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCAAATGAATTGACGGG
3I8G:A          CGTTAAGCGCGCCGCCTGGGGAGTACGGCCGCAAGGCTGAAACTCAAAGGAATTGACGGG
3J20:2          CGTTAAGCCCGCCGCCTGGGGAGTACGGCCGCAAGGCTGAAACTTAAAGGAATTGGCGGG
                ******    ************************ * ***** *** ****** ****
```

Figure 3.1: The output of a multiple sequence alignment using Clustal X, where the conventional asterisk (*) designates a conserved nucleotide.

The normal output of a MSA (Figure 3.1) was modified to highlight X circular code motifs found in the sequence with an alternation between green and red for each motif, e.g. in Figure 3.2 (b) we see the switch of colors between two consecutive motifs. The color blue (Figure 3.1 (a)) was used to highlight shared nucleotides between two overlapping motifs (due to length being less than 12 or different frames). Finally, the asterisk (*) in the modified output will designated a nucleotide base that belongs to a circular code motif in all the aligned sequences.

3.4.2 Molecule viewer

The reason behind conducting a study on PDB entries is the ability to visualize the results, which is made possible using a molecule viewer. Jmol is an open-source Java viewer (Hanson, 2010; Herráez, 2006; Willighagen, 2001) for chemical structures in 3D (`www.jmol.org`). It allows the reading of a variety of file formats

```
3J20:2  AATCCTATAATCCCGGGGGGACCGCCAGTGGCGAAGGCGCCCGGCTGGAACGGGTCCGAC  708
3J5T:A  AATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACGAAGACTGAC  754
3I8G:A  AATGCGCAGATACCGGGAGGAACGCCGATGGCGAAGGCAGCCACCTGGTCCACCCGTGAC  734
                                    *********
```

(a)

```
3J20:2  CGTTAAGCCCGCCGCCTGGGGAGTACGGCCGCAAGGCTGAAACTTAAAGGAATTGGCGGG  885
3J5T:A  CGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCAAATGAATTGACGGG  927
3I8G:A  CGTTAAGCGCGCCGCCTGGGGAGTACGGCCGCAAGGCTGAAACTCAAAGGAATTGACGGG  901
                                ********
```

(b)

Figure 3.2: The modified output of the multiple sequence alignment. The asterisk (*) designates a nucleotide belonging to X circular code motif in all sequences. Coloured nucleotides are X circular code motifs.

and high-performance 3D rendering with no hardware requirements. This allows us to examine the spatial location a motif occupies and its proximity to important or interesting regions of the ribosome.

Our application allows us to generate Jmol scripts for any PDB entry while highlighting the X circular code motifs found. Our scripts will hide protein sequences since they do not interest us in this study. It will give different structural shapes and colors to the rRNA, tRNA and mRNA sequences to better differentiate them.

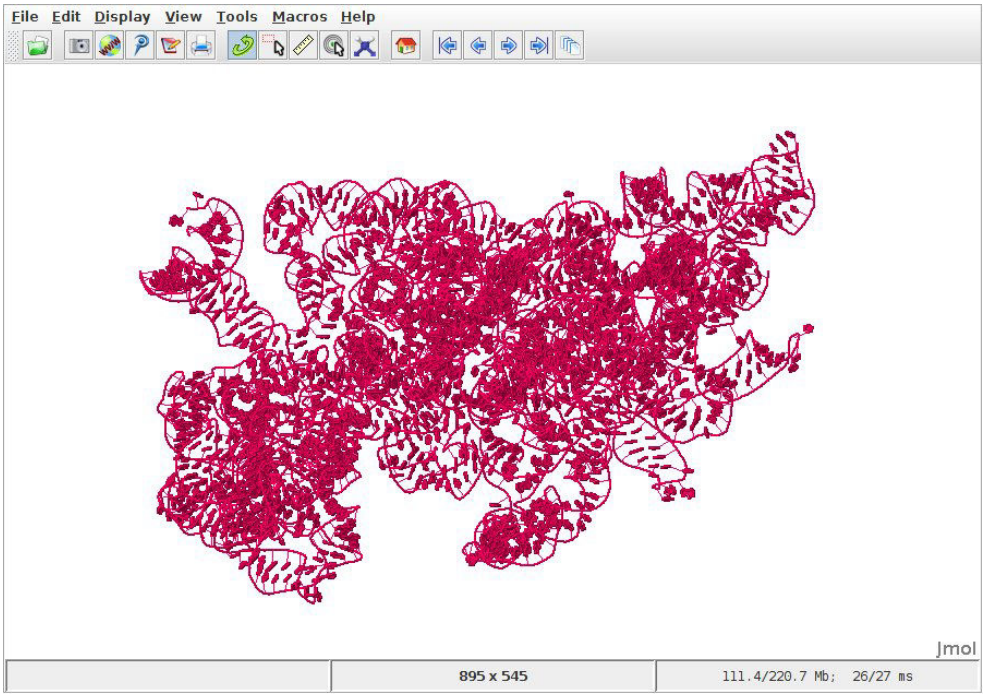

Figure 3.3: The window of Jmol with an unscripted and unformatted PDB entry of the *Homo sapiens* ribosome on display (PDB ID: 3J3D).

3.5 Repeated motifs

The unitary circular code motifs (*UCC* motifs or repeated motifs) are generated from the unitary circular codes (*UCC* defined in 2.6). They are defined by two parameters: their equivalence classes and their length in nucleotides.

3.5.1 Dinucleotide unitary circular code motifs

A repeated dinucleotide $di^+ = (l_1l_2)^+$ with $l_1l_2 \in B$ and $l_1 \neq l_2$ belongs to one of the $\frac{16-4}{2} = 6$ equivalence classes $\{(l_1l_2)^+, (l_2l_1)^+\}$ (Definition 2.13): $\{(AC)^+, (CA)^+\}$, $\{(AG)^+, (GA)^+\}$, $\{(AT)^+, (TA)^+\}$, $\{(CG)^+, (GC)^+\}$, $\{(CT)^+, (TC)^+\}$ and $\{(GT)^+, (TG)^+\}$. By convention, the dinucleotide UCC motifs are defined by the six repeated dinucleotide which are the 1st repeated motif in lexicographical order in each equivalence class:

$$Di = \{(AC)^+, (AG)^+, (AT)^+, (CG)^+, (CT)^+, (GT)^+\}. \tag{3.10}$$

The repeated dinucleotides di^n studied have length $l = n \times |di| \geq 30$ ($|di|$ being the number of letters of di), i.e. $n \geq 15$.

3.5.2 Trinucleotide unitary circular code motifs

A repeated trinucleotide $tri^+ = (l_1l_2l_3)$ with $l_1l_2l_3 \in B$ and $l_1l_2 \neq l_2l_3$ belongs to one of the $\frac{64-4}{3} = 20$ equivalence classes $\{(l_1l_2l_3)^+, (l_2l_3l_1)^+, (l_3l_1l_2)^+\}$ (Definition 2.13): $\{(AAC)^+, (ACA)^+, (CAA)^+\}, \ldots, \{(GTT)^+, (TTG)^+, (TGT)^+\}$. By convention and similarly as Di motifs, the trinucleotide UCC motifs, Tri, are defined by:

$$\begin{aligned} Tri^+ = \{&(AAC)^+, (AAG)^+, (AAT)^+, (ACC)^+, (ACG)^+, (ACT)^+, (AGC)^+, \\ &(AGG)^+, (AGT)^+, (ATC)^+, (ATG)^+, (ATT)^+, (CCG)^+, (CCT)^+, \\ &(CGG)^+, (CGT)^+, (CTG)^+, (CTT)^+, (GGT)^+, (GTT)^+\}. \end{aligned} \tag{3.11}$$

The repeated trinucleotides tri^n studied have length $l = n \times |tri| \geq 30$ nucleotides, i.e. $n \geq 10$.

3.5.3 Tetranucleotide unitary circular code motifs

A repeated tetranucleotide $tetra^+ = (l_1l_2l_3l_4)$ with $l_1, l_2, l_3, l_4 \in B$ and $l_1l_2 \neq l_3l_4$ belongs to one of the $\frac{256-16}{4} = 60$ equivalence classes $\{(l_1l_2l_3l_4)^+, (l_2l_3l_4l_1)^+, (l_3l_4l_1l_2)^+, (l_4l_1l_2l_3)^+\}$ (Definition 2.13). By convention and similarly as *Di* motifs, the tetranucleotide *UCC* motifs, *Tetra*, are defined by:

$$Tetra = \{(AAAC)^+, \ldots, (GTTT)^+\} \tag{3.12}$$

The repeated tetranucleotides $tetra^n$ studied have length $l = n \times |tetra| \geq 28$ nucleotides, i.e. $n \geq 7$.

3.5.4 Statistical analysis of repeated motifs

3.5.4.1 Occurrence number of unitary circular code motifs

Let $r^n \in \{di^n, tri^n, tetra^n\}$ be a repeated motif r of nucleotide length $n \times |r|$ (with $|r| \in \{2, 3, 4\}$ being the number of letters in r) where $r = di$ for repeated dinucleotide di^n, $r = tri$ for a repeated trinucleotide tri^n and $r = tetra$ for a repeated tetranucleotide $tetra^n$. The number $N(r^n, \mathcal{G})$ counts the occurrences of a repeated motif r^n for a given number n in a eukaryotic genome $\mathcal{G}$. Then, the occurrence number $N(r^+)$ or a repeated motif $r^+ \in \{di^+, tri^+, tetra^+\}$ in the genomes of eukaryotes $\mathbb{E}$ is obtained by summing for all genomes $\mathcal{G}$ in $\mathbb{E}$ and for all n

$$N(r^+) = \sum_{\mathcal{G} \in \mathbb{E}} \sum_{n} N(r^n, \mathcal{G}) \tag{3.13}$$

With n varying between the different UCC, being $n \geq 15$ for computing $N(di^+)$ of a repeated dinucleotide di^+, $n \geq 10$ for computing $N(tri^+)$ of a repeated trinucleotide tri^+ and $n \geq 7$ for computing $N(tetra^+)$ of a repeated tetranucleotide $tetra^+$. These occurrence numbers $N(di^+)$, $N(tri^+)$ and $N(tetra^+)$ are computed in the genomes of eukaryotes using the following algorithm.

The algorithm searches for repeated motifs in a DNA sequence such that their lengths are greater than or equal to the parameter minsize and returns a frequency map for the association of a word and how many times it was successively repeated. The algorithm is generic in regards to the input set as a parameter. It determines the number of frames required with respect to word

Algorithm 3.3 The algorithm RepeatsFinder gives the occurrence numbers of repeats in any sequence.

```
1.  Read sequence
2.  INIT Y AS a set of words
3.  INIT minsize AS the minimum number of words in a repeats motif
4.  INIT wordsize from Y
5.  INIT mapFreq AS a map using the association of a word and number
      of repeats as key with their frequency as value
6.  FOR EACH frame in wordsize
7.    INIT wordCurrent AS empty
8.    INIT streak AS 0, number of successive wordCurrent
9.    FOR EACH word in sequence starting from frame AS wordSeq
10.     IF Y contains wordSeq THEN
11.       IF wordCurrent equals wordSeq THEN
12.                increment streak by 1
13.       ELSE
14.         IF streak is greater than or equal to minsize THEN
15.           increment the frequency of wordCurrent with size equal
                to streak in mapFreq by 1
16.         ENDIF
17.         INIT wordCurrent As wordSeq
18.         INIT streak AS 1
19.       ENDIF
20.     ELSE
21.       IF streak is greater than or equal to minsize THEN
22.         increment the frequency of wordCurrent with streak in
              mapFreq by 1
23.       ENDIF
24.       INIT wordCurrent As empty
25.       INIT streak AS 0
26.     ENDIF
27.   ENDFOR
28. ENDFOR
```

length (two frames for dinucleotides, three frames for trinucleotides and four frames for tetranucleotides). This approach allows us to retrieve all the repeated motifs without the issue of overlaps between different frames because of the nature of unitary circular code.

Example 3.2. If the trinucleotide $tri = AAC$ occurs with two repeats tri^{n_1} with $n_1 = 10$ in a genome $\mathcal{G}_1$, i.e. $N(tri^{10}, \mathcal{G}_1) = 2$, and three repeats tri^{n_2} with $n_2 = 20$ in a genome $\mathcal{G}_2$, i.e. $N(tri^{20}, \mathcal{G}_2) = 3$, the occurrence number $N(tri^{+})$ of the repeated trinucleotide $(AAC)^{+}$ in the genomes of eukaryotes $\mathbb{E}$ is equal to $N(tri^{+}) = N(t^{10}, \mathcal{G}_1) + N(tri^{20}, \mathcal{G}_2) = 2 + 3 = 5$ (Equation 3.13).

3.5.4.2 Base number of unitary circular code motifs

The base number $B(r^+)$ of a repeated motif $r^+ \in \{di^+, tri^+, tetra^+\}$ in the genomes of eukaryotes $\mathbb{E}$ is:

$$B(r^+) = |r| \sum_{\mathcal{G} \in \mathbb{E}} \sum_{n} N(r^n, \mathcal{G}) \times n \tag{3.14}$$

where $N(r^n, \mathcal{G})$ is defined in Section 3.5.4.1 and with $n \geq 15$ for computing $B(di^+)$ of a repeated dinucleotide di^+, $n \geq 10$ for computing $B(tri^+)$ of a repeated trinucleotide tri^+ and $n \geq 7$ for computing $B(tetra^+)$ of a repeated tetranucleotide $tetra^+$, $|r| \in \{2, 3, 4\}$ being the number of letters of r.

Example 3.3. If the trinucleotide $tri = AAC$ occurs with two repeats tri^{n_1} with $n_1 = 10$ in a genome $\mathcal{G}_1$, i.e. $N(r^{n_1}, \mathcal{G}_1) = 2$, and three repeats t^{n_2} with $n_2 = 20$ in a genome $\mathcal{G}_2$, i.e. $N(r^{n_2}, \mathcal{G}_2) = 3$, then the base number $B(tri^+)$ of the repeated trinucleotide AAC^+ in the genomes of eukaryotes $\mathbb{E}$ is equal to $B(tri^+) = |tri|(N(r^{n_1}, \mathcal{G}_1) \times n_1 + N(r^{n_2}, \mathcal{G}_2) \times n_2) = 3(2 \times 10 + 3 \times 20) = 240$ (Equation 3.14).

3.5.4.3 Total base number of unitary circular code motifs

The total base number $B(R^+, \mathcal{G})$ of all repeated motifs $R^+ \in \{Di^+, Tri^+, Tetra^+\}$ (Equations 3.10, 3.11 and 3.12)in a genome is:

$$B(R^+, \mathcal{G}) = |r| \sum_{r^+ \in R^+} \sum_{n} N(r^n, \mathcal{G}) \times n \tag{3.15}$$

where $N(r^n, \mathcal{G})$ is defined in Section 3.5.4.1 and with $n \geq 15$ for computing $B(Di^+, \mathcal{G})$ of all repeated dinucleotide $di^+ \in Di^+$ (Equation 3.10), $n \geq 10$ for computing $B(Rri^+, \mathcal{G})$ of all repeated trinucleotide $tri^+ \in Tri^+$ (Equation 3.11) and $n \geq 7$ for computing $B(Tetra^+, \mathcal{G})$ of all repeated tetranucleotide $tetra^+ \in Tetra^+$ (Equation 3.12), $|r| \in \{2, 3, 4\}$ being the number of letters of r.

Example 3.4. If the trinucleotide $tri_1 = AAC$ occurs with two repeats $tri_1^{n_1}$ with $n_1 = 10$ in a genome $\mathcal{G}$, i.e. $N((AAC)^{n_1}, \mathcal{G}) = 2$, and three repeats $t_1^{n_2}$ with $n_2 = 20$ in a genome $\mathcal{G}_2$, i.e. $N((AAC)^{n_2}, \mathcal{G}) = 3$, and if the trinucleotide $tri_2 = AAG$ occurs with four repeats $tri_2^{n_3}$ with $n_3 = 30$ in the same genome $\mathcal{G}$, i.e. $N((AAG)^{n_3}, \mathcal{G}) = 3$, then the total base number $B(Tri^+, \mathcal{G})$ of repeated motifs Tri^+ in the genome $\mathcal{G}$ is equal to $B(Tri^+, \mathcal{G}) = |tri|(N((AAC)^{n_1}, \mathcal{G}) \times n_1 +$

$N((AAC)^{n_2}, \mathcal{G}) \times n_2 + N((AAG)^{n_3}, \mathcal{G}) \times n_3) = 3(2 \times 10 + 3 \times 20 + 4 \times 30) = 600$ (Equation 3.15).

In order to normalize the total number $B(R^+, \mathcal{G})$ (Equation 3.15) for eukaryotic genomes of different sizes, the ratio $r(R^+, \mathcal{G})$ gives the proportion of the total base number $B(R^+, \mathcal{G})$ of all repeated motifs R^+ in a eukaryotic genomes $\mathcal{G}$ of size $N(\mathcal{G})$ (Table 3.2) is defined by

$$r(R^+, \mathcal{G}) = \frac{B(R^+, \mathcal{G})}{N(\mathcal{G})}. \tag{3.16}$$

Finally, $\overline{r}(R^+)$ is the mean of the ratios $r(R^+, \mathcal{G})$ in the genomes of eukaryotes $\mathbb{E}$

$$\overline{r}(R^+) = \frac{1}{|\mathbb{E}|} \prod_{\mathcal{G} \in \mathbb{E}} r(R^+, \mathcal{G}) \tag{3.17}$$

where $|\mathbb{E}|$ is the number of genomes in $\mathbb{E}$ and $\tilde{r}(R^+)$ is the median of the ratios $r(R^+, \mathcal{G})$ in the genomes of eukaryotes $\mathbb{E}$.

3.6 Occurrence number of trinucleotide pairs

In order to identify a new property of the circular code X, we study the occurrence of the two consecutive trinucleotides $t_1t_2 \in B^6$, called trinucleotide pairs, where $t_1, t_2 \in B^3$ ($|B^6| = 4096$ t_1t_2 motifs) in the eukaryotic gene sequences. The trinucleotide pairs $t_1t_2 \in X^2$ where $t_1, t_2 \in X$ ($|X|^2 = 400$ t_1t_2 motifs) are associated to the circular X.

The number $N(t_1t_2, \mathcal{G})$ counts the occurrences of a trinucleotide pair $t_1t_2 \in B^6$ in (all) the gene sequences of a eukaryotic genome $\mathcal{G}$. Note that the number $N(t_1t_2, \mathcal{G}) = N(t^n, \mathcal{G})$ is a repeated trinucleotide t^n where $n = 2$ and of nucleotide length $n \times |t| = 2 = 6$ (Section 3.5.4.1). Then, the occurrence number $N(t_1t_2)$ of a trinucleotide pair $t_1t_2 \in B^6$ in the gene sequences of eukaryotes $\mathbb{E}$ is

$$N(t_1t_2) = \sum_{\mathcal{G} \in \mathbb{E}} N(t_1t_2, \mathcal{G}). \tag{3.18}$$

The observed probability $P(t_1t_2)$ of a trinucleotide pair $t_1t_2 \in B^6$ in the gene sequences of $\mathbb{E}$ is

$$P(t_1t_2) = \frac{N(t_1t_2)}{\sum_{t_1t_2 \in B^6} N(t_1t_2)}. \tag{3.19}$$

Due to the codon usage, in particular, this probability $P(t_1t_2)$ must be normalized. The observed probability $P(t)$ of a trinucleotide $t \in B^3$ in the gene sequences of $\mathbb{E}$ is

$$P(t) = \frac{N(t)}{\sum_{t \in B^3} N(t)} \tag{3.20}$$

with $N(t) = \sum_{\mathcal{G} \in \mathbb{E}} N(t, \mathcal{G})$ where $N(t, \mathcal{G})$ (a repeated trinucleotide t^n where $n = 1$) is the occurrence number of t in the gene sequences of a eukaryotic genome $\mathcal{G}$. By taking the hypothesis of independent events then the estimated theoretical probability $\hat{P}(t_1t_2)$ of a trinucleotide pair $t_1t_2 \in B^6$ in the gene sequences of $\mathbb{E}$ is

$$\hat{P}(t_1t_2) = P(t_1) \times P(t_2). \tag{3.21}$$

Therefore, the observed/theoretical ratio $r(t_1t_2)$ of trinucleotide pair $t_1t_2 \in B^6$ in the genes of $\mathbb{E}$ is equal to

$$r(t_1t_2) = \frac{P(t_1t_2)}{\hat{P}(t_1t_2)}. \tag{3.22}$$

Two other ratios also analyse the occurrence of trinucleotide pairs in the eukaryotic gene sequences. The observed probability $P(t_1t_2, \mathcal{G})$ of a trinucleotide pair $t_1t_2 \in B^6$ in the gene sequences of a eukaryotic genome $\mathcal{G}$ is

$$P(t_1t_2, \mathcal{G}) = \frac{N(t_1t_2, \mathcal{G})}{\sum_{t_1t_2 \in B^6} N(t_1t_2, \mathcal{G})} \tag{3.23}$$

Equation 3.23 for the gene sequences of a eukaryotic genome $\mathcal{G}$ is similar to Equation 3.19 for the gene sequences of eukaryotic $\mathbb{E}$. Similarly as previously, the observed probability $P(t, \mathcal{G})$ of a trinucleotide $t \in B^3$ in the gene sequences of genome $\mathcal{G}$ is

$$P(t, \mathcal{G}) = \frac{N(t, \mathcal{G})}{\sum_{t \in B^3} N(t, \mathcal{G})} \tag{3.24}$$

where $N(t, \mathcal{G})$ defined in Equation 3.20 is the occurrence number of t in the gene

sequences of genome $\mathcal{G}$. By taking the hypothesis of independent events then theoretical probability $\hat{P}(t_1t_2, \mathcal{G})$ of a trinucleotide pair $t_1t_2 \in B^6$ in the gene sequences of genome $\mathcal{G}$ is:

$$\hat{P}(t_1t_2, \mathcal{G}) = P(t_1, \mathcal{G}) \times P(t_2, \mathcal{G}) \tag{3.25}$$

Then, the observed/theoretical ratio $r(t_1t_2, \mathcal{G})$ of a trinucleotide pair $t_1t_2 \in B^6$ in the gene sequences of $\mathcal{G}$ is equal to:

$$r(t_1t_2, \mathcal{G}) = \frac{P(t_1t_2, \mathcal{G})}{\hat{P}(t_1t_2, \mathcal{G})}. \tag{3.26}$$

Finally, $\overline{r}(t_1t_2)$ is the mean of the observed/theoretical ratios $r(t_1t_2, \mathcal{G})$ of a trinucleotide pair $t_1t_2 \in B^6$ in the gene sequences of eukaryotic $\mathbb{E}$

$$\overline{r}(t_1t_2) = \frac{1}{|\mathbb{E}|} \sum_{\mathcal{G} \in \mathbb{E}} r(t_1t_2, \mathcal{G}) \tag{3.27}$$

where $|\mathbb{E}|$ is the number of genomes in $\mathbb{E}$ and $\tilde{r}(t_1t_2)$ is the median of the observed/theoretical ratios $r(t_1t_2, \mathcal{G})$ of a trinucleotide pair $t_1t_2 \in B^6$ in the gene sequences of eukaryotes $\mathbb{E}$.

Remark 3.6. The three observed/theoretical ratios $r(t_1t_2)$, $\overline{r}(t_1t_2)$, and $\tilde{r}(t_1t_2)$ of a trinucleotide pair $t_1t_2 \in B^6$ have the same following statistical property. When they are greater than 1, the trinucleotide pair t_1t_2 is over-represented in the eukaryotic gene sequences, and conversely when they below 1.

The three observed/theoretical ratios $r(t_1t_2)$, $\overline{r}(t_1t_2)$, and $\tilde{r}(t_1t_2)$ of trinucleotide pairs will lead to the same statistical results with circular code X (see section 4.5).

3.7 Summary

We presented in this chapter the two types of data we have, which separated this work into several studies along the type of the data and how the data was handled. During this thesis we were able to develop a new algorithm that can search for motifs in any nucleic acid sequences, this algorithm is generic by being able to use any code, be it X circular code, bijective transformation codes or randomly generated codes.

The first study focused on the relatively smaller sequences of rRNA and tRNA found in ribosomes, it covered seven organisms belonging to bacteria, archaea and eukaryote. We used the biinfinite word in search of motifs here using the Frames algorithm with a minimum length of 9 nucleotides and no restriction on the number of unique trinucleotides (cardinal) from X. The aim of the study was to study spatial significance of the X motifs.

The second study dealt with huge amount of data, the complete chromosomes of 138 genomes, obtained from RefSeq (Table 3.2), which was coupled with the mapping of their coding region, obtained from GenBank. This study aimed to find statistical significance of the X motifs compared with random sets, bijective transformations of X and its first and second permuted codes (X_1 and X_2). The motifs here collected with a minimum length of 10 trinucleotides (30 nucleotides) and a minimum of 5 unique trinucleotides.

The third study also was conducted on the complete chromosomes of 138 genomes, but it focused on searching for simple repeats which are in fact unitary circular codes. This required a new algorithm that retrieves the occurrence of each repeat in eukaryotic genomes.

The last and fourth study examines the pairing of trinucleotides in gene sequences of the eukaryotic genomes. This was conducted by using the GenBank annotation files to retrieve all gene sequences from a chromosome and study it separately in contrast to the second study where coding regions were drawn on the chromosome.

In the following chapter we will detail the results obtained from these studies while discussing their significance.

4

Results and Discussion

4.1 Introduction 63
4.2 X circular code motifs in the ribosome 63
4.2.1 X circular code motifs in the ribosomal decoding center . . . 63
4.2.1.1 The conserved A1492 and A1493 nucleotides 63
4.2.1.2 The conserved G530 nucleotide 68
4.2.2 Spatial study of X circular code motifs near the ribosomal decoding center 73
4.2.2.1 A conserved X motif near the ribosomal decoding center 73
4.2.2.2 Conserved X motifs in the rRNA of prokaryotes . . . 74
4.2.2.3 Conserved X motifs in the rRNA of eukaryotes . . . 78
4.2.3 X circular code motifs in prokaryotic tRNAs 80
4.2.3.22 Summary of X circular code motifs in tRNA sequences 99
4.2.3.23 Coverage of X circular code motifs in prokaryotic tRNAs 100
4.3 Analysis of X circular code motifs in eukaryotic genomes 100
4.3.1 Occurrence of large randoms code motifs in eukaryotic genomes 101
4.3.2 Occurrence of large motifs from X, X_1, X_2 and the 23 bijective transformations of X in eukaryotic genomes 102
4.3.3 Largest X motifs in eukaryotic genomes 105
4.3.4 Largest X motifs in *Homo sapiens* 108

4.3.5 X motifs in coding regions versus non-coding regions in eukaryotic genomes 108
4.3.6 X motifs in coding regions versus non-coding regions in *Homo sapiens* 110
4.4 Analysis of unitary circular code motifs in eukaryotic genomes . . . 111
4.4.1 Occurrence of repeated dinucleotides in eukaryotic genomes . 112
4.4.2 Occurrence of repeated trinucleotides in eukaryotic genomes . 112
4.4.3 Occurrence of repeated tetranucleotides in eukaryotic genomes 114
4.4.4 Largest repeated motifs in eukaryotic genomes 116
4.4.5 Scarcity of repeated trinucleotides in eukaryotic genomes . . 119
4.5 Identical trinucleotide pairs of the X circular code in eukaryotic gene sequences 120
4.6 Summary 123

4.1 Introduction

This chapter will be divided into two sections along the data type and biological environment mentioned before.

The first section will show the results obtained from the spatial study of the circular code motifs found in the ribosome, whether in the important decoding center or around that region, along with a study for the tRNA sequences.

The second section presents a comparative study of the X_0 circular code against its 2nd and 3rd permutations, X_1 and X_2 respectively, the 23 bijective circular code transformations and 30 random generated codes. We will show the importance and uniqueness of the X_0 circular code.

The third section involves the study of simple repeats and their ties to unitary circular codes, and a comparison between various sizes of a repeated word, dinucleotide, trinucleotide and tetranucleotide.

The fourth section presents the result from a study on the trinucleotide pairs in gene sequences of the eukaryotic genomes and the significance of identical trinucleotide pairs.

4.2 X circular code motifs in the ribosome

The data used in this study are presented in section 3.2.1 and the algorithm used to process this data is shown in section 3.2. The tools mentioned in section 3.4 were used to further analyse the sequences and the 3D structure of the ribosome.

4.2.1 X circular code motifs in the ribosomal decoding center

The universally conserved nucleotides A1492 and A1493 have an experimentally proven biological function in the codon-anti-codon binding of tRNA as the A-site in the ribosome (Moazed and Noller, 1990; Powers and Noller, 1994; Yoshizawa, Fourmy, and Puglisi, 1999). **Unexpectedly, the A1492 and A1493 nucleotides that belong to the ribosomal decoding center were found in X circular code motifs**.

4.2.1.1 The conserved A1492 and A1493 nucleotides

We will detail the X motifs containing the universally conserved A1492 and A1493 nucleotides.

Table 4.1: Identification of X circular code motifs m_{AA} containing the universally conserved nucleotides A1492 and A1493 (in bold) in all studied rRNAs of bacteria, archaea, nuclear eukaryotes, and chloroplasts.

PDB ID	Kingdom	Organism	X motif (m_{AA})	Start	End	Length
3J5T	Bacteria	*E. coli*	G,GGT,G**AA**,GTC,GTA,AC	1487	1501	15
3I8G	Bacteria	*T. thermophilus*	G,G**AA**,GGT,GC	1490	1498	10
3J20	Archaea	*P. furiosus*	A,G**AA**,GTC,GTA,AC	1445	1456	9
3IZE	Eukaryote (nuclear)	*S. cerevisiaie*	**AA**,GTC,GTA,AC	1755	1764	10
3J5Z	Eukaryote (nuclear)	*T. aestivum*	A,G**AA**,GTC,GTA,AC	1763	1774	12
3J3D	Eukaryote (nuclear)	*H. sapiens*	**AA**,GTC,GTA,AC	1824	1833	10
3BBN	Eukaryote (chloroplast)	*S. oleracea*	GT,G**AA**,GTC,GTA,AC	1438	1450	13

4.2.1.1.1 The A1492 and A1493 nucleotides in X motifs in bacterial rRNA

In the rRNAs of both *E. coli* and *T. thermophilus*, the conserved nucleotides A1492 and A1493 occur at the 2nd and 3rd sites of the trinucleotide $GAA \in X$. In rRNA of *E. coli*, it belongs to the X motif m_{AA} (*E. coli*, 1487, 1501, 15) $= G, GGT, G\boldsymbol{AA}, GTC, GTA, AC$ of 15 nucleotide length starting with the nucleotide G suffix of $CAG, CTG, GAG \in X$ followed by four trinucleotides $GAA, GGT, GTA, GTC \in X$ (given in lexicographical order) and ending with the dinucleotide AC prefix of $ACC \in X$. In rRNA of *T. thermophilus*, it belongs to the X motif m_{AA} (*T. thermophilus*, 1490, 1498, 9) $= G, G\boldsymbol{AA}, GGT, GC$ of nine nucleotide length starting with the nucleotide G, as in *E. coli*, followed by two trinucleotides $GAA, GGT \in X$ and ending with the dinucleotide GC prefix of $GCC \in X$.

These two rRNA X motifs m_{AA}(*E. coli*) and m_{AA}(*T. thermophilus*) have completely different primary structures. Thus, the classical bioinformatics methods, such as sequence alignment or phylogenetic inference, are not able to identify these motifs which were only revealed by the circular code theory.

In the rRNA of *T. thermophilus*, the following X motif which we will call m^*_{AA} (*T. thermophilus*,1461,1475,15) $= G, GGC, GAA, GTC, GTA, AC$ of 15 nucleotide length is aligned with the X motif m_{AA} (*E. coli*) with only one nucleotide difference (T in GGT replaced by C in *T. thermophilus*). However, the X motif m^*_{AA}(*T. thermophilus*) has a spatial structure far from the decoding center and probably has no function in modern rRNA of *T. thermophilus*.

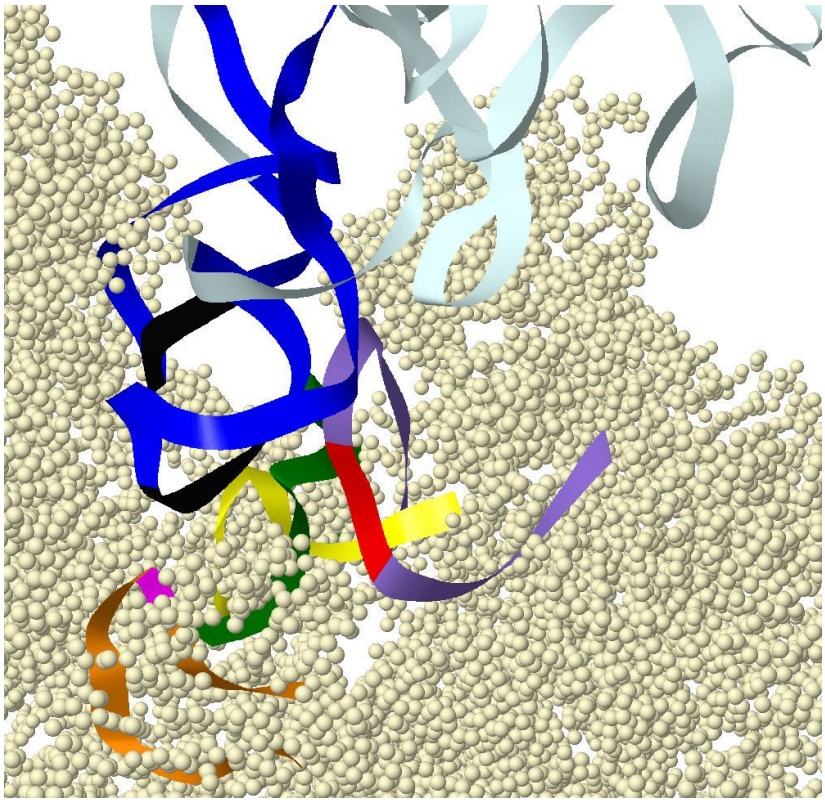

Figure 4.1: X circular code motifs involved in the bacterial ribosome decoding center of *Escherichia coli* (crystallographic structure PDB 3J5T): the mRNA X motifs (green), the rRNA X motif m_{AA}(*E. coli*, 1487, 1501, 15) (purple with the conserved A1492 and A1493 nucleotides in red), the rRNA X motif m_G(*E. coli*, 527, 536, 10) (orange with the conserved nucleotide G530 in fuchsia), the rRNA X motif m(*E. coli*, 1396, 1404, 9) (yellow) and the tRNA X motifs (blue with the anti-codon in black). The remaining rRNA (lemonchiffon) is outside the neighborhood of these X motifs.

4.2.1.1.2 The A1492 and A1493 nucleotides in X motifs in archaea rRNA

In the rRNA of *P. furiosus*, the conserved A1429 and A1493 nucleotides occur at the 2nd and 3rd sites of the trinucleotide $GAA \in X$ and belongs to the X motif m_{AA}(*P. furiosus*, 1445, 1456, 12) $= A, G\boldsymbol{AA}, GTC, GTA, AC$ of 12 nucleotide length starting with the nucleotide A suffix of $GAA, GTA \in X$ and then with a suffix of 11 nucleotides identical to the X motif m_{AA}(*E. coli*).

4.2.1.1.3 The A1492 and A1493 nucleotides in X motifs in nuclear eukaryotic rRNA

There are significant differences between prokaryotic and eukaryotic rRNAs, in particular eukaryotic 18S rRNAs are about 40% larger than the prokaryotic 16S rRNAs. Nevertheless, some rRNA sites are conserved. In particular, the universally conserved A1429 and A1493 nucleotides of bacterial rRNAs occur in

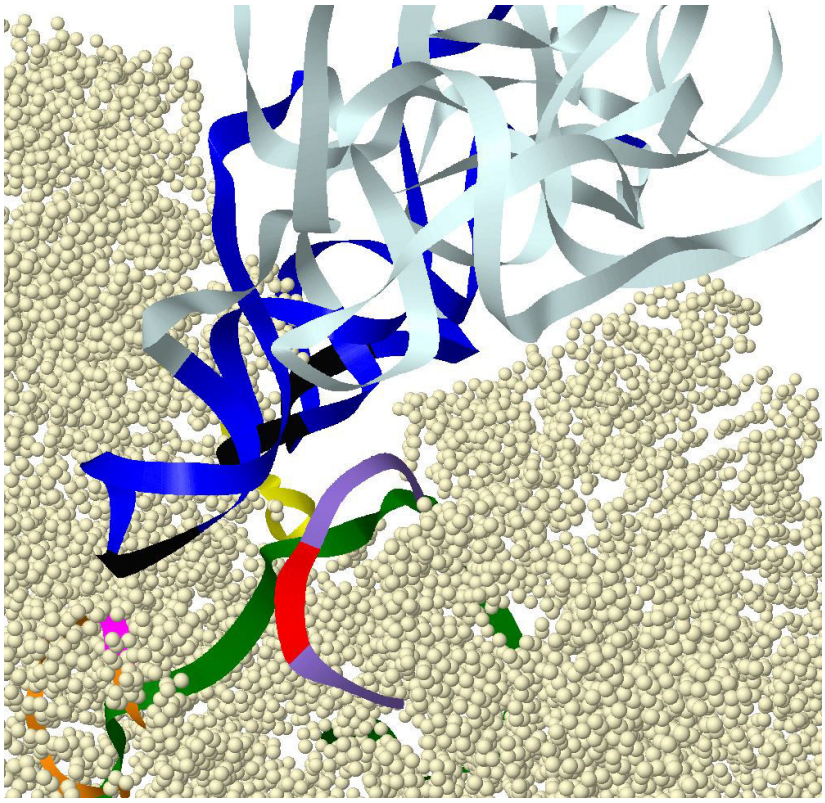

Figure 4.2: X circular code motifs involved in the bacterial ribosome decoding center of *Thermus thermophilus* (crystallographic structure PDB 3I8G): the mRNA X motifs (green), the rRNA X motif m_{AA}(*T. thermophilus*, 1490, 1498, 9) (purple with the conserved A1492 and A1493 nucleotides in red), the rRNA X motif m_G(*T. thermophilus*, 528, 536, 9) (orange with the conserved nucleotide G530 in fuchsia), the rRNA X motif m(*T. thermophilus*, 1375, 1383, 9) (yellow) and the tRNA X motifs (blue with the anti-codon in black). The remaining rRNA (lemonchiffon) is outside the neighbourhood of these X motifs.

eukaryotic rRNAs but at different positions: 1755 and 1756 in *S. cerevisiae*, 1765 and 1766 in T. aestivum (Fan-Minogue and Bedwell, 2007) and 1824 and 1825 in *H. sapiens* (Bulygin et al., 2009).

In the rRNA of *S. cerevisiae* and *H. sapiens*, the conserved A1492 and A1493 nucleotides are the prefix of the X motif m_{AA}(*S. cerevisiae*, 1755, 1764, 10) = m_{AA}(*H. sapiens*, 1824, 1833, 10) = $\boldsymbol{AA}, GTC, GTA, AC$ of 10 nucleotides length followed by two trinucleotides $GTA, GTC \in X$ and ending with the dinucleotide AC prefix of $ACC \in X$. In rRNA of *T. aestivum*, they occur occur at the 2nd and 3rd sites of the trinucleotide $GAA \in X$ and belong to the X motif m_{AA}(*T. aestivum*, 1763, 1774, 12) = $A, G\boldsymbol{AA}, GTC, GTA, AC$ of 12 nucleotides length which is identical to the archaeal X motif m_{AA}(*P. furiosus*).

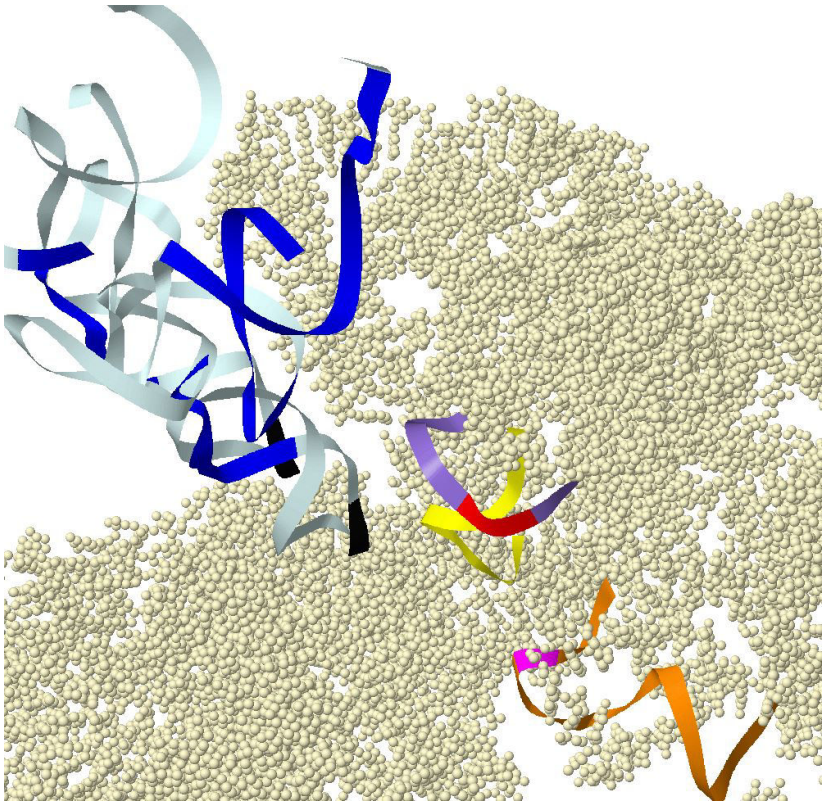

Figure 4.3: X circular code motifs involved in the archaeal ribosome decoding center of *Pyrococcus furiosus* (crystallographic structure PDB 3J20): the rRNA X motif m_{AA}(P. furiosus, 1445, 1456, 12) (purple with the conserved A1492 and A1493 nucleotides in red), the rRNA X motif m_G(P. furiosus, 480, 497, 18) (orange with the conserved nucleotide G530 in fuchsia), the rRNA X motif m(P. furiosus, 1356, 1364, 9) (yellow) and the tRNA X motifs (blue with the anti-codon in black). The remaining rRNA (lemonchiffon) is outside the neighbourhood of these X motifs and the mRNA is missing (Table 3.1).

4.2.1.1.4 The A1492 and A1493 nucleotides in X motifs in chloroplast rRNA

In the rRNA of *S. oleracea*, the conserved A1429 and A1493 nucleotides occur at the 2nd and 3rd sites of the trinucleotide $GAA \in X$ and belong to the X motif m_{AA}(*S. oleracea*, 1438, 1450, 13) $= GT, G\boldsymbol{AA}, GTC, GTA, AC$ which is a suffix of 13 nucleotides of the bacterial X motif m_{AA}(*E. coli*).

4.2.1.1.5 Summary of the A1492 and A1493 nucleotides in X motifs

In all the studied rRNAs, the universally conserved nucleotides A1492 and A1493 precede the trinucleotide $GTC \in X$, except in *T. thermophilus* where GTC is replaced by GGT. Thus, it always occurs at the 2nd and 3rd sites of a trinucleotide which is always GAA when the trinucleotide belongs to X. For one X motif m_{AA}(*S. cerevisiae*) $= m_{AA}$(*H. sapiens*), it is a suffix of X.

Figures 4.1-4.7 show that the rRNA X motifs m_{AA}(*E. coli*), m_{AA}(*T. ther-*

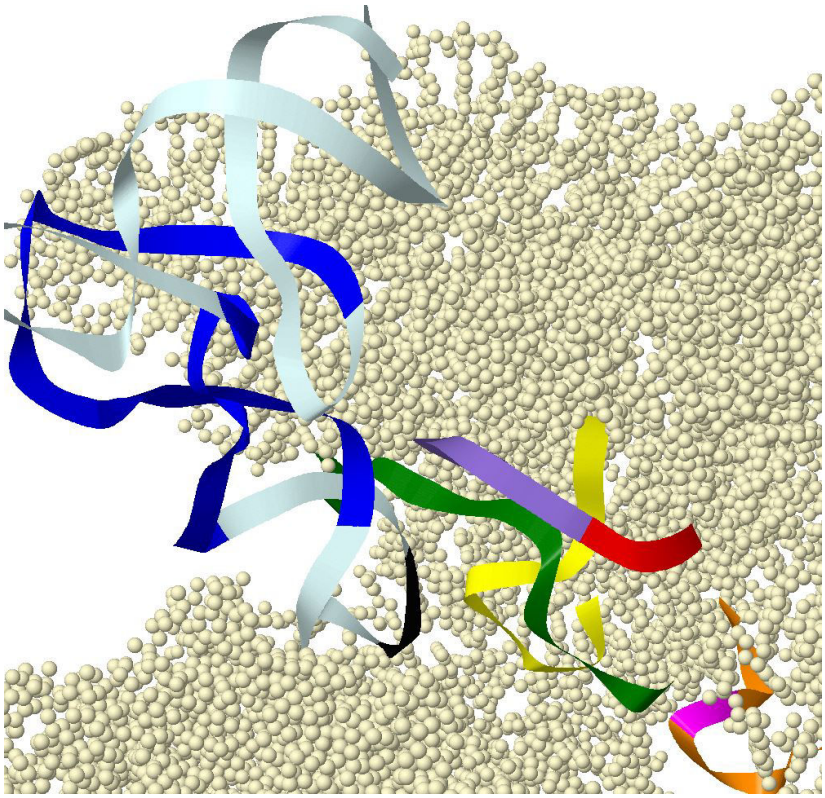

Figure 4.4: X circular code motifs involved in the nuclear eukaryotic ribosome decoding center of *Saccharomyces cerevisiae* (crystallographic structure PDB 3IZE): the mRNA X motifs (green), the rRNA X motif m_{AA}(*S. cerevisiae*, 1755, 1764, 10) (purple with the conserved A1492 and A1493 nucleotides in red), the rRNA X motif m_G(*S. cerevisiae*, 574, 582, 9) (orange with the conserved nucleotide G530 in fuchsia), the rRNA X motif m(*S. cerevisiae*, 1633, 1641, 9) (yellow) and the tRNA X motifs (blue with the anti-codon in black). The remaining rRNA (lemonchiffon) is outside the neighbourhood of these X motifs.

mophilus), m_{AA}(*P. furiosus*), m_{AA}(*S. cerevisiae*), m_{AA}(*T. aestivum*), m_{AA}(*H. sapiens*), and m_{AA}(*S. oleracea*) (purple with the conserved dinucleotide AA in red) of bacteria, archaea, nuclear eukaryotes and chloroplasts belong to the ribosome decoding center with spatial relations with mRNA (green) and tRNA X motifs (blue with the anti-codon in black). Which allows us to observe possible interactions between the $m_{AA}X$ circular code motifs from the rRNA and the tRNA and mRNA sequences.

4.2.1.2 The conserved G530 nucleotide

We will detail the motifs containing the conserved G530 nucleotide. Unlike in the prokayotic organisms studied, the G530 was not identified experimentally in eukaryotes. **We will show the potential G530 nucleotide that is responsible for that role in eukaryotes (which was not identified yet**

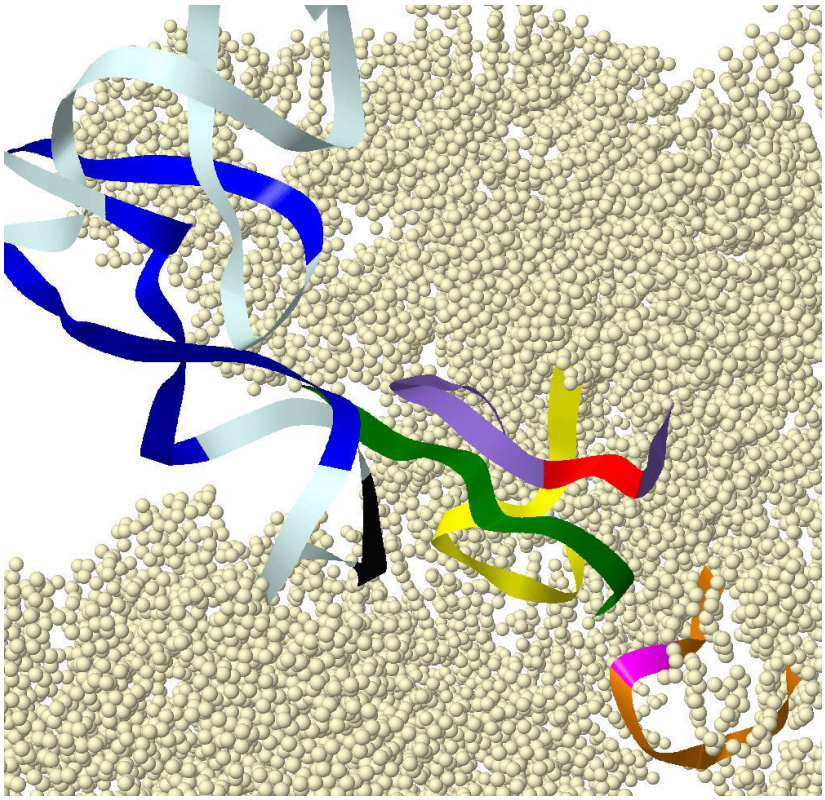

Figure 4.5: X circular code motifs involved in the nuclear eukaryotic ribosome decoding center of *Triticum aestivum* (crystallographic structure PDB 3J5Z): the mRNA (green), the rRNA X motif m_{AA}(*T. aestivum*, 1763, 1774, 12) (purple with the conserved A1492 and A1493 nucleotides in red), the rRNA X motif m_G(*T. aestivum*, 578, 586, 9) (orange with the conserved nucleotide G530 in fuchsia), the rRNA X motif m(*T. aestivum*, 1641, 1649, 9) (yellow) and the tRNA X motifs (blue with the anti-codon in black). The remaining rRNA (lemonchiffon) is outside the neighbourhood of these X motifs.

experimentally) using circular code theory.

4.2.1.2.1 The G530 nucleotide in X motifs in bacterial rRNA

In the rRNA of *E. coli*, the conserved nucleotide G occurs at the 2nd site of the trinucleotide $GGT \in X$ and belongs to the X motif m_G(*E. coli*, 527, 536, 10) = GC, GGT, AAT, AC of 10 nucleotide length starting with the dinucleotide GC suffix of $GGC \in X$ followed by two trinucleotides $AAT, GGT \in X$ and ending with the dinucleotide AC prefix of $ACC \in X$. In the rRNA of *T. thermophilus*, the conserved nucleotide G occurs at the 1st site of $GTT \in X$ and belongs to the X motif m_G(*T. thermophilus*, 528, 536, 9) = GC, GTT, ACC, C of nine nucleotide length starting with the dinucleotide GC, as in *E. coli*, followed by two trinucleotides $ACC, GTT \in X$ and ending with the nucleotide C prefix of $CAG, CTC, CTG \in X$. As with the conserved A1492 and A1493 nucleotides,

Table 4.2: Identification of X circular code motifs m_G containing the conserved nucleotide G530 (in bold) in rRNAs of bacteria and archaea. The bottom half of the table shows the X circular code motifs potentially containing the equivalent G530 in nuclear eukaryotes and chloroplasts.

PDB ID	Kingdom	Organism	X motif (m_G)	Start	End	Length
3J5T	Bacteria	*E. coli*	GC,G**G**T,AAT,AC	527	536	10
3I8G	Bacteria	*T. thermophilus*	GC,**G**TT,ACC,C	528	536	9
3J20	Archaea	*P. furiosus*	GC,G**G**T,AAT,ACC,GGC,GGC,C	480	497	18
3IZE	Eukaryote (nuclear)	*S. cerevisiaie*	GC,GGT,AAT,T	574	582	9
3J5Z	Eukaryote (nuclear)	*T. aestivum*	GC,GGT,AAT,T	578	586	9
3J3D	Eukaryote (nuclear)	*H. sapiens*	GC,GGT,AAT,T	623	631	9
3BBN	Eukaryote (chloroplast)	*S. oleracea*	GC,GGT,AA	475	481	7

the two rRNA X motifs m_G(*E. coli*) and m_G(*T. thermophilus*) have completely different primary structures and can only be revealed by the circular code theory.

4.2.1.2.2 The G530 nucleotide in X motifs in archaea rRNA

In rRNA of *P. furiosus*, the conserved nucleotide G occurs at the 2nd site of the trinucleotide $GGT \in X$, as in m_G(*E. coli*), and belongs to the X motif m_G(*P. furiosus*, 480, 497, 18) $= GC, GGT, AATACC, GGC, GGC, C$ of 18 nucleotide length starting with the dinucleotide GC, as in m_G(*E. coli*) and m_G(*T. thermophilus*), followed by five trinucleotides $AAT, ACC, GGC, GGT \in X$ and ending with the nucleotide C, as in m_G(*T. thermophilus*).

4.2.1.2.3 The G530 nucleotide in X motifs in eukaryotic rRNA

By applying our modifications mentioned in section 3.4.1 on the results of the global multiple sequence alignment ClustalX, the X motifs m_G are found in rRNAs of nuclear eukaryotes *S. cerevisiae*, *T. aestivum*, *H. sapiens* and chloroplasts *S. oleracea*. Very surprisingly, a common X motif m_G is identified in rRNAs of nuclear eukaryotes: m_G(S. cerevisiae, 574, 582, 9) $= m_G$(T. aestivum, 578, 586, 9) $= m_G$(H. sapiens, 623, 631, 9) $= m_G$(Nuclear eukaryotes) $= GC, GGT, AAT, T$ (Table 4.2). Furthermore, a X motif m_G is also identified in rRNA of chloroplasts: m_G(S. oleracea, 475, 481, 7) $= m_G$(Chloroplasts) $= GC, GGT, AA$ (Table 4.2) which is a prefix of seven nucleotides of m_G(Nuclear eukaryotes). As the common X motif m_G(Nuclear eukaryotes, Chloroplasts) $= GC, GGT, AA$ is a prefix of the common X motif m_G(*E. coli*, *P. furiosus*) $= GC, GGT, AAT, AC$, we can make the realistic hypothesis that the conserved

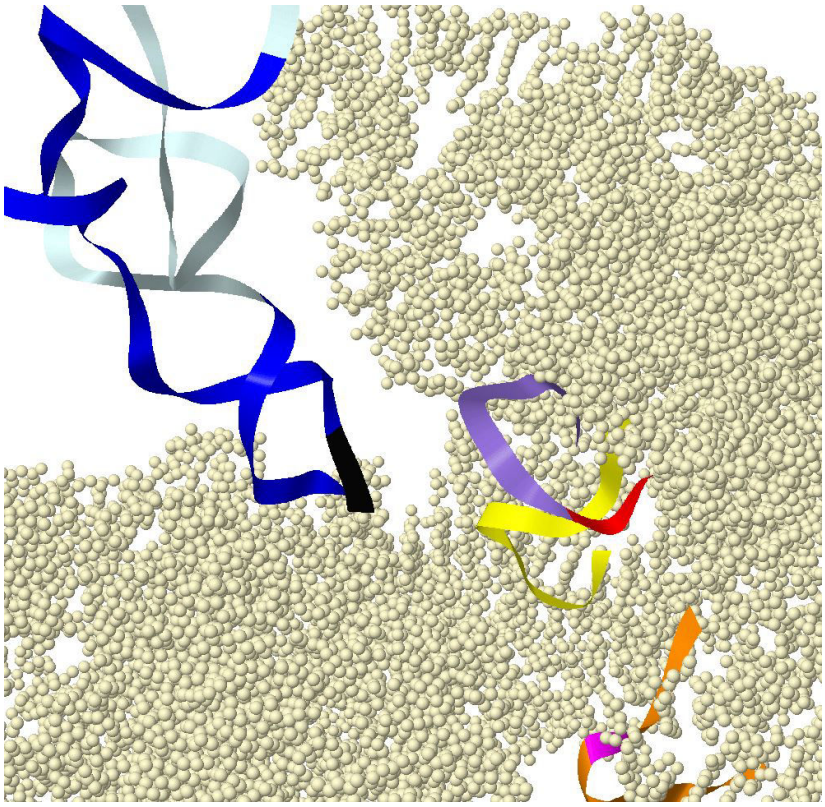

Figure 4.6: X circular code motifs involved in the nuclear eukaryotic ribosome decoding center of *Homo sapiens* (crystallographic structure PDB 3J3D): the rRNA X motif m_{AA}(*H. sapiens*, 1824, 1833, 10) (purple with the conserved A1492 and A1493 nucleotides in red), the rRNA X motif m_G(*H. sapiens*, 623, 631, 9) (orange with the conserved nucleotide G530 in fuchsia), the rRNA X motif m(*H. sapiens*, 1697, 1705, 9) (yellow) and the tRNA X motifs (blue with the anti-codon in black). The remaining rRNA (lemonchiffon) is outside the neighbourhood of these X motifs and the mRNA is missing (Table 3.1).

nucleotide G in nuclear and chloroplast rRNAs occurs at the 2nd site of the trinucleotide $GGT \in X$.

Furthermore, Figures 4.4-4.7 show that the rRNA X motifs $m_G(S.\ cerevisiae)$, $m_G(T.\ aestivum)$, $m_G(H.\ sapiens)$ and $m_G(S.\ oleracea)$ (orange with the conserved nucleotide G in fuchsia) of nuclear eukaryotes and chloroplasts belong to the ribosome decoding center with spatial relations with mRNA (green) and tRNA X motifs (blue with the anti-codon in black).

4.2.1.2.4 Summary of the G530 nucleotide in X motifs

The bacterial X motif $m_G(E.\ coli)$ is a prefix of 10 nucleotides of the archaeal X motif $m_G(P.\ furiosus)$. The conserved nucleotide G530 occurs at the 2nd site of $GGT \in X$ in $m_G(E.\ coli)$ and $m_G(P.\ furiosus)$, and at the 1st site of $GTT \in X$ in $m_G(T.\ thermophilus)$. Figures 4.1-4.3 show that the rRNA X

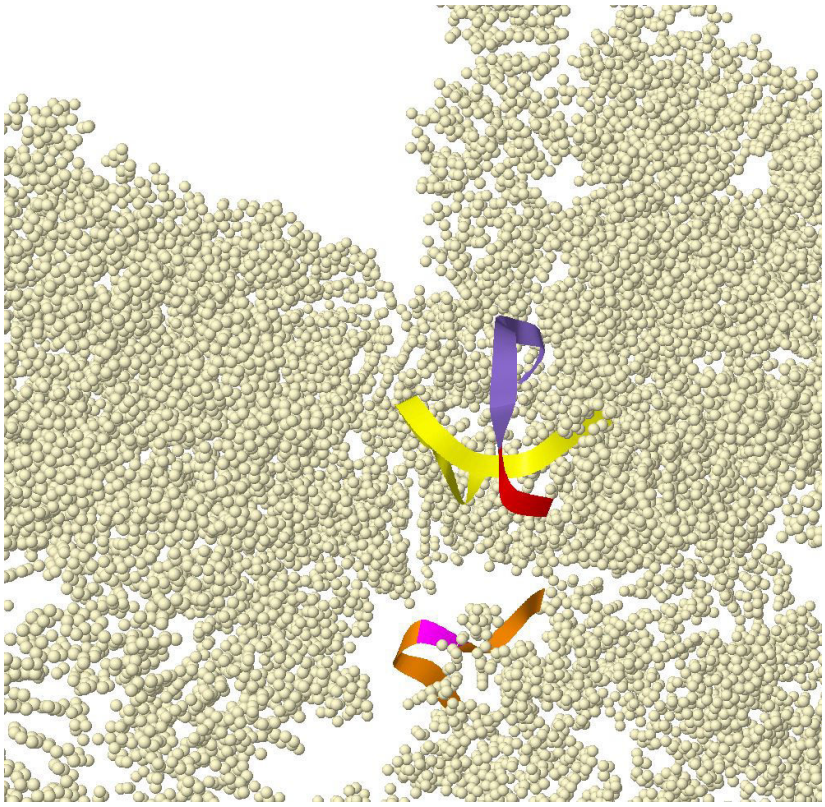

Figure 4.7: X circular code motifs involved in the chloroplast eukaryotic ribosome decoding center of *Spinacia oleracea* (crystallographic structure PDB 3BBN): the rRNA X motif m_{AA}(*S. oleracea*, 1438, 1450, 13) (purple with the conserved A1492 and A1493 nucleotides in red), the rRNA X motif m_G(*S. oleracea*, 475, 481, 7) (orange with the conserved nucleotide G530 in fuchsia) and the rRNA X motif m(*S. oleracea*, 1345, 1353, 9) (yellow). The remaining rRNA (lemonchiffon) is outside the neighbourhood of these X motifs, the mRNA and tRNA are missing (Table 3.1).

motifs m_G(*E. coli*), m_G(*T. thermophilus*) and m_G(*P. furiosus*) (orange with the conserved nucleotide G in fuchsia) of bacteria and archaea belong to the ribosome decoding center with spatial relations with mRNA (green) and tRNA X motifs (blue with the anti-codon in black).

Using prior knowledge of the G530 in bacteria, we were able to identify a possible location of its equivalent in eukaryotes. This was based on similarities of motifs in terms of composition and their spatial location in the ribosome when comparing between prokaryotes and eukaryotes. The eukaryotic X motifs m_G are identical for the nuclear eukaryotes GC, GGT, AAT, T, where as the chloroplast X motif m_G is two nucleotides shorter having possibly lost the third nucleotide T of the trinucleotide $AAT \in X$ to mutation same as the following T nucleotide (Table 4.2).

Table 4.3: Identification of a conserved X circular code motifs m in all studied rRNAs of bacteria, archaea, nuclear eukaryotes, and chloroplasts

PDB ID	Kingdom	Organism	X motif (m)	Start	End	Length
3J5T	Bacteria	*E. coli*	AC,ACC,GCC,C	1396	1404	9
3I8G	Bacteria	*T. thermophilus*	AC,ACC,GCC,C	1375	1383	9
3J20	Archaea	*P. furiosus*	AC,ACC,GCC,C	1356	1364	9
3IZE	Eukaryote (nuclear)	*S. cerevisiaie*	AC,ACC,GCC,C	1633	1641	9
3J5Z	Eukaryote (nuclear)	*T. aestivum*	AC,ACC,GCC,C	1641	1649	9
3J3D	Eukaryote (nuclear)	*H. sapiens*	AC,ACC,GCC,C	1697	1705	9
3BBN	Eukaryote (chloroplast)	*S. oleracea*	AC,ACC,GCC,C	1345	1353	9

4.2.2 Spatial study of X circular code motifs near the ribosomal decoding center

At this stage we shifted our region of interest to around the decoding center to examine possible conserved X motif in the vicinity of mRNA and tRNAs that could have a possible role in the translation process by interaction.

4.2.2.1 A conserved X motif near the ribosomal decoding center

We were able to identify an X motif m which is universally conserved in rRNAs of bacteria, archaea, nuclear eukaryotes, and chloroplasts (Table 4.3): $m =$ AC,ACC,GCC,C of nine nucleotide length starting with the dinucleotide AC suffix of AAC,GAC, TAC $\in X$ followed by two trinucleotides ACC,GCC $\in X$ and ending with the nucleotide C prefix of CAG,CTC,CTG $\in X$. The start and end positions of the X motif m in the seven studied organisms are given in Table 4.3. Very unexpectedly, Figures 4.1-4.7 show that the universally conserved rRNA X motif m (yellow) of bacteria, archaea, nuclear eukaryotes, and chloroplasts belongs to the ribosome decoding center with spatial relations with mRNA (green) and tRNA X motifs (blue with the anti-codon in black). With the sole exception being the *T. thermophilus* where this conserved motif is located outside the decoding center, but as mentioned before this organism has several differences in terms of structure than the others.

Table 4.4: Identification of seven X circular code motifs $PrRNAXm$ in 16S rRNAs of prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*) near the ribosome decoding center.

Alias	X circular code motif	Organism	Start	End	Length
	G,GAG,GGT,GC	*E. coli* (3J5T)	537	545	9
$PrRNAXm_1$	G,GAG,GGC,GC	*T. thermophilus* (3I8G)	517	525	9
	GC,GGT,AAT,ACC,GGC,GGC,C	*P. furiosus* (3J20)	480	497	18
	GC,GGT,GAA,AT	*E. coli* (3J5T)	688	697	10
$PrRNAXm_2$	GC,GGT,GAA,AT	*T. thermophilus* (3I8G)	668	677	10
	G,GGT,GAA,ATC,CT	*P. furiosus* (3J20)	643	654	12
	G,AAT,ACC,GGT,GGC,GAA,GGC,GGC,C	*E. coli* (3J5T)	714	736	23
$PrRNAXm_3$	G,AAC,GCC,GAT,GGC,GAA,GGC,A	*T. thermophilus* (3I8G)	694	713	20
	GT,GGC,GAA,GGC,GCC,C	*P. furiosus* (3J20)	676	690	15
	TA,GAT,ACC,CTG,GTA,GTC,CA	*E. coli* (3J5T)	789	807	19
$PrRNAXm_4$	TA,GAT,ACC,C	*T. thermophilus* (3I8G)	769	777	9
	TA,GAT,ACC,C	*P. furiosus* (3J20)	743	751	9
	G,GAT,GAC,GTC,AA	*E. coli* (3J5T)	1186	1197	12
$PrRNAXm_5$	G,GAC,GAC,GTC,T	*T. thermophilus* (3I8G)	1164	1174	11
	G,GGC,GAC,GGT,A	*P. furiosus* (3J20)	1146	1188	9
	T,TAC,GAC,CAG,GGC,TAC,AC	*E. coli* (3J5T)	1211	1228	18
$PrRNAXm_6$	T,TAC,GGC,CTG,GGC,GAC,AC	*T. thermophilus* (3I8G)	1189	1206	18
	G,GGC,TAC,AC	*P. furiosus* (3J20)	1180	1188	9
	AC,GGT,GAA,TAC,GTT,C	*E. coli* (3J5T)	1368	1382	15
$PrRNAXm_7$	GC,GGT,GAA,TAC,GTT,C	*T. thermophilus* (3I8G)	1347	1361	15
	GC,GGC,GAA,TAC,GTC,C	*P. furiosus* (3J20)	1328	1342	15

4.2.2.2 Conserved X motifs in the rRNA of prokaryotes

The circular code theory identified seven X circular code motifs, $PrRNAXm$, that are conserved in the prokaryotic 16s rRNA of bacteria *E. coli*(3J5T) and *T. thermophilus*(3I8G), and archaea *P. furiosus*(3J20) (Table 4.4).

(i) $PrRNAXm_1$(*E. coli*, 537, 545, 9) = $PrRNAXm_1$(*T. thermophilus*, 517, 525, 9) = G, GAG, GGY, GC of nine nucleotides starts with the nucleotide G suffix of $\{CAG, CTG, GAG\} \in X$, has two trinucleotides $GAG, GGY \in X$ where $Y = T$ in *E. coli* and $Y = C$ in *T. thermophilus*, and ends with the dinucleotide GC prefix of $GCC \in X$. The large X motif $PrRNAXm_1$(*P. furiosus*, 480, 497, 18) = $GC, GGT, AAT, ACC, GGC, GGC, C$ of 18 nucleotides starts with the dinucleotide GC suffix of $GGC \in X$, has five trinucleotides $GGT, AAT, ACC, GGC, GGC \in X$ and ends with the nucleotide C prefix of $\{CAG, CTC, CTG\} \in X$. $PrRNAXm_1$ of *E. coli* and *T. thermophilus* are partial suffixes of $PrRNAXm_1$ of *P. furiosus*.

(ii) $PrRNAXm_2$(*E. coli*, 688, 697, 10), $PrRNAXm_2$(*T. thermophilus*, 668, 677, 10) and $PrRNAXm_2$(*P. furiosus*, 643, 654, 12) have the com-

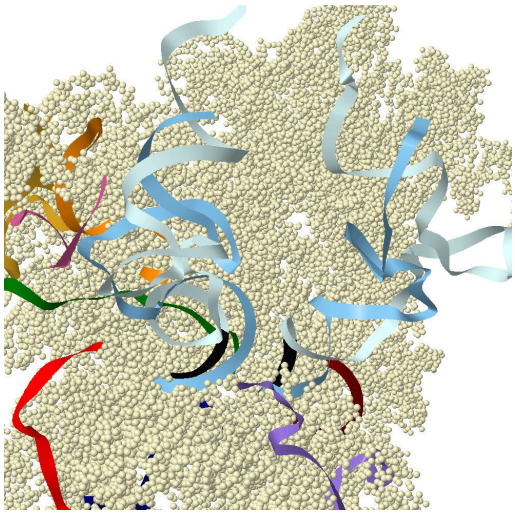

Figure 4.8: X circular code motifs near the bacterial ribosome decoding center of *Escherichia coli* (PDB 3J5T): the mRNA (green), the rRNA X motifs $PrRNAXm_1$(537, 545, 9) (maroon), $PrRNAXm_2$(688, 697, 10) (pink), $PrRNAXm_3$(714, 736, 23) (gold), $PrRNAXm_4$(789, 807, 19) (orange), $PrRNAXm_5$(1186, 1197,12) (navy), $PrRNAXm_6$(1211,1228,18) (purple), $PrRNAXm_7$(1368,1382,15) (red), and the tRNA (50 region in dark blue, 30 region in clearer blue and anti-codon in black). The remaining rRNA (lemonchiffon) is outside the neighborhood of these X motifs.

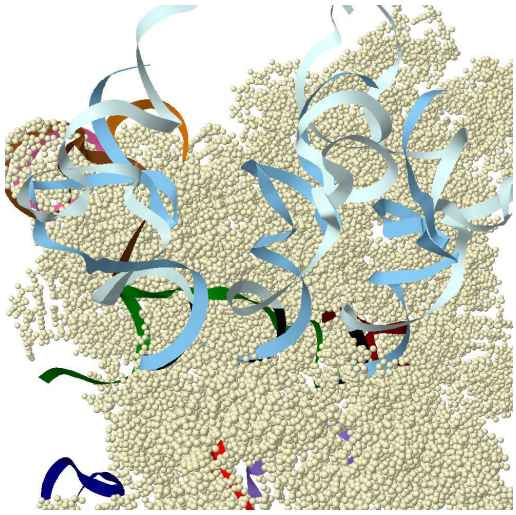

Figure 4.9: X circular code motifs near the bacterial ribosome decoding center of *Thermus thermophilus* (PDB ID: 3I8G): the mRNA (green), the rRNA X motifs $PrRNAXm_1$(517, 525, 9) (maroon), $PrRNAXm_2$(668, 677, 10) (pink), $PrRNAXm_3$(694, 713, 20) (gold), $PrRNAXm_4$(769, 777, 9) (orange), $PrRNAXm_5$(1164, 1174, 11) (navy), $PrRNAXm_6$(1189, 1206, 18) (purple), $PrRNAXm_7$(1347, 1361, 15) (red), and the tRNA (50 region in dark blue, 30 region in clearer blue and anti-codon in black). The remaining rRNA (lemonchiffon) is outside the neighborhood of these X motifs.

mon X motif GGT, GAA, AT of eight nucleotides with $GGT, GAA \in X$. $PrRNAXm_2$ of *E. coli* and *T. thermophilus* start with the dinucleotide GC suffix of $GGC \in X$ and end with the dinucleotide AT prefix of $\{ATC, ATT\} \in X$. $PrRNAXm_2$ of *P. furiosus* starts with the nucleotide G suffix of $\{CAG, CTG, GAG\} \in X$ and ends with the dinucleotide CT prefix of $\{CTC, CTG\} \in X$.

(iii) $PrRNAXm_3$(*E. coli*, 714, 736, 23) and $PrRNAXm_3$(*T. thermophilus*, 694, 713, 20) have the large common X motif $G, AAY, R_1CC, GR_2T, GGC, GAA, GGC$ of 19 nucleotides starting with the nucleotide G suffix of $\{CAG, CTG, GAG\} \in X$ followed by six trinucleotides $AAY, R_1CC, GR_2T, GGC, GAA, GGC \in X$ where Y = T, R_1 = A and R_2 = G in *E. coli*, while Y = C, R_1 = G and R_2 = A in *T. thermophilus*. $PrRNAXm_3$of *E. coli* ends with the nucleotide C prefix of $\{CAG, CTC, CTG\} \in X$ whereas

$PrRNAXm_3$ of *T. thermophilus* ends with the nucleotide A prefix of $\{AAC, AAT, ACC, ATC, ATT\} \in X$. The X motif $PrRNAXm_3$(P. furiosus, 676, 690, 15) $= GT, GGC, GAA, GGC, GCC, C$ of 15 nucleotides is a conserved suffix of $PrRNAXm_3$of *E. coli* (14 identical letters among 15) starting with the dinucleotide GT suffix of GGT in $PrRNAXm_3$ of *E. coli*.

(iv) The large X motif $PrRNAXm_4$(*E. coli*, 789, 807, 19) $= TA, GAT, ACC, CTG, GTA, GTC, CA$ of 19 nucleotides starts with the dinucleotide TA suffix of $GTA \in X$, has five trinucleotides $GAT, ACC, CTG, GTA, GTC \in X$ and ends with the dinucleotide CA prefix of $CAG \in X$. $PrRNAXm_4$(*T. thermophilus*, 769, 777, 9) $= PrRNAXm_4$(*P. furiosus*, 743, 751, 9) $= TA, GAT, ACC, C$ of nine nucleotides is a prefix of $PrRNAXm_4$ of *E. coli* ending with the nucleotide C prefix of CTG in $PrRNAXm_4$ of *E. coli*.

(v) $PrRNAXm_5$(*E. coli*, 1186, 1197, 12), $PrRNAXm_5$(*T. thermophilus*, 1164, 1174, 11) and $PrRNAXm_5$(*P. furiosus*, 1146, 1156, 11) have the common X motif G, GRY_1, GAC, GKY_2, W of 11 nucleotides starting with the nucleotide G suffix of $\{CAG, CTG, GAG\} \in X$ followed by three trinucleotides $GRY_1, GAC, GKY_2 \in X$ where $R = A$, $Y_1 = T$, $K = T$, $Y_2 = C$ and $W = A$ in *E. coli*, $R = A$, $Y_1 = C$, $K = T$, $Y_2 = C$ and $W = T$ in *T. thermophilus* while $R = G$, $Y_1 = C$, $K = G$, $Y_2 = T$ and $W = A$ in *P. furiosus*. $PrRNAXm_5$ of *E. coli* ends with the dinucleotide AA prefix of $\{AAC, AAT\} \in X$, $PrRNAXm_5$ of *T. thermophilus* ends with the nucleotide T prefix of $\{TAC, TTC\} \in X$ and $PrRNAXm_5$ of *P. furiosus* ends with the nucleotide A prefix of $\{AAC, AAT, ACC, ATC, ATT\} \in X$.

(vi) The large common X motif $PrRNAXm_6$(*E. coli*, 1211, 1228, 18) $= PrRNAXm_6$(*T. thermophilus*, 1189, 1206, 18) $= T, TAC, GRC, CWG, GGC, KAC, AC$ of 18 nucleotides starts with the nucleotide T suffix of $\{AAT, ATT, GAT, GGT, GTT\} \in X$, has five trinucleotides $TAC, GRC, CWG, GGC, KAC \in X$ where $R = A$, $W = A$ and $K = T$ in *E. coli* while $R = G$, $W = T$ and $K = G$ in *T. thermophilus*, and ends with the dinucleotide AC prefix of $ACC \in X$. $PrRNAXm_6$(*P. furiosus*, 1180, 1188, 9) $= G, GGC, TAC, AC$ of nine nucleotides is a suffix of $PrRNAXm_6$ of *E. coli* starting with the nucleotide G suffix of CAG in

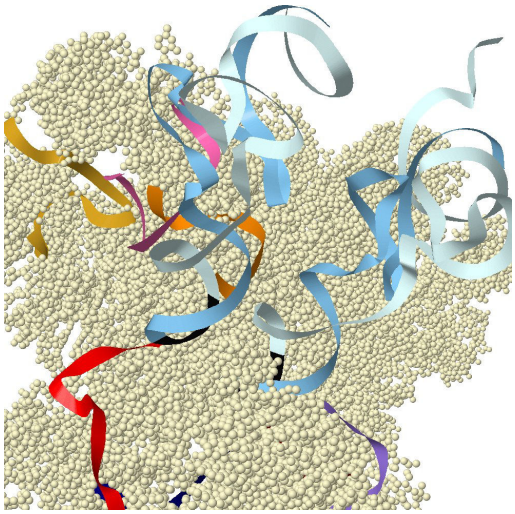

Figure 4.10: X circular code motifs near the archaeal ribosome decoding center of *Pyrococcus furiosus* (PDB ID: 3J20): the rRNA Xmotifs $PrRNAXm_1$(480, 497, 18) (maroon), $PrRNAXm_2$(643, 654, 12) (pink), $PrRNAXm_3$(676, 690, 15) (gold), $PrRNAXm_4$(743, 751, 9) (orange), $PrRNAXm_5$(1146, 1156, 11) (navy), $PrRNAXm_6$(1180, 1188, 9) (purple), $PrRNAXm_7$(1328, 1342, 15) (red), and the tRNA (50 region in dark blue, 30 region in clearer blue and anti-codon in black). The remaining rRNA (lemonchiffon) is outside the neighbourhood of these X motifs and the mRNA is missing (Table 3.1).

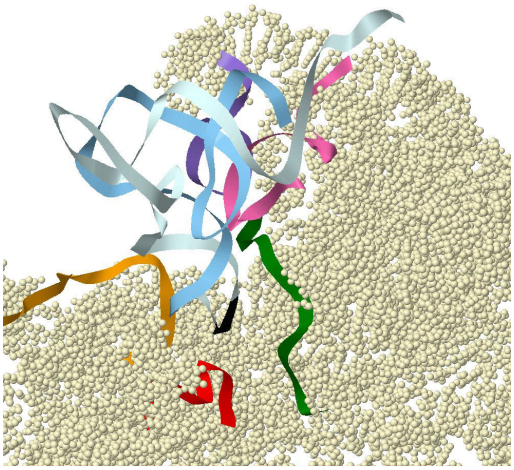

Figure 4.11: X circular code motifs near the (nuclear) eukaryotic ribosome decoding center of *Saccharomyces cerevisiae* (PDB ID: 3IZE): the mRNA (green), the rRNA X motifs $ErRNAXm_1$(900, 911, 12) (purple), $ErRNAXm_2$(987, 1004, 18) (pink), $ErRNAXm_3$(1189, 1197, 9) (red), $ErRNAXm_4$(1564, 1582, 19) (orange), and the tRNA (50 region in dark blue, 30 region in clearer blue and anti-codon in black). The remaining rRNA (lemonchiffon) is outside the neighbourhood of these X motifs.

$PrRNAXm_6$ of *E. coli*.

(vii) The common X motif $PrRNAXm_7$(*E. coli*, 1368, 1382, 15) = $PrRNAXm_7$(*T. thermophilus*, 1347, 1361, 15) = $PrRNAXm_7$(*P. furiosus*, 1328, 1342, 15) = $RC, GGY, GAA, TAC, GTY, C$ of 15 nucleotides start with the dinucleotide AC ($R = A$) suffix of $\{AAC, GAC, TAC\} \in X$ in *E. coli* and with the dinucleotide GC ($R = G$) suffix of $GGC \in X$ in *T. thermophilus* and *P. furiosus*, has four trinucleotides $GGY, GAA, TAC, GTY \in X$ where $Y = T$ in *E. coli* and *T. thermophilus* while $Y = C$ in *P. furiosus*, and ends with the nucleotide C prefix of $\{CAG, CTC, CTG\} \in X$.

Figures 4.8-4.10 show the prokaryotic rRNA X motifs $PrRNAXm_1$ in maroon, $PrRNAXm_2$ in pink, $PrRNAXm_3$ in gold, $PrRNAXm_4$ in orange, $PrRNAXm_5$ in navy blue, $PrRNAXm_6$ in purple and $PrRNAXm_7$ in red

of *E. coli*, *T. thermophilus* and *P. furiosus* are near the ribosome decoding center (50 regions of tRNAs in dark blue, 30 regions of tRNAs in clearer blue and anti-codons of tRNAs in black).

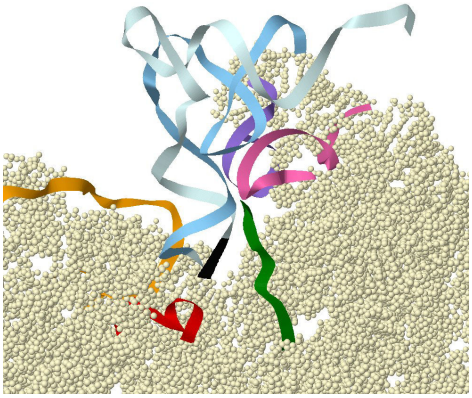

Figure 4.12: X circular code motifs near the (nuclear) eukaryotic ribosome decoding center of *Triticum aestivum* (PDB ID: 3J5Z): the mRNA (green), the rRNA X motifs $ErRNAXm_1$(905, 916, 12) (purple), $ErRNAXm_2$(992, 1009, 18) (pink), $ErRNAXm_3$(1193, 1201, 9) (red), $ErRNAXm_4$(1575, 1596, 12) (orange), and the tRNA (50 region in dark blue, 30 region in clearer blue and anti-codon in black). The remaining rRNA (lemonchiffon) is outside the neighbourhood of these X motifs.

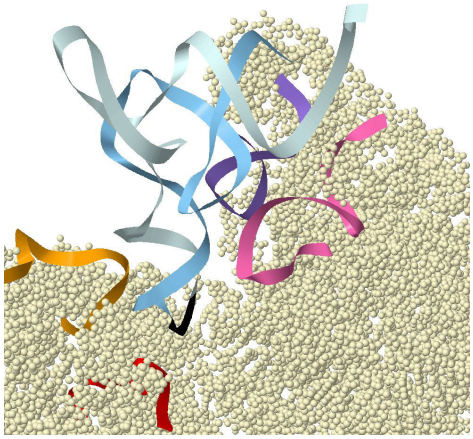

Figure 4.13: X circular code motifs near the (nuclear) eukaryotic ribosome decoding center of *Homo Sapiens* (PDB ID: 3J3D): the rRNA X motifs $ErRNAXm_1$(957, 968, 12) (purple), $ErRNAXm_2$(1044, 1061, 18) (pink), $ErRNAXm_3$(1246, 1254, 9) (red), $ErRNAXm_4$(1631, 1645, 15) (orange), and the tRNA (50 region in dark blue, 30 region in clearer blue and anti-codon in black). The remaining rRNA (lemonchiffon) is outside the neighborhood of these X motifs and the mRNA is missing (Table 3.1).

4.2.2.3 Conserved X motifs in the rRNA of eukaryotes

The circular code theory identified four X circular code motifs, $PrRNAXm$, that are conserved in the eukaryotic 18s rRNA of *S. cerevisiae* (3IZE), *T. aestivum* (3J5Z) and *H. sapiens* (3J3D) (Table 4.5)

(i) $ErRNAXm_1$(*S. cerevisiae*, 900, 911, 12) = $ErRNAXm_1$(*T. aestivum*, 905, 916, 12) = $ErRNAXm_1$(*H. sapiens*, 957, 968, 12) = A, GGT, GAA, ATT, CT of 12 nucleotides starts with the nucleotide A suffix of $\{GAA, GTA\} \in X$, has three trinucleotides $GGT, GAA, ATT \in X$ and ends with the dinucleotide CT prefix of $\{CTC, CTG\} \in X$.

(ii) The large common X motif $ErRNAXm_2$(*S. cerevisiae*, 987, 1004, 18) = $ErRNAXm_2$(*T. aestivum*, 992, 1009, 18) = $ErRNAXm_2$(*H. sapiens*,

Table 4.5: Identification of four X circular code motifs $ErRNAXm$ in 18s rRNAs of (nuclear) eukaryotes (*S. cerevisiae*, *T. aestivum*, *H. sapiens*) near the ribosome decoding center.

Alias	X circular code motif	Organism	Start	End	Length
$ErRNAXm_1$	A,GGT,GAA,ATT,CT	*S. cerevisiae* (3IZE)	900	911	12
	A,GGT,GAA,ATT,CT	*T. aestivum* (3J5Z))	905	916	12
	A,GGT,GAA,ATT,CT	*H. sapiens* (3J3D)	957	968	12
$ErRNAXm_2$	G,ATC,GAA,GAT,GAT,CAG,AT	*S. cerevisiae* (3IZE)	987	1004	18
	G,CTC,GAA,GAC,GAT,CAG,AT	*T. aestivum* (3J5Z))	662	1009	18
	G,TTC,GAA,GAC,GAT,CAG,AT	*H. sapiens* (3J3D)	1044	1061	18
$ErRNAXm_3$	A,CTC,AAC,AC	*S. cerevisiae* (3IZE)	1189	1197	9
	A,CTC,AAC,AC	*T. aestivum* (3J5Z))	1193	1201	9
	A,CTC,AAC,AC	*H. sapiens* (3J3D)	1246	1254	9
$ErRNAXm_4$	TC,TTC,AAC,GAG,GAA,TTC,CT	*S. cerevisiae* (3IZE)	1564	1582	19
	TC,AAC,GAG,GAA,T	*T. aestivum* (3J5Z))	1575	1596	12
	TG,AAC,GAG,GAA,TTC,C	*H. sapiens* (3J3D)	1631	1645	15

1044, 1061, 18) $= G, NTC, GAA, GAY, GAT, CAG, AT$ of 18 nucleotides starts with the nucleotide G suffix of $\{CAG, CTG, GAG\} \in X$, has five trinucleotides $NTC, GAA, GAY, GAT, CAG \in X$ where $N = A$ and $Y = T$ in *S. cerevisiae*, $N = C$ and $Y = C$ in *T. aestivum*, while $N = T$ and $Y = C$ in *H. sapiens*, and ends with the dinucleotide AT prefix of $\{ATC, ATT\} \in X$.

(iii) $ErRNAXm_3$(*S. cerevisiae*, 1189, 1197, 9) $= ErRNAXm_3$ (*T. aestivum*, 1193, 1201, 9) $= ErRNAXm_3$(*H. sapiens*, 1246, 1254, 9) $= A, CTC, AAC, AC$ of nine nucleotides starts with the nucleotide A suffix of $\{GAA, GTA\} \in X$, has two trinucleotides $CTC, AAC \in X$ and ends with the dinucleotide AC prefix of $ACC \in X$.

(iv) The large X motif $ErRNAXm_4$(*S. cerevisiae*, 1564, 1582, 19) $= TC, TTC, AAC, GAG, GAA, TTC, CT$ of 19 nucleotides starts with the dinucleotide TC suffix of $\{ATC, CTC, GTC, TTC\} \in X$, has five trinucleotides $TTC, AAC, GAG, GAA, TTC \in X$ and ends with the dinucleotide CT prefix of $\{CTC, CTG\} \in X$. $ErRNAXm_4$(*T. aestivum*, 1575, 1596, 12) $= TC, AAC, GAG, GAA, T$ is a factor of $ErRNAXm_4$ of *S. cerevisiae* starting with the dinucleotide TC suffix of the 1st $TTC \in X$ in $ErRNAXm_4$ of *S. cerevisiae* and ending with the nucleotide T prefix of the 2nd TTC $\in X$ in $ErRNAXm_4$ of *S. cerevisiae*. The X motif $ErRNAXm_4$(*H. sapiens*, 1631, 1645, 15) =

$TG, AAC, GAG, GAA, TTC, C$ is an almost exact suffix of $ErRNAXm_4$ of *S. cerevisiae*.

Figures 4.11-4.13 show the (nuclear) eukaryotic rRNA X motifs $ErRNAXm_1$ in purple, $ErRNAXm_2$ in pink, $ErRNAXm_3$ in red and $ErRNAXm_4$ in orange of *S. cerevisiae*, *T. aestivum* and *H. sapiens* are near the ribosome decoding center (50 regions of tRNAs in dark blue, 30 regions of tRNAs in clearer blue and anti-codons of tRNAs in black), except $ErRNAXm_1$ in *S. cerevisiae* and *T. aestivum*.

4.2.3 X CIRCULAR CODE MOTIFS IN PROKARYOTIC TRNAS

We give the main features of X motifs for each isoaccepting tRNA of prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*), more details on the X motifs are found in Tables 4.6-4.26. We use the classical genetic alphabet convention to be able to engulf X motifs that are fairly similar within a unified pattern. Let $R = \{A, G\}$, $Y = \{C, T\}$, $S = \{C, G\}$, $W = \{A, T\}$, $K = \{G, T\}$, $M = \{A, C\}$ and $N = \{A, C, G, T\}$. Furthermore, in term of X motif length, we are distinguishing three classes of X motifs: very large X motifs greater or equal to 20 nucleotides (remember that the average lengths of prokaryotic tRNAs range typically from 71 to 91 nucleotides for Cys and Ser, respectively, see Section 2.4.1 and Fig. 2 in Michel, 2013), large X motifs between 16 and 19 nucleotides and X motifs between 9 and 15 nucleotides. X motifs of lengths equal to 9 nucleotides already retrieve the reading frame with a probability of 99.9% and X motifs of lengths greater or equal to 12 nucleotides always retrieve, by definition, the reading frame, i.e. with a probability of 100% (Table 3 and Fig. 4 in Michel, 2012). Moreover, the underline in an X motif signifies that the underlined nucleotides are in common with one or more other motifs.

Finally, the X motifs are studied according to three regions of tRNAs: X motifs between the 5' end of tRNAs and the anti-codon (called here 5' region), X motifs having at least one nucleotide in the anti-codon (anti-codon regions) and X motifs between the anti-codon and the 3' end of tRNAs (3' region). The results below will identify X motifs, a few of them being very large, and different relations, in particular a shifting by 0, +1 or +2 mod 3 nucleotides with other X motifs or the anti-codon.

We would like to notify the reader that the following sections offer a exhaustive description of the X circular code motifs in tRNA sequences are presented.

These are particularly important as a reference for future studies targeting tRNAs.

4.2.3.1 X circular code motifs in Ala-tRNAs (Table 4.6)

(i) 5' region of Ala-tRNAs: The X motif T, CAG, CTG, GG and the class of X motifs GC, CTG, GWA, K are shifted in frame (modulo 3 according to their suffix-prefix). The class of X motifs GC, GCC, GCC, YT occurs before (5') the anti-codons GGC and TGC.

(ii) anti-codon regions of Ala-tRNAs: The X motif GC, GCC, GCC, CTC, GC is in a different frame than the anti-codon CGC. The very large X motif $Ala - tRNAXm_1$ $GC, CTC, AAT, GGC, ATT, GAG, GAG, GTC, A$ of 24 nucleotides in *T. thermophilus* is in frame with the anti-codon GGC. The X motif GC, CTG, AAT, C which is prefix of $Ala - tRNAXm_1$ is in frame with the anti-codon CGC.

(iii) 3' region of Ala-tRNAs: The class of X motifs $Ala - tRNAXm_2$ K, SAG, GAG, GTC, W is suffix of $Ala - tRNAXm_1$. The class of X motifs R, GAG, GYC, R is suffix of $Ala - tRNAXm_2$. Two X motifs are also observed: A, GGT, CAG, GG and CC, CTC, GGC, T.

Table 4.6: Identification of X circular code motifs ala-tRNAXm in tRNAs of Alanine (Ala) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
TGC 34	T,CAG,CTG,GG		*E. coli* C08004469	12	20	9
TGC 34	T,CAG,CTG,GG		*E. coli* C08004522	12	20	9
GGC 34	T,CAG,CTG,GG		*E. coli* C08004535	12	20	9
TGC 36	GC,CTG,GTA,T		*P. furiosus* At1825	15	23	9
GGC 36	GC,CTG,GTA,T		*P. furiosus* At1834	15	23	9
CGC 35	GC,CTG,GAA,GAG,C		*P. furiosus* At1833	15	26	12
TGC 36	GC,GCC,GCC,CT		*P. furiosus* At1825	26	35	10
GGC 36	GC,GCC,GCC,TT		*P. furiosus* At1834	26	35	10
CGC 35	GC,GCC,GCC,CT	C,GC	*P. furiosus* At1833	25	37	13
GGC 34	GC,CTC,AAT,	GGC, ATT,GAG,GAG,GTC,A	*T. thermophilus* C025943	26	49	24
CGC 34	GC,CTG,AAT,	C	*T. thermophilus* C025964	26	34	9
TGC 34		G,CAG,GAG,GTC,T	*E. coli* C08004469	39	49	11
TGC 34		G,CAG,GAG,GTC,T	*E. coli* C08004522	39	49	11
CGC 34		T,CAG,GAG,GTC,A	*T. thermophilus* C025964	39	49	11
GGC 34		AA,GAG,GTC,A	*E. coli* C08004535	41	49	9
CGC 35		G,GAG,GCC,GC	*P. furiosus* At1833	43	51	9
GGC 34		A,GGT,CAG,GG	*T. thermophilus* C025943	44	52	9
CGC 34		A,GGT,CAG,GG	*T. thermophilus* C025964	44	52	9
GGC 34		CC,CTC,GGC,T	*T. thermophilus* C025943	62	70	9
CGC 34		CC,CTC,GGC,T	*T. thermophilus* C025964	62	70	9

Table 4.7: Identification of X circular code motifs arg-tRNAXm in tRNAs of Arginine (Arg) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
TCG 36	G,GCC,GGT,GGC,CT		*P. furiosus* At1842	2	13	12
GCG 36	CC,GGT,GGC,CT		*P. furiosus* At1850	4	13	10
ACG 35	CC,GTA,GTT,CAG,CTG,GAT,A		*E. coli* C08004532	5	22	18
TCT 35	CC,GTA,GCC,TA		*P. furiosus* At1826	5	14	10
CCT 33	G,GTA,GCC,TA		*P. furiosus* At1852	6	14	9
CCG 36	G,GTA,GTT,TA		*P. furiosus* At1832	6	14	9
CCG 35	T,CAG,CTG,GAT,A		*E. coli* C08004500	12	22	11
ACG 35	T,CAG,CTG,GAT,A		*E. coli* C08004529	12	22	11
ACG 35	T,CAG,CTG,GAT,A		*T. thermophilus* C025939	12	22	11
CCG 35	T,CAG,CTG,GAT,A		*T. thermophilus* C025957	12	22	11
CCT 35	T,CAG,CAG,GAT,A		*T. thermophilus* C025920	12	22	11
TCT 35	AG,CAG,GAT,A		*P. furiosus* At1826	14	22	9
CCG 36	GC,CAG,GAG,A		*P. furiosus* At1832	15	23	9
CCT 33	GC,CAG,GAT,A		*P. furiosus* At1852	15	23	9
TCG 36	GC,CTG,GAT,GG		*P. furiosus* At1842	15	24	10
GCG 36	GC,CTG,GAT,A		*P. furiosus* At1850	15	23	9
CCT 31	AT,AAC,GAG,C		*E. coli* C08004487	21	29	9
ACG 35	TA,GAG,TAC,T		*E. coli* C08004529	21	29	9
ACG 35	TA,GAG,TAC,T		*E. coli* C08004532	21	29	9
TCT 35	AG,GGC,GCC,GGC,CT		*P. furiosus* At1826	22	34	13
TCG 36	G,GGC,GTC,GGC,CT		*P. furiosus* At1842	24	35	12
TCT 35	GC,AAC,GAC,CT		*E. coli* C08004473	25	34	10
CCG 36	A,GAG,AAC,GCC,GCC,CT	C,C	*P. furiosus* At1832	21	37	17
CCG 35	GC,GTC,GGC,CT	C,C	*T. thermophilus* C025957	25	36	12
CCT 36	AG,GGC,GGC,GGC,CT	C,CT	*P. furiosus* At1852	23	38	16
CCT 35	G,GGC,TT	C,CT	*T. thermophilus* C025920	29	37	9
TCT 34	A,GCC,GCC,T	TC,T A	*T. thermophilus* C025953	26	37	12
TCT 35	G,GCC,T	TC,T A	*P. furiosus* At1826	30	38	9
TCT 35	G,ACC,T	TC,T A	*E. coli* C08004473	30	38	9
TCG 36	G,GCC,T	TC,G G	*P. furiosus* At1842	31	39	9
ACG 35	A,GTA,CTC,GGC,T	AC,G AA,C	*E. coli* C08004529	24	40	17
ACG 35	A,GTA,CTC,GGC,T	AC,G AA,C	*E. coli* C08004532	24	40	17
ACG 35	T,GAC,T	AC,G G	*T. thermophilus* C025939	30	38	9
GCG 36		G, GAC,CTC,GAG,GTC,C	*P. furiosus* At1850	38	51	14
ACG 35		G, GAT,CAG,CAG,GTC,GG	*T. thermophilus* C025939	37	51	15
CCT 31		TG,CAG,GTT,C	*E. coli* C08004487	47	55	9
CCG 35		A,GCC,GAA,GGT,CAG,A	*T. thermophilus* C025957	39	52	14
TCG 36		A,GCC,GAA,GGT,C	*P. furiosus* At1842	40	50	11
CCT 31		A,TTC,CTG,CAG,GG	*E. coli* C08004487	57	68	12
TCT 35		A,ATC,CTG,CAG,GGC,GC	*E. coli* C08004473	59	73	15

Table 4.8: Identification of X circular code motifs asn-tRNAXm in tRNAs of Asparagine (Asn) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
GTT 34	GCC,GCC,GTA,GC		*P. furiosus* At1856	1	11	11
GTT 34	T,GTA,GTT,CAG,T		*E. coli* C08004483	6	16	11
GTT 34	T,GTA,GTT,CAG,T		*E. coli* C08004537	6	16	11
GTT 34	T,CAG,CAG,GTA,GAG,CAG,C		*T. thermophilus* C025962	12	28	17
GTT 34	AG,AAC,GGC,GG		*E. coli* C08004483	21	30	10
GTT 34	AG,AAC,GGC,GG		*E. coli* C08004537	21	30	10
GTT 34		T A,ACC,GGT,A	*T. thermophilus* C025962	36	44	9
GTT 34		CC,GGC,GGT,C	*P. furiosus* At1856	40	48	9
GTT 34		G,GGC,GGC,GGC,GCC	*P. furiosus* At1856	63	75	13

Table 4.9: Identification of X circular code motifs asp-tRNAXm in tRNAs of Aspartic (Asp) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
GTC 37	G,GGT,GGT,GTA,GCC,C		*P. furiosus* At1869	5	18	14
GTC 35	GT,GGT,GTA,GTT,GGT,TA		*T. thermophilus* C025932	7	22	16
GTC 35	G,GTA,GTT,CAG,T		*E. coli* C08004470	6	16	11
GTC 35	A,GTC,GGT,TA		*E. coli* C08004470	14	22	9
GTC 35	TG,GTT,AAC,AC		*T. thermophilus* C025932	17	26	10
GTC 35	TA,GAA,TAC,CTG,C		*E. coli* C08004470	21	32	12
GTC 35	A,GAA,TAC,CT		*E. coli* C08004470	22	30	9
GTC 35		G,GAG,ATC,GC	*T. thermophilus* C025932	43	51	9

Table 4.10: Identification of X circular code motifs cys-tRNAXm in tRNAs of Cysteine (Cys) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
GCA 33	GGC,GCC,GTA,GCC,AA		*T. thermophilus* C025925	1	14	14
GCA 33	GC,GTT,AAC,AA		*E. coli* C08004540	4	13	10
GCA 33	TA,GAG,GCC,A		*P. furiosus* At1859	13	21	9
GCA 33	A,GGC,CAG,GC		*P. furiosus* At1859	16	24	9
GCA 33	A,GGT,CT G,CA		*T. thermophilus* C025925	27	35	9
GCA 33	A,CT G,CA G,ATC,C		*P. furiosus* At1859	30	40	11
GCA 33	A A,AAC,CTC,CA		*T. thermophilus* C025925	35	44	10
GCA 33		A,TTC,GCC,GGT,T	*T. thermophilus* C025925	44	54	11
GCA 33		G,GCC,GGC,GCC,T	*T. thermophilus* C025925	62	72	11

Table 4.11: Identification of X circular code motifs gln-tRNAXm in tRNAs of Glutamine (Gln) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
TTG 33	G,GGT,ATC,GCC,AA		*E. coli* C08004549	3	14	12
CTG 33	G,GGT,ATC,GCC,AA		*E. coli* C08004552	3	14	12
CTG 33	G,GGT,GTC,GTC,TA		*T. thermophilus* C025940	3	14	12
TTG 33	G,GGC,GTC,GTC,TA		*T. thermophilus* C025948	3	14	12
TTG 35	GT,GGT,GTA,GC		*P. furiosus* At1827	7	16	10
CTG 35	GT,GGT,GTA,GC		*P. furiosus* At1849	7	16	10
TTG 33	AA,GGC,ACC,GGT,TT		*E. coli* C08004549	20	32	13
CTG 33	AA,GGC,ACC,GG		*E. coli* C08004552	20	29	10
CTG 33	G,ATT, CTG, ATT,C		*E. coli* C08004552	29	39	11
TTG 33	TT,G AT,ACC,GGC,ATT,C		*E. coli* C08004549	33	47	15
CTG 33	CC,GGC,ATT,C		*E. coli* C08004552	39	47	9
TTG 33	CC,GCC,GGT,GGT,GGT,T		*T. thermophilus* C025948	39	53	15
CTG 33	CC,GCC,GGT,C		*T. thermophilus* C025940	39	47	9
TTG 33		CC,CTG,GTT,C	*E. coli* C08004549	47	55	9
CTG 33		CC,GAG,GTT,C	*E. coli* C08004552	47	55	9
TTG 33		A,ATC,CAG,GTA,C	*E. coli* C08004549	57	67	11
CTG 33		A,ATC,CTC,GTA,C	*E. coli* C08004552	57	67	11

Table 4.12: Identification of X circular code motifs glu-tRNAXm in tRNAs of Glutamic acid (Glu) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
CTC 35	CC,GGT,GGT,GTA,GCC,C		*P. furiosus* At1830	4	18	15
TTC 37	CC,GGT,GGT,GTA,GCC,C		*P. furiosus* At1847	4	18	15
TTC 35	CC,TTC,GTC,TA		*E. coli* C08004497	5	14	10
TTC 35	CC,TTC,GTC,TA		*E. coli* C08004534	5	14	10
TTC 34	CC,ATC,GAC,TA		*T. thermophilus* C025941	5	14	10
CTC 32	CC,ATC,GTC,TA		*T. thermophilus* C025923	5	14	10
CTC 33	CC,ATC,GTC,TA		*T. thermophilus* C025942	5	14	10
TTC 35	TA,GAG,GCC,CAG,GAC,ACC,GCC,CT		*E. coli* C08004497	13	34	22
TTC 35	TA,GAG,GCC,CAG,GAC,ACC,GCC,CT		*E. coli* C08004534	13	34	22
CTC 33	TA,GAG,GCC,TA		*T. thermophilus* C025942	13	22	10
TTC 34	AG,GTC,ACC,GGC,CT		*T. thermophilus* C025941	21	33	13
TTC 34		AA,GCC,GGC,GGC,GG	*T. thermophilus* C025941	37	49	13
TTC 35		AC,GGC,GGT,AAC,A	*E. coli* C08004497	38	49	12
TTC 35		AC,GGC,GGT,AAC,A	*E. coli* C08004534	38	49	12
CTC 32		AG,GCC,GAA,AC	*T. thermophilus* C025923	38	47	10
CTC 33		AG,GCC,GAG,AC	*T. thermophilus* C025942	39	48	10

4.2.3.2 X CIRCULAR CODE MOTIFS IN ARG-TRNAS (TABLE 4.7)

(i) 5' region of Arg-tRNAs: The X motif CC,GGT,GGC,CT is found. The large X motif $Arg-tRNAXm_1$ $CC, GTA, GTT, CAG, CTG, GAT, A$ of 18 nucleotides is identified in *E. coli*. The class of X motifs S, GTA, GYY, TA is prefix of $Arg-tRNAXm_1$, and the class of X

motifs T, CAG, CWG, GAT, A and $Arg-tRNAXm_2$ RS, CWG, GAT, R are suffix of $Arg-tRNAXm_1$. The X motif AT, AAC, GAG, C and $Arg-tRNAXm_2$ are shifted in frame. The X motif TA, GAG, TAC, T is observed. The class of X motifs G, GGC, GYC, GGC, CT and the X motif GC, AAC, GAC, CT occur before the anti-codons TCG and TCT.

(ii) anti-codon regions of Arg-tRNAs: The large X motif $Arg-tRNAXm_3$ $A, GAG, AAC, GCC, GCC, CTC, C$ of 17 nucleotides in *P. furiosus* is in a different frame than the anti-codon CCG. The X motif GC, GTC, GGC, CTC, C which is suffix of $Arg-tRNAXm_3$ is in a different frame than the anti-codon CCG. The large X motif $Arg-tRNAXm_4$ $AG, GGC, GGC, GGC, CTC, CT$ of 16 nucleotides in *P. furiosus* is in a different frame than the anti-codon CCT. The X motif G, GGC, TTC, CT which is suffix of $Arg-tRNAXm_4$ is in a different frame than the anti-codon CCT. The X motif $Arg-tRNAXm_5$ A, GCC, GCC, TCT, TA is in a different frame than the anti-codon TCT. The class of X motifs G, RCC, TCT, TA which is suffix of $Arg-tRNAXm_5$ is in a different frame than the anti-codon TCT. The X motif G, GCC, TTC, GG is in a different frame than the anti-codon TCG. The large X motif $Arg-tRNAXm_6$ $A, GTA, CTC, GGC, TAC, GAA, C$ of 17 nucleotides in *E. coli* is in a different frame than the anti-codon ACG. The X motif T, GAC, TAC, GG is in a different frame than the anti-codon ACG. The class of X motifs G, GAY, CWS, SAG, GTC, S is in frame with the anti-codons ACG and GCG.

(iii) 3' region of Arg-tRNAs: The X motifs A, GCC, GAA, GGT, CAG, A and the class of X motifs A, WTC, CTG, CAG, GG are observed.

4.2.3.3 X CIRCULAR CODE MOTIFS IN ASN-TRNAS (TABLE 4.8)

(i) 5' region of Asn-tRNAs: The X motif $Asn-tRNAXm_1$ GCC, GCC, GTA, GC is observed. The X motif $Asn-tRNAXm_2$ T, GTA, GTT, CAG, T and $Asn-tRNAXm_1$ are shifted in frame. The large X motif $Asn-tRNAXm_3$ $T, CAG, CAG, GT\underline{A}, \underline{GA}G, \underline{C}A\underline{G}, \underline{C}$ of 17 nucleotides in *T. thermophilus* and $Asn-tRNAXm_2$ are shifted in frame. The X motif $\underline{AG}, \underline{A}AC, G\underline{GC}, GG$ is shifted by +2 nucleotides from $Asn-tRNAXm_3$ (underlined nucleotides).

(ii) anti-codon regions of Asn-tRNAs: The X motif $Asn - tRNAXm_4$ TA, ACC, GGT, A is in a different frame than the anti-codon GTT.

(iii) 3' region of Asn-tRNAs: The X motif CC,GGC,GGT,C and $Asn - tRNAXm_4$ are shifted in frame. The X motif $G, GGC, GGC, GGC, GCCisobserved.$

4.2.3.4 X circular code motifs in Asp-tRNAs (Table 4.9)

(i) 5' region of Asp-tRNAs: The X motif $Asp - tRNAXm_1$ G, GGT, GGT, GTA, GCC, C is observed. The large X motif $Asp - tRNAXm_2$ $GT, GGT, GTA, GTT, GGT, TA$ of 16 nucleotides in *T. thermophilus* and $Asp - tRNAXm_1$ are shifted in frame. The X motif $Asp - tRNAXm_3$ $G, GTA, GTT, C\underline{AG, T}$ is suffix of $Asp - tRNAXm_2$. The X motif $Asp - tRNAXm_4$ $\underline{A, GT}C, \underline{GGT, TA}$ is shifted by $+1$ nucleotide from $Asp - tRNAXm_3$. The X motif $T\underline{G, GTT, A}AC, AC$ is shifted by $+2$ nucleotides from $Asp - tRNAXm_4$. The X motif TA, GAA, TAC, CTG, C is observed.

(ii) 3' region of Asp-tRNAs: The X motif G, GAG, ATC, GC is observed.

4.2.3.5 X circular code motifs in Cys-tRNAs (Table 4.10)

(i) 5' region of Cys-tRNAs: The X motif $Cys - tRNAXm_1$ GGC, GCC, GTA, GCC, AA is observed. The X motif GC, GTT, AAC, AA is suffix of $Cys - tRNAXm_1$. The X motif $\underline{A, GGC, CA}G, GC$ is shifted by $+1$ nucleotide from $TA, G\underline{AG, GCC, A}$.

(ii) anti-codon regions of Cys-tRNAs: The X motif $Cys - tRNAXm_2$ A, GGT, CTG, CA is in a different frame than the anti-codon GCA. The X motif $Cys - tRNAXm_3$ A, CTG, CAG, ATC, C and $Cys - tRNAXm_2$ are shifted in frame, thus $Cys - tRNAXm_3$ is in a different frame than the anti-codon GCA. The X motifs AA, AAC, CTC, CA and $Cys - tRNAXm_3$ are shifted in frame, thus they are in a different frame than the anti-codon GCA.

(iii) 3' region of Cys-tRNAs: Two X motifs A, TTC, GCC, GGT, T and G, GCC, GGC, GCC, T are observed.

4.2.3.6 X circular code motifs in Gln-tRNAs (Table 4.11)

(i) 5' region of Gln-tRNAs: The class of X motifs $Gln-tRNAXm_1$ G, GGY, RTC, GYC, WA is identified. The X motif $\underline{GT}, G\underline{GT}, G\underline{TA}, GC$ is shifted by $+1$ nucleotide from the class of X motifs $G, GGY, \underline{GT}C, \underline{GT}C, \underline{TA}$ belonging to $Gln-tRNAXm_1$. The X motif AA, GGC, ACC, GGT, TT occurs before the anti-codon.

(ii) anti-codon regions of Gln-tRNAs: The X motif G, ATT, CTG, ATT, C is in frame with the anti-codon CTG. The X motif $Gln-tRNAXm_2$ $TT, GAT, ACC, GGC, ATT, C$ is in a different frame than the anti-codon TTG.

(iii) 3' region of Gln-tRNAs: The X motif $CC, GCC, GGT, GGT, GGT, T$ and $Gln-tRNAXm_2$ are shifted in frame. Two classes of X motifs CC, SWG, GTT, C and A, ATC, CWS, GTA, C are observed.

4.2.3.7 X circular code motifs in Glu-tRNAs (Table 4.12)

(i) 5' region of Glu-tRNAs: The X motif $Glu-tRNAXm_1$ $C\underline{C}, GG\underline{T}, G\underline{GT}, G\underline{TA}, GCC, C$ is observed. The class of X motifs $\underline{C}C, W\underline{T}C, \underline{G}WC, \underline{TA}$ is shifted by $+2$ nucleotides from $Glu-tRNAXm_1$. A very large X motif $Glu-tRNAXm_2$ $TA, GAG, GCC, CAG, GAC, ACC, GCC, CT$ of 22 nucleotide in *E. coli* and $Glu-tRNAXm_1$ are shifted in frame. The X motif TA, GAG, GCC, TA is prefix of $Glu-tRNAXm_2$ and the X motif AG, GTC, ACC, GGC, CT is suffix of $Glu-tRNAXm_2$.

(ii) 3' region of Glu-tRNAs: The X motifs AA, GCC, GGC, GGC, GG and $Glu-tRNAXm_3$ $AC, \underline{GGC}, G\underline{GT}, A\underline{AC}, A$ occur after (3') the anti-codon TTC. The class of X motifs $A\underline{G, GCC}, \underline{G}AR, \underline{AC}$ is shifted by $+2$ nucleotides from $Glu-tRNAXm_3$.

4.2.3.8 X circular code motifs in Gly-tRNAs (Table 4.13)

(i) 5' region of Gly-tRNAs: Several classes of X motifs are identified: GC, GGT, GGT, A, $Gly-tRNAXm_1$ G, GGC, RTM, GTW, YA, TG, GTA, GTC, TA suffix of $Gly-tRNAXm_1$, WY, AAT, GGY, W, GC, CTG, GTC, TA, G, GTA, GAR, C and AK, WAC, SWS, A. The

Table 4.13: Identification of X circular code motifs gly-tRNAXm in tRNAs of Glycine (Gly) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
CCC 37	GC,GGT,GGT,A		*P. furiosus* At1837	1	9	9
TCC 37	GC,GGT,GGT,A		*P. furiosus* At1844	1	9	9
GCC 37	GC,GGT,GGT,A		*P. furiosus* At1851	1	9	9
TCC 34	G,GGC,ATC,GTA,TA		*E. coli* C08004509	3	14	12
CCC 33	G,GGC,GTA,GTT,CA		*E. coli* C08004527	3	14	12
CCC 37	TG,GTA,GTC,TA		*P. furiosus* At1837	5	14	10
TCC 37	TG,GTA,GTC,TA		*P. furiosus* At1844	5	14	10
GCC 37	TG,GTA,GTC,TA		*P. furiosus* At1851	5	14	10
CCC 33	AG,TTC,AAT,GGT,A		*E. coli* C08004527	9	20	12
TCC 34	AT,AAT,GGC,TA		*E. coli* C08004509	12	21	10
CCC 37	GC,CTG,GTC,TA		*P. furiosus* At1837	15	24	10
TCC 37	GC,CTG,GTC,TA		*P. furiosus* At1844	15	24	10
GCC 37	GC,CTG,GTC,TA		*P. furiosus* At1851	15	24	10
CCC 33	TG,GTA,GAA,C		*E. coli* C08004527	16	24	9
GCC 34	TG,GTA,GAG,CA		*E. coli* C08004512	17	26	10
GCC 34	TG,GTA,GAG,CA		*E. coli* C08004539	17	26	10
CCC 34	TG,GTA,GAG,CA		*T. thermophilus* C025960	17	26	10
TCC 35	TG,GTA,GAG,CA		*T. thermophilus* C025952	18	27	10
GCC 34	G,GTA,GAG,CA		*T. thermophilus* C025918	18	26	9
CCC 33	AG,AAC,GAG,A		*E. coli* C08004527	20	28	9
TCC 34	AT,TAC,CTC,A		*E. coli* C08004509	21	29	9
TCC 37	AG,GAC,GCC,GGC,CT		*P. furiosus* At1844	24	36	13
TCC 35	GC,ATC,GGC,CT		*T. thermophilus* C025952	25	34	10
CCC 37	AG,GAC,GCC,GGC,CT	C,C	*P. furiosus* At1837	24	38	15
CCC 34	GC,ATC,GGC,TT	C,C	*T. thermophilus* C025960	24	35	12
GCC 37	AG,GAC,GCC,ACC,CT	G,C	*P. furiosus* At1851	24	38	15
TCC 34	A,GCC,T	TC,C A	*E. coli* C08004509	29	37	9
TCC 35	G,GCC,T	TC,C A	*T. thermophilus* C025952	30	38	9
CCC 34		AA,GCC,GAG,GGT,C	*T. thermophilus* C025960	37	48	12
TCC 35		AA,GCC,GAG,GGT,C	*T. thermophilus* C025952	38	49	12
CCC 37		AA,GCC,GGC,GAC,C	*P. furiosus* At1837	40	51	12
TCC 37		A,GCC,GGC,GAC,C	*P. furiosus* At1844	41	51	11
TCC 34		T,GAT,GAT,GC	*E. coli* C08004509	41	49	9
CCC 33		TA,TAC,GAG,GGT,T	*E. coli* C08004527	42	53	12
GCC 37		TG,GAG,ACC,C	*P. furiosus* At1851	44	52	9
GCC 34		GC,GAG,TTC,GAG,T	*E. coli* C08004512	49	60	12
GCC 34		GC,GAG,TTC,GAG,T	*E. coli* C08004539	49	60	12
GCC 34		GT,CTC,GTT,T	*E. coli* C08004512	59	67	9
GCC 34		GT,CTC,GTT,T	*E. coli* C08004539	59	67	9
CCC 33		CC,TTC,GCC,C	*E. coli* C08004527	60	68	9
CCC 37		G,GCC,ACC,GC	*P. furiosus* At1837	66	74	9
TCC 37		G,GCC,ACC,GC	*P. furiosus* At1844	66	74	9
GCC 37		G,GCC,ACC,GC	*P. furiosus* At1851	66	74	9

Table 4.14: Identification of X circular code motifs his-tRNAXm in tRNAs of Histidine (His) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
GTG 36	G,GGT,GGT,GTA,GCC,T		*P. furiosus* At1843	5	18	14
GTG 35	T,CAG,CTG,GTT,A		*T. thermophilus* C025927	12	22	11
GTG 36	GC,CTG,GTT,A		*P. furiosus* At1843	15	23	9
GTG 34	TG,GTA,GAG,C		*E. coli* C08004501	17	25	9
GTG 34	A,GCC,CTG,GAT,T		*E. coli* C08004501	23	33	11
GTG 34		A,TTC,CAG,TT	*E. coli* C08004501	37	45	9
GTG 34		A,GTT,GTC,GT	*E. coli* C08004501	42	50	9
GTG 36		CC,CTG,GCC,C	*P. furiosus* At1843	43	51	9

two X motifs $Gly - tRNAXm_2$ AG, GAC, GCC, GGC, CT and $Gly - tRNAXm_3$ GC, ATC, GGC, CT occur before the anti-codon TCC.

(ii) anti-codon regions of Gly-tRNAs: The two X motifs $AG, GAC, GCC, GGC, CTC, C$ whose prefix is $Gly - tRNAXm_2$ and GC, ATC, GGC, TTC, C whose prefix is $Gly - tRNAXm_3$ are in a different frame than the anti-codon CCC. The X motif

Table 4.15: Identification of X circular code motifs ile-tRNAXm in tRNAs of Isoleucine (Ile) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
GAT 35	G,GGC,GAT,TA		*T. thermophilus* C025963	1	9	9
GAT 35	T,CAG,CTG,GTT,A		*T. thermophilus* C025963	12	22	11
GAT 36	GC,CTG,GTC,A		*P. furiosus* At1831	15	23	9
GAT 35	A,GGT,GGT,TA		*E. coli* C08004468	14	22	9
GAT 35	A,GGT,GGT,TA		*E. coli* C08004521	14	22	9
GAT 35		AG,GGT,GAG,GTC,GGT,GGT,T	*E. coli* C08004468	39	56	18
GAT 35		AG,GGT,GAG,GTC,GGT,GGT,T	*E. coli* C08004521	39	56	18
GAT 35		GT,GAG,GTC,GGT,GGT,T	*T. thermophilus* C025963	42	56	15
GAT 36		G,TTC,GAA,GCC,C	*P. furiosus* At1831	55	65	11
GAT 35		CC,ACC,ATC,GCC,CA	*T. thermophilus* C025963	62	74	13
GAT 35		T,CAG,GCC,TAC	*E. coli* C08004468	66	75	10

Table 4.16: Identification of X circular code motifs leu-tRNAXm in tRNAs of Leucine (Leu) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
CAG 35	GC,GAA,GGT,GGC,GG		*E. coli* C08004502	1	13	13
CAG 35	GC,GAA,GGT,GGC,GG		*E. coli* C08004516	1	13	13
CAG 35	GC,GAA,GGT,GGC,GG		*E. coli* C08004517	1	13	13
TAG 35	GC,GAG,GAT,GGC,GG		*T. thermophilus* C025928	1	13	13
TAA 34	G,GGT,GGC,GG		*T. thermophilus* C025929	5	13	9
CAG 35	G,GGT,GGC,GG		*T. thermophilus* C025947	5	13	9
CAA 35	G,GGT,GGC,GG		*T. thermophilus* C025954	5	13	9
GAG 35	A,GGT,GGT,GG		*E. coli* C08004524	5	13	9
GAG 35	G,GGT,GGT,GG		*T. thermophilus* C025921	5	13	9
TAA 35	G,GAT,GGT,GG		*E. coli* C08004541	5	13	9
CAG 37	G,GTT,GCC,GAG,C		*P. furiosus* At1867	6	16	11
TAG 37	G,GTT,GCC,GAG,C		*P. furiosus* At1838	6	16	11
CAA 37	G,GTT,GCC,GAG,C		*P. furiosus* At1848	6	16	11
GAG 37	G,GTT,GCC,GAG,C		*P. furiosus* At1853	6	16	11
CAA 35	GT,GGC,GAA,ATC,GGT,A		*E. coli* C08004515	7	21	15
TAG 35	GT,GGC,GAA,ATT,GGT,A		*E. coli* C08004548	7	21	15
GAG 35	TG,GAA,CTG,GTA,GAC,AC		*T. thermophilus* C025921	11	26	16
CAG 37	GC,CTG,GTC,AA		*P. furiosus* At1867	15	24	10
TAG 37	GC,CTG,GTC,AA		*P. furiosus* At1838	15	24	10
CAA 37	GC,CTG,GTC,AA		*P. furiosus* At1848	15	24	10
GAG 37	GC,CTG,GTC,AA		*P. furiosus* At1853	15	24	10
TAA 37	GC,CTG,GCC,AA		*P. furiosus* At1862	15	24	10
CAG 35	TG,GTA,GAC,GC		*E. coli* C08004502	17	26	10
CAG 35	TG,GTA,GAC,GC		*E. coli* C08004516	17	26	10
CAG 35	TG,GTA,GAC,GC		*E. coli* C08004517	17	26	10
GAG 35	TG,GTA,GAC,AC		*E. coli* C08004524	17	26	10
TAG 35	TG,GTA,GAC,GC		*E. coli* C08004548	17	26	10
CAG 35	TG,GTA,GAC,GC		*T. thermophilus* C025947	17	26	10
TAA 34	G,GTA,GAC,GC		*T. thermophilus* C025929	17	25	9
CAA 35	G,GTA,GAC,GC		*E. coli* C08004515	18	26	9
TAA 35	G,GTA,GAC,AC		*E. coli* C08004541	18	26	9
TAG 35	G,GTA,GAC,GC		*T. thermophilus* C025928	18	26	9
CAA 35	G,GTA,GAC,GC		*T. thermophilus* C025954	18	26	9
GAG 35	AC,GCC,ATC,TT		*T. thermophilus* C025921	25	34	10
TAG 35	AC,CAG,ATT,TA		*E. coli* C08004548	27	36	10
CAG 35	TG,ATT,CAG,GGT,CA		*T. thermophilus* C025947	30	42	13
CAG 37	G,ATT,CAG,GGT,C		*P. furiosus* At1867	33	43	11
GAG 37	G,ATT,GAG,GGT,C		*P. furiosus* At1853	33	43	11
TAG 35	AG,GTT,CTG,GC		*E. coli* C08004548	36	45	10
CAA 35	AA,AAT,CTG,CTG,T		*T. thermophilus* C025954	36	47	12
CAA 35	AA,ATC,AAC,C		*E. coli* C08004515	37	45	9
CAA 35	A,ACC,GTA,GAA,AT		*E. coli* C08004515	42	53	12
TAA 35	CC,CTC,GGC,GTT,C		*E. coli* C08004541	41	52	12
TAG 35	T,GGC,GCC,GC		*E. coli* C08004548	42	50	9
GAG 35	G,GGT,GGT,GCC,C		*T. thermophilus* C025921	39	49	11
TAA 37	CC,GGT,GCC,GTA,GG		*P. furiosus* At1862	43	55	13
CAA 35	GT,GCC,GGT,T		*E. coli* C08004515	56	64	9
TAG 35	GC,GAG,TTC,AA		*E. coli* C08004548	58	67	10
TAG 35	GT,CTC,GCC,T		*E. coli* C08004548	68	76	9
CAA 35	G,GCC,TTC,GGC,ACC		*E. coli* C08004515	72	84	13
TAG 35	G,CTC,CTC,GC		*T. thermophilus* C025928	73	81	9

$AG, GAC, GCC, ACC, CTG, C$ is in a different frame than the anti-codon GCC. The class of X motifs R, GCC, TTC, CA is in a different frame than the anti-codon TCC.

(iii) 3' region of Gly-tRNAs: The class of X motifs AA, GCC, GRS, GRY, C oc-

Table 4.17: Identification of X circular code motifs lys-tRNAXm in tRNAs of Lysine (Lys) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
CTT 34	AG,CTC,AAC,T		*T. thermophilus* C025946	9	17	9
TTT 33	T,CAG,CTG,GC		*T. thermophilus* C025922	12	20	9
TTT 35	GC,CTG,GTT,A		*P. furiosus* At1828	15	23	9
CTT 36	GC,CTG,GTT,A		*P. furiosus* At1858	15	23	9
CTT 34	AA,CTG,GTA,GAG,CA		*T. thermophilus* C025946	14	26	13
TTT 33	TG,GTA,GAG,CAG,TT		*E. coli* C08004474	17	29	13
TTT 33	A,GTT,GAC, TT		*E. coli* C08004474	26	34	9
TTT 33	A,ACC,GAC, TT		*T. thermophilus* C025922	26	34	9
CTT 34	T A,ATC,GGT,GG		*T. thermophilus* C025946	36	45	10
TTT 33	TA,ATC,GGT,A		*T. thermophilus* C025922	36	44	9
TTT 33	TA,ATC,AAT,T		*E. coli* C08004474	36	44	9
CTT 34	TG,GGT,TAC,A		*T. thermophilus* C025946	43	51	9
TTT 35	CC,GGT,GGT,C		*P. furiosus* At1828	42	50	9
CTT 34	TA,CAG,GTT,C		*T. thermophilus* C025946	48	56	9
CTT 36	G,CAG,GTC,GG		*P. furiosus* At1858	44	52	9
TTT 33	A,ATC,CTG,CA		*E. coli* C08004474	58	66	9
TTT 33	A,ATC,CTG,CA		*T. thermophilus* C025922	58	66	9

Table 4.18: Identification of X circular code motifs met-tRNAXm in tRNAs of Methionine (Met) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
CAT 35	GGC,TAC,GTA,GC		*E. coli* C08004547	1	11	11
CAT 35	GGC,GGC,GTA,GC		*T. thermophilus* C025934	1	11	11
CAT 35	TG,GAG,CAG,C		*E. coli* C08004492	8	16	9
CAT 35	TG,GAG,CAG,C		*E. coli* C08004525	8	16	9
CAT 35	TG,GAG,CAG,C		*T. thermophilus* C025937	8	16	9
CAT 34	AG,CTC,AAC,GGT,CAG,A		*T. thermophilus* C025955	9	23	15
CAT 35	A,GGT,GGT,CAG,A		*T. thermophilus* C025934	14	24	11
CAT 35	A,GTT,GGT,TA		*E. coli* C08004547	14	22	9
CAT 36	GC,CTG,GTC,AA		*P. furiosus* At1857	15	24	10
CAT 36	GC,CTG,GTC,A		*P. furiosus* At1855	15	23	9
CAT 35	GC,CTG,GTA,GC		*E. coli* C08004492	15	24	10
CAT 35	GC,CTG,GTA,GC		*E. coli* C08004525	15	24	10
CAT 35	GC,CTG,GTA,GC		*T. thermophilus* C025937	15	24	10
CAT 34	TA,GAG,CAG,GC		*E. coli* C08004495	20	29	10
CAT 34	TA,GAG,CAG,GC		*E. coli* C08004533	20	29	10
CAT 35	AG,CTC,GTC,GG		*E. coli* C08004492	22	31	10
CAT 35	AG,CTC,GTC,GG		*E. coli* C08004525	22	31	10
CAT 35	AG,CTC,GTC,GG		*T. thermophilus* C025937	22	31	10
CAT 35	AT, AAT,GAT,GG		*E. coli* C08004547	36	45	10
CAT 34	T A,ACC,GGT,A		*T. thermophilus* C025955	36	44	9
CAT 35	CC,GAA,GGT,C		*E. coli* C08004492	41	49	9
CAT 35	CC,GAA,GGT,C		*T. thermophilus* C025937	41	49	9
CAT 35	CC,GAA,GAT,C		*E. coli* C08004525	41	49	9
CAT 35	GT,GGT,GTC,GT		*T. thermophilus* C025934	42	51	10
CAT 35	AG,GTC,GTC,GGT,T		*E. coli* C08004492	45	56	12
CAT 35	AG,ATC,GTC,GGT,T		*E. coli* C08004525	45	56	12
CAT 34	TG,CAG,GTT,C		*T. thermophilus* C025955	48	56	9
CAT 36	CC,GAG,GTT,CA		*P. furiosus* At1846	50	59	10
CAT 34	G,CTG,GTT,CA		*E. coli* C08004495	49	57	9
CAT 34	G,CTG,GTT,CA		*E. coli* C08004533	49	57	9
CAT 36	G,TTC,GAA,GCC,C		*P. furiosus* At1857	55	65	11
CAT 34	AA,GTC,CAG,CAG,GG		*E. coli* C08004495	57	69	13
CAT 34	AA,GTC,CAG,CA		*E. coli* C08004533	57	66	10
CAT 36	AA,ATC,CTC,GGC,C		*P. furiosus* At1846	59	70	12
CAT 34	A,ATC,CTG,CA		*T. thermophilus* C025955	58	66	9
CAT 35	CC,GTC,GTA,GCC		*E. coli* C08004547	63	73	11
CAT 35	CC,ACC,GCC,GCC,ACC		*T. thermophilus* C025934	63	76	14

curs after the anti-codons CCC and TCC. Several X motifs are also found: TA, TAC, GAG, GGT, T, TG, GAG, ACC, C, GC, GAG, TTC, GAG, T, GT, CTC, GTT, T, CC, TTC, GCC, C and G, GCC, ACC, GC where no obvious relation could have been identified between them so far.

Table 4.19: Identification of X circular code motifs phe-tRNAXm in tRNAs of Phenylalanine (Phe) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
GAA 34	GCC,GAG,GTA,GC		*T. thermophilus* C025933	1	11	11
GAA 34	TG,GTA,GAG,CA		*T. thermophilus* C025933	17	26	10
GAA 34	G,GTA,GAG,CAG,GG		*E. coli* C08004519	18	29	12
GAA 34	G,ATT, GAA, AAT,C		*E. coli* C08004519	30	40	11
GAA 35	G,GGT,GTC,GG		*P. furiosus* At1845	43	51	9
GAA 34	GT,GTC,GGC,GGT,T		*T. thermophilus* C025933	44	55	12
GAA 34	CC,CTC,GGC,ACC		*T. thermophilus* C025933	65	75	11

4.2.3.9 X CIRCULAR CODE MOTIFS IN HIS-TRNAS (TABLE 4.14)

(i) 5' region of His-tRNAs: Several classes of X motifs are identified: G, GGT, GGT, GTA, GCC, T, $His - tRNAXm_1$ RS, CTG, GTT, A, TG, GTA, GAG, C shifted in frame with $His - tRNAXm_1$ and A, GCC, CTG, GAT, T.

(ii) 3' region of His-tRNAs: The X motif $His - tRNAXm_2$ $\underline{A, GTT}, GTC, GT$ is shifted by $+1$ nucleotide from $A, TTC, C\underline{AG, TT}$. The X motif CC, CTG, GCC, C is shifted in frame with $His - tRNAXm_2$.

4.2.3.10 X CIRCULAR CODE MOTIFS IN ILE-TRNAS (TABLE 4.15)

(i) 5' region of Ile-tRNAs: Several X motifs are identified: G, GGC, GAT, TA and $Ile - tRNAXm_1$ $RS, \underline{CTG, GTY}, A$. The X motif $A, GG\underline{T, GGT}, T\underline{A}$ is shifted by $+1$ nucleotide from $Ile - tRNAXm_1$.

(ii) 3' region of Ile-tRNAs: A large X motif $Ile - tRNAXm_2$ $AG, GGT, GAG, GTC, GGT, G\underline{GT, T}$ of 18 nucleotides is identified in *E. coli*. A suffix of $Ile - tRNAXm_2$ of 15 nucleotides is found in *T. thermophilus*. The X motif $\underline{G, TTC}, GAA, GCC, C$ is shifted by $+1$ nucleotide from $Ile - tRNAXm_2$. The X motifs are observed: CC, AAC, ATC, GCC, CA and T, CAG, GCC, TAC.

4.2.3.11 X CIRCULAR CODE MOTIFS IN LEU-TRNAS (TABLE 4.16)

(i) 5' region of Leu-tRNAs: Several classes of X motifs are shifted in frame, in series: GC, GAR, GRT, GGC, GG, R, GRT, GGY, GG, G, GTT, GCC, GAG, C, $GT, GGC, GAA, ATY, GGT, A$, a large X motif $Leu - tRNAXm_1$ $TG, GAA, CTG, GTA, GAC, AC$ of 16 nucleotides in *T. thermophilus*, GC, CTG, GYC, AA, G, GTA, GAC, RC and AC, GCC, ATC, TT which occurs before the anti-codon GAG.

(ii) anti-codon regions of Leu-tRNAs: The X motif $Leu - tRNAXm_2$ AC, CAG, ATT, TA is in frame with the anti-codon TAG. The class of X motifs G, ATT, SAG, GGT, C in frame with $Leu - tRNAXm_2$ except with its suffix TA of the anti-codon TAG, is in frame with both the anti-codons CAG and GAG. The X motifs AG, GTT, CTG, GC, AA, AAT, CTG, CTG, T and $Leu - tRNAXm_3$ $AA, ATC, \underline{AAC, C}$ are in frame with the anti-codons TAG, CAA and CAA, respectively.

(iii) 3' region of Leu-tRNAs: The X motif $Leu - tRNAXm_4$ $\underline{A, ACC}, GTA, GAA, AT$ is shifted by $+2$ nucleotides from $Leu - tRNAXm_3$. Several classes of X motifs are shifted in frame: $Leu - tRNAXm_4$, CC, CTC, GGC, GTT, C, T, GGC, GCC, GC and SY, GGT, GCC, S. The X motif $\underline{GC}, GA\underline{G, TTC}, AA$ is shifted by $+1$ nucleotide from $GT, \underline{GC}C, G\underline{GT, T}$. The X motifs GT, CTC, GCC, T and G, GCC, TTC, GGC, ACC are shifted in frame.

4.2.3.12 X circular code motifs in Lys-tRNAs (Table 4.17)

(i) 5' region of Lys-tRNAs: The X motif $His{-}tRNAXm_1$ $\underline{T, CA}G, \underline{CT}G, GC$ is shifted by $+1$ nucleotide from $AG, \underline{CTC, AA}\underline{C, T}$. The X motifs $His - tRNAXm_1$, GC, CTG, GTT, A, AA, CTG, GTA, GAG, CA and $His - tRNAXm_2$ $TG, GTA, GAG, C\underline{AG, TT}$ are shifted in frame.

(ii) anti-codon regions of Lys-tRNAs: The class of X motifs $\underline{A, RYY}, GAC, TT$ which is in frame with the anti-codon TTT, is shifted by $+1$ nucleotide from $His{-}tRNAXm_2$. The X motif $His{-}tRNAXm_3$ TA, ATC, GGT, GG in *T. thermophilus* C025946 is in frame with the anti-codon CTT.

(iii) 3' region of Lys-tRNAs: Interestingly, the X motifs TA, ATC, GGT, A in *T. thermophilus* C025922 which is identical to $His{-}tRNAXm_3$ (except its last letter) and TA, ATC, AAT, T occur after the anti-codon TTT and are not involved in the anti-codon of Lys- tRNAs. Several classes of X motifs are observed: TG, GGT, TAC, A, CC, GGT, GGT, C, R, CAG, GTY, S and A, ATC, CTG, CA.

4.2.3.13 X circular code motifs in Met-tRNAs (Table 4.18)

(i) 5' region of Met-tRNAs: Several classes of X motifs are shifted in frame: GGC, KRC, GTA, GC, TG, GAG, CAG, C and $Met - tRNAXm_1$

$AG, CTC, AAC, GGT, CAG, A$. The X motif $Met - tRNAXm_2$ A, GGT, GGT, CAG, A is suffix of $Met - tRNAXm_1$. The X motif $Met - tRNAXm_3$ A, GTT, GGT, TA is prefix of $Met - tRNAXm_2$. The class of X motifs $Met - tRNAXm_4$ $GC, \underline{CTG, GTM}, R$ is shifted by $+2$ nucleotides from $Met - tRNAXm_3$ $A, \underline{GTT, GGT}, TA$. The class of X motifs $Met - tRNAXm_4$, TA, GAG, CAG, GC and AG, CTC, GTC, GG are shifted in frame.

(ii) anti-codon regions of Met-tRNAs: The X motif AT, AAT, GAT, GC is in frame with the anti-codon CAT. The X motif TA, ACC, GGT, A is in a different frame than the anti-codon CAT.

(iii) 3' region of Met-tRNAs: Several classes of X motifs are shifted in series: CC, GGA, GRT, C, GT, GGT, GTC, GT, $Met - tRNAXm_5$ $AG, RTC, GTC, \underline{GGT, T}$, $Met - tRNAXm_6$ $S, SW\underline{G, GTT, C}$ shifted by $+2$ nucleotides from $Met - tRNAXm_5$, $\underline{G, TTC}, GAA, GCC, C$ shifted by $+1$ nucleotide from $Met - tRNAXm_6$, AA, GTC, CAG, CAG, GG, A, ATC, CTS, S, CC, GTC, GTA, GCC and CC, ACC, GCC, GCC, ACC.

Table 4.20: Identification of X circular code motifs pro-tRNAXm in tRNAs of Proline (Pro) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
TGG 35	C,GGC,GAG,TA		*E. coli* C08004503	1	9	9
TGG 35	G,GGC,GTA,GC		*T. thermophilus* C025951	3	11	9
CGG 37	G,GCC,GTA,GG		*P. furiosus* At1870	3	11	9
TGG 37	G,GCC,GTA,GG		*P. furiosus* At1864	3	11	9
TGG 35	G,CAG,GCC,GGT,A		*T. thermophilus* C025951	12	22	11
GGG 35	GC,CTG,GTA,GC		*E. coli* C08004486	15	24	10
CGG 35	GC,CTG,GTA,GC		*E. coli* C08004520	15	24	10
GGG 35	GC,CTG,GTA,GC		*T. thermophilus* C025919	15	24	10
CGG 37	CC,ATC,CTG,C		*P. furiosus* At1870	22	30	9
TGG 37	CC,ATC,CTG,C		*P. furiosus* At1864	22	30	9
GGG 35	GC,ACC,GTC,AT		*E. coli* C08004486	25	34	10
CGG 35	AC,TTC,GTT,C		*E. coli* C08004520	27	35	9
CGG 35	AC,CTC,GTT,C		*T. thermophilus* C025931	27	35	9
TGG 35	AA,CTG,GTT,T		*E. coli* C08004503	27	35	9
GGG 36	CC,GGC,CT G,GG		*P. furiosus* At1866	29	38	10
CGG 37	G,GGC,TT C,GG		*P. furiosus* At1870	31	39	9
GGG 35	G,G GT,GTC,GG		*E. coli* C08004486	36	44	9
GGG 36	G,G GC,GCC,GG		*P. furiosus* At1866	37	45	9
CGG 35	G, GAC,GAA,GG		*E. coli* C08004520	37	45	9
CGG 35	G, GAC,GAG,GG		*T. thermophilus* C025931	37	45	9
TGG 35	G, GAG,CAG,GG		*T. thermophilus* C025951	37	45	9
GGG 35	G,GTC,GTC,GGT,T		*T. thermophilus* C025919	46	56	11
GGG 35	G,GAG,GTT,CA		*E. coli* C08004486	50	58	9
CGG 35	G,CTG,GTT,CA		*T. thermophilus* C025931	50	58	9
GGG 35	AA,ATC,CTC,T		*E. coli* C08004486	58	66	9
CGG 35	AA,ATC,CAG,T		*T. thermophilus* C025931	58	66	9
CGG 35	A,ATC,CTC,TA		*E. coli* C08004520	59	67	9
GGG 35	CC,GGC,CTC,C		*T. thermophilus* C025919	62	70	9
GGG 36	CC,GGC,GGC,C		*P. furiosus* At1866	63	71	9
CGG 37	CC,GGC,GGC,C		*P. furiosus* At1870	64	72	9
TGG 37	CC,GGC,GGC,C		*P. furiosus* At1864	64	72	9
TGG 35	T,CTC,GCC,GAC		*E. coli* C08004503	66	75	10

Table 4.21: Identification of X circular code motifs sec-tRNAXm in tRNAs of selenocysteine (Sec) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
TCA 35	AG,ATC,GTC,GTC,T		*E. coli* C08004496	4	15	12
TCA 35	CC,GGT,GAG,GC		*E. coli* C08004496	16	25	10
TCA 35	G,CTG,GAC,T	TC,A A	*E. coli* C08004496	27	38	12
TCA 35		A A,ATC,CAG,TT	*E. coli* C08004496	37	46	10
TCA 35		TG,ATC,TTC,C	*E. coli* C08004496	83	91	9

Table 4.22: Identification of X circular code motifs ser-tRNAXm in tRNAs of Serine (Ser) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
CGA 35	A,GAT,GCC,GG		*E. coli* C08004538	5	13	9
CGA 35	G,GGT,GCC,GG		*T. thermophilus* C025935	5	13	9
CGA 35	G,GTA,GCC,TA		*P. furiosus* At1829	6	14	9
GGA 35	G,GTA,GCC,TA		*P. furiosus* At1839	6	14	9
GCT 35	G,GTA,GCC,TA		*P. furiosus* At1860	6	14	9
GCT 35	TG,GCC,GAG,A		*E. coli* C08004528	8	16	9
TGA 35	TG,GCC,GAG,C		*E. coli* C08004545	8	16	9
TGA 35	TG,GCC,GAG,C		*T. thermophilus* C025944	8	16	9
GCT 35	TG,GCC,GAG,T		*T. thermophilus* C025945	8	16	9
TGA 35	GC,CTG,GTA,GG		*P. furiosus* At1868	15	24	10
GCT 35	GC,CTG,GTA,GG		*P. furiosus* At1860	15	24	10
CGA 35	G,CTG,AAC,GG		*E. coli* C08004538	18	26	9
TGA 35	G,GTT,GAA,GGC,GGC,GGT,CT		*T. thermophilus* C025944	17	34	18
TGA 35	G,GTT,GAA,GGC,ACC,GGT,CT		*E. coli* C08004545	17	34	18
CGA 35	G,GAA,GGC,GC		*P. furiosus* At1829	20	28	9
GGA 35	G,GAA,GGC,GC		*P. furiosus* At1839	20	28	9
GCT 35	T,GAA,GGC,GC		*E. coli* C08004528	20	28	9
GCT 35	TG,GTC,GAA,GGC,GGC,ACC,CT	G,CT	*T. thermophilus* C025945	16	37	22
GCT 35	AG,GGC,GCC,GGC,CT	G,CT	*P. furiosus* At1860	22	37	16
CGA 35	G,ACC,GGT,CT	C,GA A,AAC,C	*E. coli* C08004538	26	42	17
CGA 35	G,GCC,GGT,CT	C,GA A,AAC,C	*T. thermophilus* C025935	26	42	17
CGA 35	A,CT	C,GA G,ATC,C	*P. furiosus* At1829	32	42	11
GGA 35	G,GGC,CT	G,GA G,A	*P. furiosus* At1839	29	39	11
GGA 35	GC,CT	G,GA A,A	*E. coli* C08004544	31	39	9
GGA 35	A,CT	G,GA A,ATC,GT	*T. thermophilus* C025950	32	43	12
TGA 35	T	T,GA A,AAC,C	*E. coli* C08004545	34	42	9
TGA 35	T	T,GA A,AAC,C	*T. thermophilus* C025944	34	42	9
CGA 35		AA,ACC,GGT,A	*T. thermophilus* C025935	38	46	9
GCT 35		AA,GCC,GGT,GG	*P. furiosus* At1860	38	47	10
GCT 35		AA,GGT,GTT,GC	*T. thermophilus* C025945	38	47	10
TGA 35		AA,ACC,GGC,GAC,C	*E. coli* C08004545	38	49	12
GGA 35		GT,GTA,TAC,GGC,AAC,GTA,T	*E. coli* C08004544	42	59	18
CGA 35		TG,GGC,GTT,C	*P. furiosus* At1829	45	53	9
GGA 35		TG,GGC,GTT,C	*P. furiosus* At1839	45	53	9
TGA 35		TG,GGC,GTT,T	*P. furiosus* At1868	45	53	9
GGA 35		G,TTC,GCC,CA	*P. furiosus* At1839	50	58	9
TGA 35		GC,GAA,GCC,CA	*T. thermophilus* C025944	52	61	10
CGA 35		AA,CTC,TAC,C	*E. coli* C08004538	54	62	9
GCT 35		AA,ACC,GGT,GCC,GC	*T. thermophilus* C025945	55	67	13
GGA 35		TA,AAC,CTC,C	*T. thermophilus* C025950	55	63	9
CGA 35		G,GGC,CTC,GC	*T. thermophilus* C025935	58	66	9
TGA 35		A,GAG,TTC,GAA,T	*E. coli* C08004545	62	72	11
TGA 35		CC,CTC,ACC,CTC,C	*T. thermophilus* C025944	76	87	12
GCT 35		CC,GCC,CTC,T	*T. thermophilus* C025945	79	87	9
GGA 35		CC,GCC,CTC,T	*T. thermophilus* C025950	80	88	9

4.2.3.14 X circular code motifs in Phe-tRNAs (Table 4.19)

(i) 5' region of Phe-tRNAs: Two X motifs GCC, GAG, GTA, GC and G, GTA, GAG, CA are observed.

(ii) anti-codon regions of Phe-tRNAs: The X motif G, ATT, GAA, AAT, C is in frame with the anti-codon GAA.

(iii) 3' region of Phe-tRNAs: The X motifs G, GGT, GTC, GG and

Table 4.23: Identification of X circular code motifs thr-tRNAXm in tRNAs of Threonine (Thr) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
GGT 35	CC,GGT,GGC,T		*P. furiosus* At1863	4	12	9
CGT 33	GCC,GGT,GTA,GC		*T. thermophilus* C025936	1	11	11
TGT 35	G,GTA,GCC,TA		*P. furiosus* At1865	6	14	9
TGT 34	AG,CTC,AAC,C		*T. thermophilus* C025958	9	17	9
GGT 34	T,CAG,CAG,GTA,GAG,CA		*T. thermophilus* C025961	12	26	15
GGT 35	GC,CTG,GTA,GAG,C		*P. furiosus* At1863	15	26	12
CGT 34	TG,GTA,GAG,CAG,C		*E. coli* C08004472	17	28	12
GGT 34	TG,GTA,GAG,C		*E. coli* C08004510	17	25	9
GGT 34	TG,GTA,GAG,C		*E. coli* C08004523	17	25	9
TGT 34	AG,GTA,GAG,CA		*E. coli* C08004507	17	26	10
CGT 33	G,GTA,GAG,CA		*T. thermophilus* C025936	17	25	9
CGT 33		GC,CT C,GT A,A	*T. thermophilus* C025936	29	37	9
CGT 34		A,TT C,GT A,AT	*E. coli* C08004472	31	39	9
CGT 35		A,CT C,GT A,ATC,C	*P. furiosus* At1861	32	42	11
TGT 34		T T,GT A,ATC,A	*E. coli* C08004507	33	41	9
TGT 35		T T,GT A,ATC,C	*P. furiosus* At1865	34	42	9
TGT 34		T T,GT A,ATC,GG	*T. thermophilus* C025958	33	42	10
TGT 34		GT, AAT,CAG,TA	*E. coli* C08004507	35	44	10
GGT 34		AG,GGT,GAG,GTC,GCC,GGT,T	*T. thermophilus* C025961	38	55	18
GGT 34		AG,GGT,GAG,GTC,GGC,A	*E. coli* C08004510	38	52	15
GGT 34		AG,GGT,GAG,GTC,C	*E. coli* C08004523	38	49	12
CGT 34		GC,GAA,GGT,C	*E. coli* C08004472	40	48	9
TGT 34		A,GTA,GGT,CA	*E. coli* C08004507	41	49	9
CGT 35		CC,CAG,GTC,C	*P. furiosus* At1861	42	50	9
TGT 35		CC,CAG,GTC,GC	*P. furiosus* At1865	42	51	10
TGT 34		AG,GTC,ACC,A	*E. coli* C08004507	44	52	9
CGT 34		AG,GTC,GTA,GGT,T	*E. coli* C08004472	44	55	12
CGT 33		AG,GCC,GCC,GGT,T	*T. thermophilus* C025936	43	54	12
TGT 34		AC,CAG,TTC,GAT,T	*E. coli* C08004507	49	60	12
GGT 34		CC,CAG,TTC,GAC,T	*E. coli* C08004523	49	60	12
GGT 34		G,CAG,TTC,GAA,T	*E. coli* C08004510	50	60	11
TGT 34		CC,CTG,GGT,GGC,T	*T. thermophilus* C025958	62	73	12
GGT 34		T,CTG,GGT,ATC,A	*E. coli* C08004523	60	70	11
CGT 33		G,GCC,ACC,GGC,T	*T. thermophilus* C025936	62	72	11
CGT 34		T,ATT,ATC,GGC,ACC	*E. coli* C08004472	63	75	13
TGT 34		G,GTA,GTC,GGC,ACC	*E. coli* C08004507	63	75	13

Table 4.24: Identification of X circular code motifs trp-tRNAXm in tRNAs of Tryptophan (Trp) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
CCA 34	G,GGC,GTA,GTT,CA		*E. coli* C08004499	3	14	12
CCA 37	GT,GGT,GTA,GCC,T		*P. furiosus* At1854	7	18	12
CCA 34	AG,TTC,AAT,T		*E. coli* C08004499	9	17	9
CCA 34	AG,CTC,AAC,T		*T. thermophilus* C025938	9	17	9
CCA 37	GC,CTG,GTC,CA		*P. furiosus* At1854	15	24	10
CCA 34	TG,GTA,GAG,CA		*E. coli* C08004499	17	26	10
CCA 37	CC,ATC,ATC,GC		*P. furiosus* At1854	22	31	10
CCA 34	GC,ACC,GGT,CT C,CA		*E. coli* C08004499	24	36	13
CCA 34	GC,ACC,GGT,CT C,CA		*T. thermophilus* C025938	24	36	13
CCA 37	G,CT C,CA G,ACC,C		*P. furiosus* At1854	34	44	11
CCA 34	G,GGT,GTT,GG		*E. coli* C08004499	42	50	9
CCA 34	TG,GAG,GTT,C		*T. thermophilus* C025938	48	56	9
CCA 34	G,GAG,TTC,GAG,T		*E. coli* C08004499	50	60	11

GT, GTC, GGC, GGT, T are shifted in frame. The X motif CC, CTC, GGC, ACC is observed.

4.2.3.15 X CIRCULAR CODE MOTIFS IN PRO-TRNAS (TABLE 4.20)

(i) 5' region of Pro-tRNAs: Several classes of X motifs are shifted in series: S, GSC, GWR, K, $Pro-tRNAXm_1$ $G, CAG, \underline{GCC}, G\underline{GT}, A$, $\underline{GC}, CT\underline{G}, GTA, GC$ shifted by $+1$ nucleotide from $Pro-tRNAXm_1$, CC, ATC, CTG, C and GC, ACC, GTC, AT.

Table 4.25: Identification of X circular code motifs tyr-tRNAXm in tRNAs of Tyrosine (Tyr) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
GTA 36	G,GTA,GCC,TA		*P. furiosus* At1836	6	14	9
GTA 36	GC,CTG,GTA,GT		*P. furiosus* At1836	15	24	10
GTA 36	GT,GGC,GGC,GG		*P. furiosus* At1836	23	32	10
GTA 35	G,GAG,CAG,AC		*E. coli* C08004508	25	33	9
GTA 35	G,GAG,CAG,AC		*E. coli* C08004542	25	33	9
GTA 35	G,GAC,GGT,CT	G,TA	*T. thermophilus* C025959	26	37	12
GTA 35	T,	GTA, AAT,CTG,C	*E. coli* C08004508	34	44	11
GTA 35	T,	GTA, AAT,CTG,C	*E. coli* C08004542	34	44	11
GTA 35		AA,ACC,GTT,GGC,GTA,T	*T. thermophilus* C025959	38	52	15
GTA 35		T,GCC,GTC,ATC,GAC,TTC,GAA,GGT,T	*E. coli* C08004542	42	64	23
GTA 35		T,GCC,GTC,AC	*E. coli* C08004508	42	50	9
GTA 35		A,GAC,TTC,GAA,GGT,T	*E. coli* C08004508	51	64	14
GTA 35		AT,GCC,TTC,GC	*T. thermophilus* C025959	51	60	10

Table 4.26: Identification of X circular code motifs val-tRNAXm in tRNAs of Valine (Val) in prokaryotes (bacteria *E. coli* and *T. thermophilus*, archaea *P. furiosus*).

AC pos	5' region	AC 3' region	Organism ID	Start	End	Length
TAC 34	G,GGT,GAT,TA		*E. coli* C08004475	1	9	9
TAC 35	G,GGC,GGC,TA		*T. thermophilus* C025924	1	9	9
CAC 33	G,GGC,GGC,TA		*T. thermophilus* C025949	1	9	9
TAC 34	T,CAG,CTG,GG		*E. coli* C08004475	12	20	9
TAC 35	T,CAG,CTG,GCC,A		*T. thermophilus* C025924	12	22	11
TAC 35	AG,CTG,GTT,AT		*P. furiosus* At1841	14	23	10
GAC 35	A,GTT,GGT,TA		*E. coli* C08004481	14	22	9
GAC 35	A,GTT,GGT,TA		*E. coli* C08004482	14	22	9
GAC 35	A,GGT,GGC,TA		*T. thermophilus* C025926	14	22	9
GAC 36	A,CTG,GTT,AT		*P. furiosus* At1835	16	24	9
CAC 36	A,CTG,GTT,AT		*P. furiosus* At1840	16	24	9
TAC 35	AT,GAC,GCC,GCC,CT		*P. furiosus* At1841	22	34	13
GAC 35	GC,ACC,ACC,TT		*E. coli* C08004481	25	34	10
GAC 35	GC,ACC,ACC,TT		*E. coli* C08004482	25	34	10
CAC 36	AT,GAC,GCC,GCC,CT	C,AC	*P. furiosus* At1840	23	38	16
GAC 36	AT,GAC,GCC,ACC,CT	G,AC	*P. furiosus* At1835	23	38	16
TAC 35	A,CTC,GCC,T	T	*T. thermophilus* C025924	27	35	9
TAC 35	T,	TAC, GAG,GC	*P. furiosus* At1841	34	42	9
CAC 33		A,GAG,GTC,GTA,GGT,T	*T. thermophilus* C025949	41	54	14
GAC 36		TG,GAG,GTC,C	*P. furiosus* At1835	43	51	9
TAC 35		A,GAG,GTT,CA	*T. thermophilus* C025924	50	58	9
TAC 34		G,GTC,GGC,GGT,T	*E. coli* C08004475	45	55	11
GAC 35		G,GTC,GTT,GGT,T	*E. coli* C08004481	46	56	11
GAC 35		G,GTC,GGT,GGT,T	*E. coli* C08004482	46	56	11
GAC 35		G,GTC,GGT,GGT,T	*T. thermophilus* C025926	46	56	11
TAC 35		AA,GTC,CTC,T	*T. thermophilus* C025924	58	66	9
TAC 34		CC,GTC,ATC,ACC,CA	*E. coli* C08004475	61	73	13
CAC 33		CC,TAC,GCC,GCC,CA	*T. thermophilus* C025949	60	72	13
TAC 35		T,GCC,GCC,CA	*T. thermophilus* C025924	66	74	9

(ii) anti-codon regions of Pro-tRNAs: The class of X motifs $Pro-tRNAXm_2$ AM, YTS, GTT, Y is in frame with the anti-codons CGG and TGG. The class of X motifs S, GGC, YTR, GG is in a different frame than the anti-codons CGG and GGG. The class of X motifs G, GGY, GYC, GG is in a different frame than the anti-codon GGG. The class of X motifs $Pro - tRNAXm_3$ G, GAS, SAR, GG is in frame with the anti-codons CGG and TGG. The two classes of X motifs $Pro - tRNAXm_2$ and $Pro - tRNAXm_3$ may derive from an ancestral class of X motifs constructed by the concatenation of $Pro-tRNAXm_2$ and $Pro-tRNAXm_3$ $AM, YTS, GTT, YAG, GAS, SAR, GG$ of 19 nucleotides. Indeed, CAG

belongs to X (Equation 2.2). Then, the nucleotide A in the middle site of CAG has mutated to G for building the anti-codon CGG.

(iii) 3' region of Pro-tRNAs: The class of X motifs $\underline{G}, SW\underline{G}, \underline{GTT}, CA$ is shifted by +2 nucleotides from the X motif $G, GTC, \underline{G}TC, \underline{GGT}, \underline{T}$. Several classes of X motifs are observed: A, ATC, CWS, T, CC, GGC, SKC, C and T, CTC, GCC, GAC.

4.2.3.16 X circular code motifs in SeC-tRNAs (Table 4.21)

(i) 5' region of SeC-tRNAs: Two X motifs AG, ATC, GTC, GTC, T and CC, GGT, GAG, GC are observed.

(ii) anti-codon regions of SeC-tRNAs: The X motif $SeC - tRNAXm$ $G, CTG, GAC, TTC, \underline{AA}$ is in a different frame than the anti-codon TCA. The X motif $\underline{AA}, ATC, CAG, TT$ shifted by +1 nucleotide from $SeC - tRNAXm$ is also in a different frame than the anti-codon TCA.

(iii) 3' region of SeC-tRNAs: One X motif TG, ATC, TTC, C is observed.

4.2.3.17 X circular code motifs in Ser-tRNAs (Table 4.22)

(i) 5' region of Ser-tRNAs: Several classes of X motifs are shifted: R, GNW, GCC, KR, $Ser - tRNAXm_1$ $TG, GCC, GA\underline{G}, \underline{C}$, $\underline{GC}, CTG, GTA, GG$ shifted by +1 nucleotide from $Ser - tRNAXm_1$, G, CTG, AAC, GG, the class of large X motifs $Ser - tRNAXm_2$ $G, GTT, GAA, GGC, RSC, GGT, CT$ of 18 nucleotides in *T. thermophilus* and *E. coli* and K, GAA, GGC, GC factor of $Ser - tRNAXm_2$.

(ii) anti-codon regions of Ser-tRNAs: The very large X motif $Ser - tRNAXm_3$ $TG, GTC, GAA, GGC, GGC, ACC, CTG, CT$ of 22 nucleotides in *T. thermophilus* is in a different frame than the anti-codon GCT. The large X motif $Ser - tRNAXm_4$ $AG, GGC, GCC, GGC, CTG, CT$ of 16 nucleotides in *P. furiosus* which is suffix of $Ser - tRNAXm_3$, is thus in a different frame than the anti-codon GCT. The class of large X motifs $Ser - tRNAXm_5$ $G, RCC, GGT, CTC, GAA, AAC, C$ of 17 nucleotides in *E. coli* and *T. thermophilus* is in a different frame than the anti-codon CGA. The X motif A, CTC, GAG, ATC, C which is suffix of $Ser - tRNAXm_5$, is thus in a different frame than the anti-codon CGA. The class of X motifs GC, CTG, GAR, A and A, CTG, GAA, ATC, GT which are shifted in

frame, are in a different frame than the anti-codon GGA. The X motif $Ser - tRNAXm_6$ TT, GAA, AAC, C is in a different frame than the anti-codon TGA.

(iii) 3' region of Ser-tRNAs: The class of X motifs $\underline{AA, RSY}, GKY, R$ is shifted by +2 nucleotides from $Ser - tRNAXm_6$ $TT, G\underline{AA, AAC, C}$. A large X motif $Ser - tRNAXm_7$ $GT, GTA, TAC, GGC, AAC, GTA, T$ of 18 nucleotides is identified in *E. coli*. The X motif $\underline{G, TTC}, GCC, CA$ is shifted by +2 nucleotides from the class of X motifs $TG, GGC, \underline{GTT, Y}$. Several classes of X motifs are observed: GC, GAA, GCC, CA, AA, MYC, KRY, S, R, RRS, YTC, S and YC, RCC, CTC, Y.

4.2.3.18 X circular code motifs in Thr-tRNAs (Table 4.23)

(i) 5' region of Thr-tRNAs: Several classes of X motifs are shifted: CC, GGT, GKM, K, $Thr - tRNAXm_1$ $G, GT\underline{A, GCC}, T\underline{A}, \underline{AG, C}TC, AAC, C$ shifted by +2 nucleotides from $\underline{Thr - tRNAXm_1}$, $T, CAG, CAG, GTA, GAG, CA$ and its suffix G, GTA, GAG, C.

(ii) anti-codon regions of Thr-tRNAs: The class of X motifs $Thr - tRNAXm_2$ M, YTC, GTA, A is in a different frame than the anti-codon CGT. The X motif $Thr - tRNAXm_3$ TT, GTA, ATC, shifted in frame with $Thr - tRNAXm_2$, is in a different frame than the anti-codon TGT. Interestingly, the X motif $\underline{GT, AAT, C}AG, TA$ shifted by +1 nucleotide from $Thr - tRNAXm_3$ $TT, \underline{GTA, ATC}$ is, in contrast, in frame with the anti-codon TGT.

(iii) 3' region of Thr-tRNAs: The large X motif $Thr - tRNAXm_4$ $AG, GGT, GAG, GTC, GCC, GGT, T$ of 18 nucleotides is identified in *T. thermophilus*. The class of X motifs AG, GGT, GAG, GTC, S is prefix of $Thr - tRNAXm_4$. Several classes of X motifs are observed: M, GWA, GGT, C, CC, CAG, GTC, S, AG, GYC, RYM, R, S, CAG, TTC, GAN, T, Y, CTG, GGT, RKC, W and K, RYN, RYC, GGC, W.

4.2.3.19 X circular code motifs in Trp-tRNAs (Table 4.24)

(i) 5' region of Trp-tRNAs: The X motif $\underline{GT}, G\underline{GT}, GT\underline{A}, GCC, T$ is shifted by +1 nucleotide from $G, GGC, \underline{GT}A, \underline{GT}T, \underline{CA}$. Several classes of X motifs are identified: AG, YTC, AAY, T, GC, CTG, GTC, CA, TG, GTA, GAG, CA and CC, ATC, ATC, GC.

(ii) anti-codon regions of Trp-tRNAs: The X motif $Trp - tRNAXm$ GC, ACC, GGT, CTC, CA is in a different frame than the anti-codon CCA. The X motif G, CTC, CAG, ACC, C shifted in frame with $Trp - tRNAXm$, is also in a different frame than the anti-codon CCA.

(iii) 3' region of Trp-tRNAs: The class of X motifs $\underline{TG}, \underline{GAG}, KTY, S$ is shifted by +2 nucleotides from the X motif $G, GGT, \underline{GT}T, \underline{GG}$.

4.2.3.20 X circular code motifs in Tyr-tRNAs (Table 4.25)

(i) 5' region of Tyr-tRNAs: The X motif G, GTA, GCC, TA is observed. The X motif $\underline{GT}, GGC, GGC, GG$ is shifted by +1 nucleotide from the X motif $GC, CTG, GTA, \underline{GT}$. The X motif G, GAG, CAG, AC occurs before the anti-codon GTA.

(ii) anti-codon regions of Tyr-tRNAs: The X motif $Tyr - tRNAXm_1$ $G, GAC, GGT, C\underline{TG}, \underline{TA}$ is in a different frame than the anti-codon GTA. Interestingly, the X motif $\underline{T}, \underline{GTA}, AAT, CTG, C$ shifted by +1 nucleotide from $Tyr - tRNAXm_1$ is, in contrast, in frame with the anti-codon GTA.

(iii) 3' region of Tyr-tRNAs: The X motif $\underline{AA}, \underline{AC}\underline{C}, GTT, GGC, GTA, T$ is shifted by +2 nucleotides from the X motif $T, GT\underline{A}, \underline{AA}T, \underline{C}TG, C$. A very large X motif $Tyr - tRNAXm_2$ $T, GCC, GTC, ATC, GAC, TTC, GAA, GGT, T$ of 23 nucleotides is identified in *E. coli*. The X motif T, GCC, GTC, AC is prefix of $Tyr - tRNAXm_2$, the X motif A, GAC, TTC, GAA, GGT, T is suffix of $Tyr - tRNAXm_2$ and the X motif AT, GCC, TTC, GC is factor of $Tyr - tRNAXm_2$.

4.2.3.21 X circular code motifs in Val-tRNAs (Table 4.26)

(i) 5' region of Val-tRNAs: Several classes of X motifs are shifted: G, GGY, GRY, TA, T, CAG, CTG, GS, $Val - tRNAXm_1$

$\underline{AG}, C\underline{TG}, GTT, \underline{AT}$, $\underline{A}, \underline{GT}\underline{T}, \underline{GGT}, \underline{TA}$ shifted by +1 nucleotide from $Val - tRNAXm_1$, A, GGT, GGC, TA, $Val - tRNAXm_2$ $A, CTG, GTT, \underline{AT}$, $\underline{AT}, GAC, GCC, GCC, CT$ shifted by +1 nucleotide from $Val - tRNAXm_2$ and GC, ACC, ACC, TT which occurs before the anti-codon GAC.

(ii) anti-codon regions of Val-tRNAs: The class of large X motifs $Val - tRNAXm_3$ $AT, GAC, GCC, RCC, CTS, AC$ of 16 nucleotides identified in *P. furiosus* is in a different frame than the anti-codons CAC and GAC. The X motif $Val - tRNAXm_4$ $A, CTC, GCC, \underline{TT}$ is in a different frame than the anti-codon TAC. Interestingly, the X motif $\underline{T}, \underline{T}AC, GAG, GC$ shifted by +2 nucleotides from $Val - tRNAXm_4$ is, in contrast, in frame with the anti-codon TAC.

(iii) 3' region of Val-tRNAs: Several classes of X motifs are observed: $Val - tRNAXm_5$ A, GAG, GTC, GTA, GGT, T, R, GAG, GTY, C prefix of $Val - tRNAXm_5$, G, GTC, GKY, GGT, T suffix of $Val - tRNAXm_5$, AA, GTC, CTC, T, CC, KWC, RYC, RCC, CA and T, GCC, GCC, CA.

4.2.3.22 Summary of X circular code motifs in tRNA sequences

We mention some of the properties of X motifs found in tRNAs:

(i) an X motif can occur at the same position in the same isoaccepting tRNA of different species, e.g. the X motif CAG, GAG, GTC is at position 40 in $Ala - tRNA$ of *E. coli* and *T. thermophilus* (Table 4.6), the X motif T, CAG, CTG, GAT, A is at position 12 in $Arg - tRNA$ of *E. coli* and *T. thermophilus* (Table 4.7), etc.

(ii) an X motif can occur at the same position in different isoaccepting tRNAs, e.g. the X motif T, CAG, CTG, G is at position 12 in $Ala - tRNA$ of *E. coli*, $Arg - tRNA$ of *E. coli* and *T. thermophilus*, $His - tRNA$ of *T. thermophilus*, $Ile - tRNA$ of *T. thermophilus* and $Lys - tRNA$ of *T. thermophilus*, $Val - tRNA$ of *E. coli* and *T. thermophilus* (Table 4.6,4.7,4.14,4.15,4.17 and 4.26), the X motif GTA, GTT, CAG is at the same position 7 in $Arg - tRNA$, $Asn - tRNA$, $Asp - tRNA$, $Gly - tRNA$, $Trp - tRNA$ of *E. coli* (Table 4.7, 4.8, 4.9, 4.13 and 4.24), etc.

(iii) an X motif can be shifted by 0, +1 or +2 mod 3 nucleotides from another X motif in the same species or in different species.

(a) an X motif can be in the same frame as the anti-codon, e.g. the very large X motif $Ala - tRNAXm_1$ $GC, CTC, AAT, GGC, ATT, GAG, GAG, GTC, A$ of 24 nucleotides in $Ala - tRNA$ of *T. thermophilus* is in frame with the anti-codon GGC (Table 4.6), the X motif G, ATT, CTG, ATT, C in $Gln - tRNA$ of *E. coli* is in frame with the anti-codon CTG (Table 4.11), etc.

(b) an X motif can be in a different frame than the anti-codon with a shift of one nucleotide, e.g. the X motif $Cys - tRNAXm_3$ A, CTG, CAG, ATC, C in $Cys - tRNA$ of *P. furiosus* is shifted by +1 nucleotide relative to the anti-codon GCA (Table 4.10), etc.

(iv) an X motif can be in a different frame than the anti-codon with a shift of two nucleotides, e.g. the large X motif $Arg - tRNAXm_6$ $A, GTA, CTC, GGC, TAC, GAA, C$ of 17 nucleotides in $Arg - tRNA$ of *E. coli* is shifted by +2 (-1) nucleotides relative to the anti-codon ACG (Table 4.7), etc.

4.2.3.23 Coverage of X circular code motifs in prokaryotic tRNAs

Table 4.27 shows that the coverage (Equation 3.5) of X motifs (Algorithm 3.2) is greater in the 5' region of tRNAs in *E. coli*, *T. thermophilus* and *P. furiosus* (mean equal to 88%, minimum equal to 62% and maximum equal to 100%) compared to their 3' region (mean equal to 71%, minimum equal to 23% and maximum equal to 98%). The coverage of X motifs is maximal (100%) in the 5' region of tRNAs of Gly, Leu, Pro and Thr, and minimal (around 30%) in the 3' region of tRNAs of Asp, Glu, His and SeC. This indicates a possible role of X circular code in the formation of tRNA sequences, or the possibility of X motifs in tRNAs interacting with X motifs in rRNA to help with positioning on the A, P and E sites of the ribosome.

4.3 Analysis of X circular code motifs in eukaryotic genomes

The motifs studied in this section are large, having a minimum length of 30 nucleotides (Equation 3.4) which were extracted using the Frames algorithm (Algorithm 3.2) with the eukaryotic genomes (Table 3.2).

Table 4.27: The lengths and the anti-codon positions are average values in each isoaccepting tRNA. The coverage of X motifs is given in rounded percentage.

tRNA	Length	anti-codon position	Coverage of X motifs(%) 5' region	3' region
Ala	76	34	67	63
Arg	77	35	97	88
Asn	76	34	91	63
Asp	78	36	80	23
Cys	74	33	94	77
Gln	75	34	85	77
Clu	77	34	91	32
Cly	76	35	100	95
His	77	35	85	35
Ile	77	35	62	93
Leu	87	36	100	94
Lys	78	34	76	69
Met	77	35	91	98
Phe	76	34	85	70
Pro	78	36	100	93
SeC	95	35	88	31
Ser	89	35	88	92
Thr	77	34	100	95
Trp	77	35	94	57
Tyr	83	35	85	59
Val	78	35	94	90
Mean	79	35	88	71
Min	74	33	62	23
Max	95	36	100	98

4.3.1 Occurrence of large randoms code motifs in eukaryotic genomes

The mean number $\overline{N}(m(R)) = \frac{1}{Card(R)} \sum_{j=1}^{Card(R)} N(m(R_j))$ and its standard deviation $\sigma(m(R))$ of large R motifs $m(R)$ from $Card(R) = 30$ random codes are determined in the 138 eukaryotic genomes. The computation leads to $\overline{N}(m(R)) = 1171$ and $\sigma(M(R)) = 1170$. By assuming a normal distribution of the population, a student t test gives a confidence interval at 99% for the mean $\overline{N}(m(R))$ equal to $[582, 1760]$ (shown in Figure 4.14). Note that the number of random codes was limited to 30 as their statistical analysis in the 138 eukaryotic genomes stretches into days.

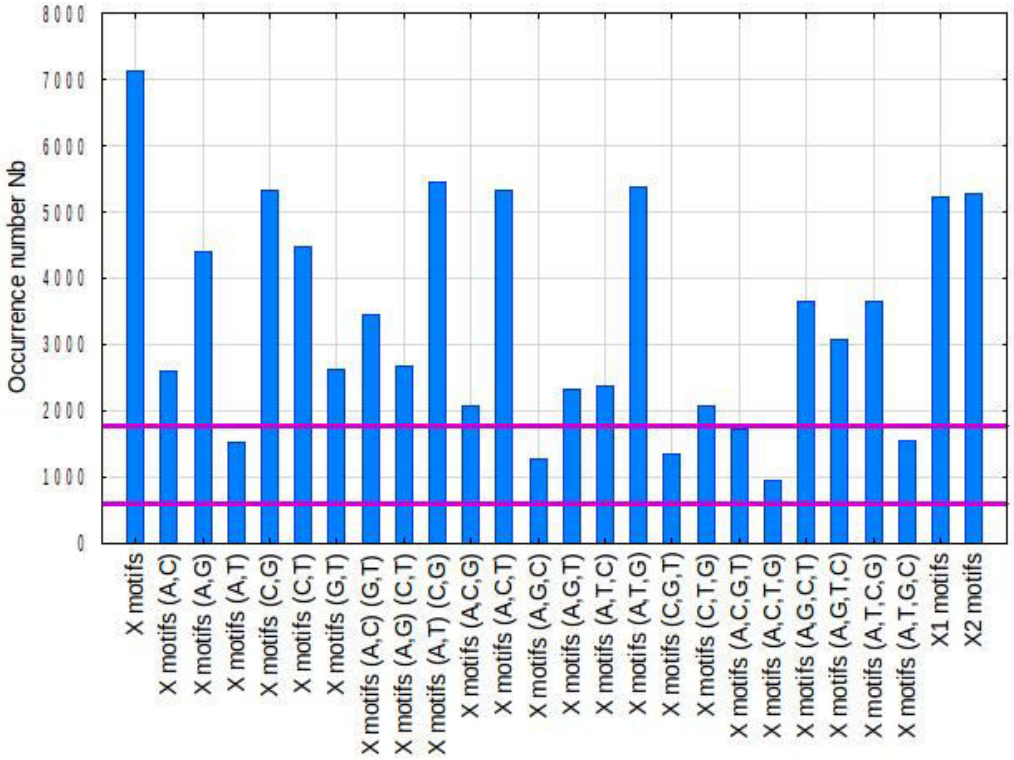

Figure 4.14: Occurrence numbers $N(m(X))$ of large X motifs $m(X)$, $N(m(\Pi(X)))$ of its 23 large bijective motifs $m(\Pi(X))$, $N(m(X_1))$ and $N(m(X_2))$ of its two large permuted motifs $m(X_1)$ and $m(X_2)$, respectively,in the 138 eukaryotic genomes. All these 26 classes of large motifs have lengths $l \geq 15$ trinucleotides and cardinality (composition) $Card \geq 10$ trinucleotides. The top horizontal line (1760) and the bottom horizontal line (582) represent the confidence interval at 99% (student t test by assuming a normal distribution of the population) of the mean occurrence number $\overline{N}(m(R)) = 1171$ (standard deviation $\sigma(m(R)) = 1170$) of large R motifs $m(R)$ from $Card(R) = 30$ random codes in the 138 eukaryotic genomes. The large X motifs $m(X)$ have the highest occurrence. The six large bijective motifs $m(\pi_2(X) : (A,G))$, $m(\pi_4(X) : (C,G))$, $m(\pi_5(X) : (C,T))$, $m(\pi_9(X) : (A,T)(C,G))$, $m(\pi_{11}(X) : (A,C,T))$ and $m(\pi_{15}(X) : (A,T,G))$, and the two large permuted motifs $m(X_1)$ and $m(X_2)$ have occurrence numbers greater than $\overline{N}(m(R)) + 2.75\sigma(m(R)) \approx 4400$.

4.3.2 Occurrence of large motifs from X, X_1, X_2 and the 23 bijective transformations of X in eukaryotic genomes

Figure 4.14 shows the occurrence numbers $N(m(X))$ of large X motifs $m(X)$, $N(m(\Pi(X)))$ of its 23 large bijective motifs $m(\Pi(X))$, $N(m(X_1))$ and $N(m(X_2))$ of its two large permuted motifs $m(X_1)$ and $m(X_2)$, respectively, in the 138 eukaryotic genomes. All these 26 classes of large motifs have lengths $l \geq 15$ trinucleotides and cardinality (composition) $Card \geq 10$ trinucleotides. The large X motifs $m(X)$ have the highest occurrence with $N(m(X)) = 7133$ compared to all the 25 other classes of large motifs $m(\Pi(X))$, $m(X_1)$ and $m(X_2)$ in genomes of eukaryotes. Eight large motifs also occur significantly with numbers

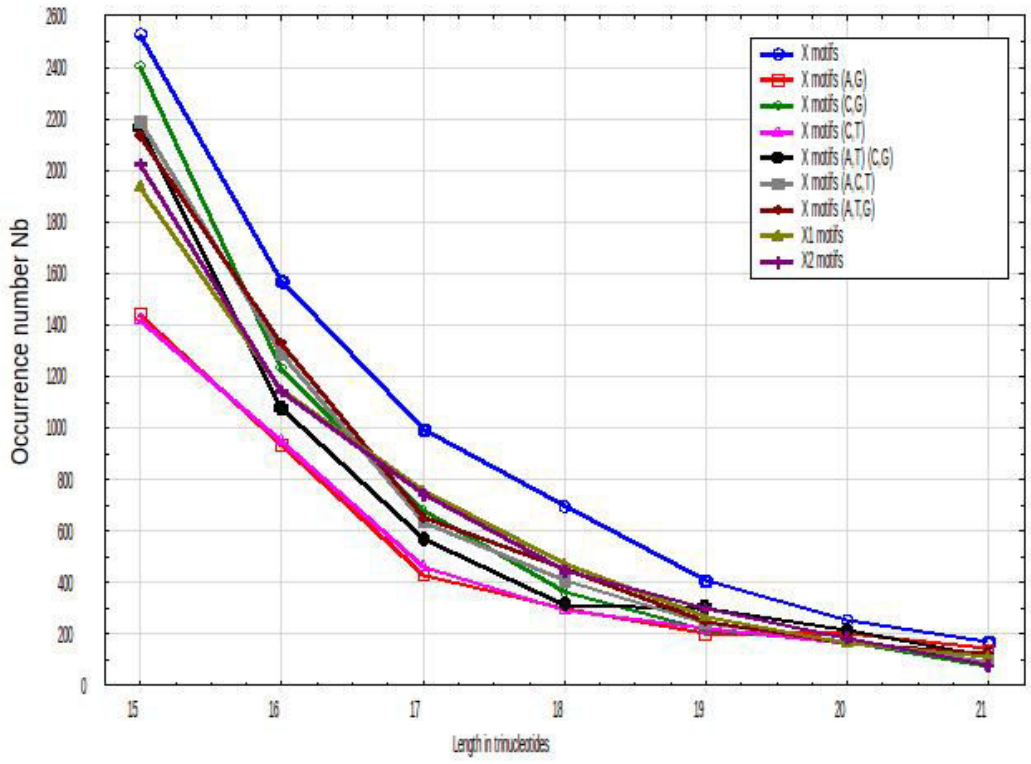

Figure 4.15: Occurrence numbers $N(m(X))$ of large X motifs $m(X), N(m(\Pi(X)))$ of its six large bijective motifs $m(\pi_2(X) : (A,G)), m(\pi_4(X) : (C,G)), m(\pi_5(X) : (C,T)), m(\pi_9(X) : (A,T)(C,G)),$ $m(\pi_{11}(X) : (A,C,T))$ and $m(\pi_{15}(X) : (A,T,G)), N(m(X_1))$ and $N(m(X_2))$ of its two large permuted motifs $m(X_1)$ and $m(X_2)$ (greater than $\overline{N}(m(R)) + 2.75\sigma(m(R)) \approx 4400$, see Figure 4.14) as a function of their lengths l varying from 15 to 21 trinucleotides in the 138 eukaryotic genomes. All these classes of large motifs have a cardinality $Card \geq 10$ trinucleotides (Equation 3.4). The large X motifs have the highest occurrence for all trinucleotide lengths.

greater than $\overline{N}(m(R)) + 2.75\sigma(m(R)) \approx 4400$ (where $\overline{N}(m(R))$ and $\sigma(m(R))$ are given in Section 4.3.1). They are in descending order: $m(\sigma_4(X) : (A,T)(C,G))$ with $N(m(\sigma_4(X))) = 5447$, $m(\sigma_{15}(X) : (A,T,G))$ with $N(m(\sigma_{15}(X))) = 5374$, $m(\sigma_{11}(X) : (A,C,T))$ with $N(m(\sigma_{11}(X))) = 5341$, $m(X_2)$ with $N(m(X_2)) = 5289$, $m(X_1)$ with $N(m(X_1)) = 5223$, $m(\pi_5(X) : (C,T))$ with $N(m(\pi_5(X)) = 4466$ and $m(\pi_2(X) : (A,G))$ with $N(m(\pi_2(X) : (A,G))$ with $N(m(\pi_2(X))) = 4404$ (Figure 4.14). Note that $\pi_2(X)$, $\pi_4(X)$ and $\pi_5(X)$ are symmetric bijective transformation circular codes $\Pi_{\mathcal{S},2}(X)$ at 2 letters, $\pi_9(X)$ is a symmetric bijective transformation circular code $\Pi_{\mathcal{S},2,2}(X)$ of two disjoint transformations at 2 letters, and $\pi_{11}(X)$ and $\pi_{15}(X)$ are asymmetric bijective transformation circular codes $\Pi_{\mathcal{A},3}(X)$ at 3 letters. Note also that $\pi_4(X)$ and $\pi_9(X)$ are C^3 self-complementary trinucleotide circular codes.

The six motifs $m(\pi_3(X) : (A,T))$, $m(_{12}(X) : (A,G,C))$, $m(\pi_{16}(X)$:

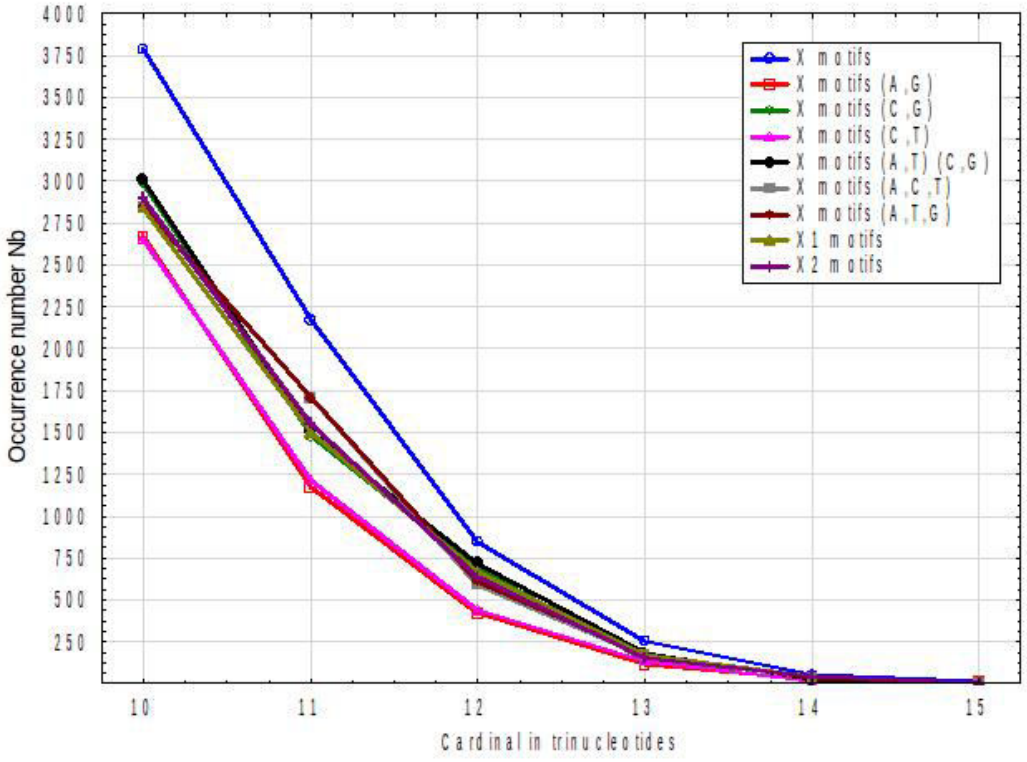

Figure 4.16: Occurrence numbers $N(m(X))$ of large X motifs $m(X)$, $N(m(\Pi(X)))$ of its six large bijective motifs $m(\pi_2(X) : (A,G)), m(\pi_4(X) : (C,G)), m(\pi_5(X) : (C,T)), m(\pi_9(X) : (A,T)(C,G))$, $m(\pi_{11}(X) : (A,C,T))$ and $m(\pi_{15}(X) : (A,T,G))$, $N(m(X_1))$ and $N(m(X_2))$ of its two large permuted motifs $m(X_1)$ and $m(X_2)$ (greater than $\overline{N}(m(R)) + 2.75\sigma(m(R)) \approx 4400$, see Figure 4.14) as a function of their cardinality (composition) $Card$ varying from 10 to 15 trinucleotides in the 138 eukaryotic genomes. All these classes of large motifs have a cardinality $Card \geq 10$ trinucleotides (Equation 3.4). The large X motifs have the highest Occurrence for all trinucleotide cardinalities.

$(C,G,T))$, $m(\pi_{18}(X) : (A,C,G,T))$, $m(\pi_{19}(X) : (A,C,T,G))$ and $m(\pi_{23}(X) : (A,T,G,C))$ occur randomly $(N(m(\pi_i(X))) \in [582, 1760]$, $i = 3, 12, 16, 18, 19, 23$, see section 4.3.1) and the four motifs $m(\pi_{10}(X) : (A,C,G))$, $m(\pi_{13}(X) : (A,G,T))$, $m(\pi_{14}(X) : (A,T,C))$ and $m(\pi_{17}(X) : (C,T,G))$ have low occurrences $(2000 < N(m(\pi_i(X)) < 2400, i = 10, 13, 14, 17)$ (Figure 4.14).

Figures 4.15 and 4.16 strengthen the above mentioned results. Indeed, 4.15 shows that the large X motifs with cardinality $Card \geq 10$ trinucleotides (Equation 3.4) have the highest occurrence compared to all the 25 other classes of large motifs $m(\Pi(X))$, $m(X_1)$ and $m(X_2)$ for all lengths l from 15 to 21 trinucleotides. Figure 4.16 shows that the large X motifs with length $l \geq 15$ trinucleotides (Equation 3.4) have the highest occurrence compared to all the 25 other classes of large motifs $m(\Pi(X))$, $m(X_1)$ and $m(X_2)$ (with length $l \geq 15$ trinucleotides) for all cardinalities $Card$ from 10 to 15 trinucleotides.

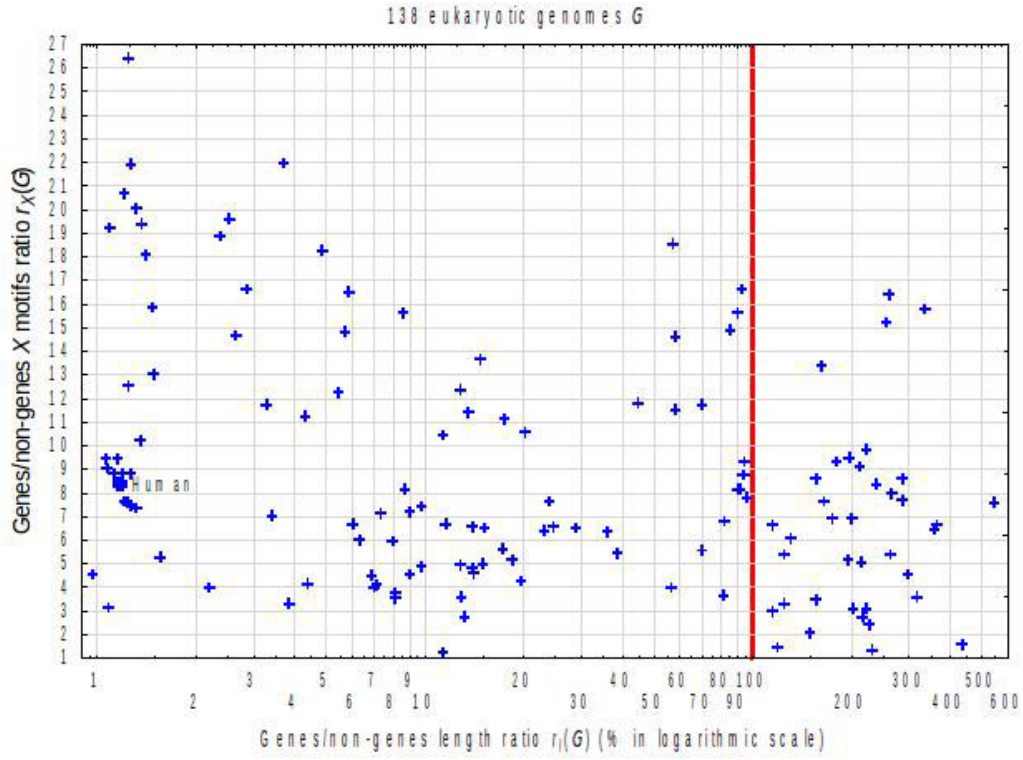

Figure 4.17: Base ratio $r(\mathcal{G}$ (Equation 3.7 in %) of coding/non-coding regions and base ratio $r_{m(X)}(\mathcal{G})$ (Equation 3.9) of X motifs $m(X)$ of length $l \geq 10$ trinucleotides and cardinality (composition) $Card \geq 5$ trinucleotides (Equation 3.8) in coding/non-coding regions of the 138 eukaryotic genomes $\mathcal{G}$. The vertical red line $r(\mathcal{G}) = 100\%$ makes a partition of genomes $\mathcal{G}$ according to their base content in coding regions. When $r(\mathcal{G}) < 100\%$, the total base number $N(\mathcal{G}_C)$ of coding regions $\mathcal{G}_C$ in genome $\mathcal{G}$ is less than the total base number $N(\mathcal{G}_{\overline{C}})$ of all non-coding regions $\mathcal{G}_{\overline{C}}$ in $\mathcal{G}$, and conversely when $r(\mathcal{G}) > 100\%$. The genome $\mathcal{G} =$ *Plasmodium falciparum* with $r_{m(X)}(\mathcal{G}) = 79.3$ is not represented in the figure. There is no correlation between $r(\mathcal{G})$ and $r_{m(X)}(\mathcal{G})$.

4.3.3 Largest X motifs in eukaryotic genomes

Table 4.28 shows the top 20 largest X motifs $m_{\mathcal{G}_{Chr}}(X)$ with cardinality (composition) $Card \geq 10$ trinucleotides (Equation 3.4) in the chromosomes $\mathcal{G}_{Chr}$ of the 138 eukaryotic genomes $\mathcal{G}$ in descending order of their length (in trinucleotides) $l \geq 45$. The 1^{st} largest X motif $m_{Solanum_3}(X)$ is observed in a non-coding region of the chromosome $Chr = 3$ in the genome $\mathcal{G} = Solanum\ pennellii$. It has a length of $l = 155$ trinucleotides (456 nucleotides) and an expectation $E\big[N\big(m_{Solanum_3}(X)\big)\big] = 10^{-71}$ (Equation 3.6). The 2^{nd} and 3^{rd} largest X motifs $m_{Salmo_{15}}(X)$ are found in non-coding regions of chromosome $Chr = 15$ in the $\mathcal{G} = Salmosalar$ genome. They have a different composition but the same length $l = 118$ trinucleotides (354 nucleotides) and an expectation $E\big[N\big(m_{Solmo_15}(X)\big)\big] = 10^{-52}$. The biological function and evolution of these

Table 4.28: The top 20 largest X motifs $m_{\mathcal{G}_{Chr}}(X)$ with cardinality (compostion) $Card \geq 10$ trinucleotides (Equation 3.4) in the chromosomes $\mathcal{G}_{Chr}$ of the 138 eukaryotic genomes $\mathcal{G}$ in descending order of length (in trinucleotide) $l \geq 45$. The 1^{st} and 2^{nd} columns give the genome $\mathcal{G}$ and its chromosome number $\mathcal{G}_{Chr}$, respectively, the 3^{rd} column gives its bases size $N(\mathcal{G}_{Chr})$, the 4^{th} and 5^{th} columns indicate the start and end positions of the large X motif, the 6^{th} column gives the length l in trinucleotide, the 7^{th} column indicates its expectation E (Equation 3.6) and the last column shows if the motif is in a coding region (Yes) or non-coding region (No).

Genome $\mathcal{G}$	$\mathcal{G}_{Chr}$	Size (in bases) $N(\mathcal{G}_{Chr})$	Start	End	Length (in trinucleotides)	Expectation E	In coding region
Solanum pennellii	3	75414019	36982714	36983178	155	10^{71}	No
Salmo salar	15	103963436	16024777	16025130	118	10^{52}	No
Salmo salar	15	103963436	17850373	17850726	118	10^{52}	No
Monodelphis domestica	2	541556283	513328228	513328533	102	10^{43}	No
Solanum lycopersicum	8	65866657	30359989	30360276	96	10^{41}	No
Monodelphis domestica	4	435153693	290107123	290107407	95	10^{40}	No
Plasmodium falciparum	11	2038337	872956	873216	87	10^{38}	Yes
Equus caballus	28	46177339	35484817	35485047	77	10^{32}	No
Bombus terrestris	14	11649563	11165956	11166153	66	10^{27}	Yes
Sorghum bicolor	4	68034345	38474677	38474856	60	10^{23}	No
Felis catus	3	140925898	2211844	2212020	59	10^{22}	No
Cynoglossus semilaevis	9	19616557	14919031	14919192	54	10^{20}	No
Plasmodium knowlesi	13	2200295	1265167	1265322	52	10^{20}	Yes
Mus musculus	1	195471971	74368813	74368968	52	10^{18}	Yes
Micromonas sp.	12	1084119	530353	530496	48	10^{19}	Yes
Dictyostelium discoideum	2	8484197	1796161	1796304	48	10^{18}	Yes
Apis mellifera	4	12718334	12440101	12440241	47	10^{17}	No
Salmo salar	19	82978132	46877047	46877184	46	10^{16}	No
Bombus terrestris	15	11467329	3286219	3286353	45	10^{16}	No
Camelina sativa	10	25316904	13177546	13177680	45	10^{16}	No

unexpected large X motifs in the eukaryotic genomes are unknown.

Table 4.29: Largest X motifs $m_{\mathcal{H}_{Chr}}(X)$ with cardinality (compostion) $Card \geq 10$ trinucleotides (Equation 3.4) and expectation $E < 1$ (Equation 3.6) in the chromosomes $\mathcal{H}_{Chr}$ of the human genome $\mathcal{G} = \mathcal{H} =$ *Homo sapiens*. The 1^{st} and 2^{nd} columns give the human chromosome number $\mathcal{H}_{Chr}$ and its base size $N(\mathcal{H}_{Chr})$, respectively, the 3^{rd} column shows the largest X motif with cardinality $Card \geq 10$ trinucleotides and expectation $E < 1$, the 4^{th} and 5^{th} columns indicate the start and end positions of the large X motif, the 6^{th} column gives the length l in trinucleotide, the 7^{th} column indicates its expectation E (Equation 3.6) and the last column shows if the motif is in a coding region (Yes) or non-coding region (No).

$\mathcal{H}_{Chr}$	size (in bases) $N(\mathcal{H}_{Chr})$	X motifs	Start	End	Length (in trinucleotides)	Expectation E	In coding region
1	248956422	GAG,GAG,GAG,CTG,CTG,GCC,CAG,CTG,GAG,GAG,TAC,GAG,CAG,GTC,ATC,CTG,GAC,TTC, CAG,TTC,AAC,CTG,GAG,GCC,ACC	3763375	3763449	25	5.9×10^{-5}	Yes
2	242193529	GTC,GAT,GAG,CAG,AAT,GCC,CAG,ACC,CAG,GAG,CAG,GAG,GGC,TTC,GTC,CTG,GGC,CTC	233449984	233450037	18	2.0×10^{-1}	Yes
4	190214555	GCC,ATC,ATT,ATC,ATT,ATC,ATC,CTC,ACC,TTC,ATC,ATT,AAT,AAC,CTG,GGC,CAG,GGT	42018853	42018906	18	1.5×10^{-1}	No
5	181538259	GAA,ATC,TTC,ATC,ATT,ACC,CTC,ACC,GCC,GCC,ATC,ATT,GAC,CTG,GTT,AAT,GTT	133306903	133306953	17	4.7×10^{-1}	No
7	159345973	ATC,ACC,CAG,GAT,GAA,GAT,GGT,CTC,ACC,CTG,CTC,ATT,GAG,GAT,GCC,GGT,GGT	30452806	30452856	17	4.1×10^{-1}	Yes
8	145138636	ACC,GTC,ACC,AAC,CTG,TTC,ATC,CTC,AAC,CTG,GCC,ATC,GCC,GAC,GAG,CTC,TTC	52940113	52940163	17	3.8×10^{-1}	Yes
9	138394717	GGT,CTC,CAG,GCC,AAT,GTC,ATT,GAC,GTC,ACC,ATC,ATC,GCC,ATC,ACC,ATC,ATT,ACC	95705686	95705739	18	1.1×10^{-1}	No
11	135086622	GAT,GAT,GCC,ACC,ACC,CTC,TAC,CTG,CAG,AAC,AAC,CAG,ATC,AAC,AAC,GCC,GGC,ATC	64116508	64116561	18	1.1×10^{-1}	Yes
13	114364328	AAT,GAG,GAC,ACC,ACC,CAG,GGC,ATC,GCC,AAC,GAG,GAA,GCC,GCC,CAG,GGC,ATC,GCC, GAG,GAC,GCC,ATC,CAG,GGC,ATC,GCC,AAC,GAG,GAG,GTT,GCC,CAG,GGC,ATC,GCC,AAT	18235684	18235791	36	7.5×10^{-11}	No
14	107043718	GCC,CAG,GAC,GAC,GAG,GGT,CTG,CTG,GAC,AAC,TTC,GTC,ACC,TTC,TTC,ATT	99716146	99716193	16	8.9×10^{-1}	Yes
15	101991189	GGC,GAA,GAA,GGT,GAA,GAT,GAA,GAG,GAT,GAA,GAT,CTG,GCC,CTC,GGT,GAC,CAG,GTA	68208355	68208408	18	8.2×10^{-2}	Yes
17	83257441	CTG,CTG,GTT,GAA,GTT,GTC,AAT,GAT,GAC,GCC,AAT,GAA,GAG,GTT,GAG,GGT,GAA,GAA	63944680	63944733	18	6.7×10^{-2}	Yes
18	80373285	ATC,GAG,CAG,AAT,GCC,ACC,AAC,ACC,TTC,CTG,GTC,TAC,ACC,GAG,GAG,GAC	49583566	49583613	16	6.6×10^{-1}	Yes
19	58617616	GAA,ACC,AAC,CAG,GTC,CTC,ATC,AAC,ATT,GGC,CTG,CTG,CTC,CTG,GCC,TTC	13959991	13960038	16	4.8×10^{-1}	Yes
20	64444167	TAC,CTG,GCC,CAG,GTC,CAG,GGT,GAC,GTT,GAC,CTC,GTT,GTA,CTC,CAG,GCC	62362396	62362443	16	5.3×10^{-1}	No
22	50818468	CAG,GTT,GAA,GAA,GTT,GTA,GTT,GCC,GGT,GAT,GAT,AAT,CAG,GAC,CTG,CAG,CAG	50505760	50505810	17	1.3×10^{-1}	Yes
X	156040895	CTC,CAG,GTA,GAG,GGC,ATT,GAG,CAG,CTC,AAT,GAT,GTC,AAC,GAG,GAC,CTG,GTT,GTC	39981361	39981414	18	1.3×10^{-1}	No

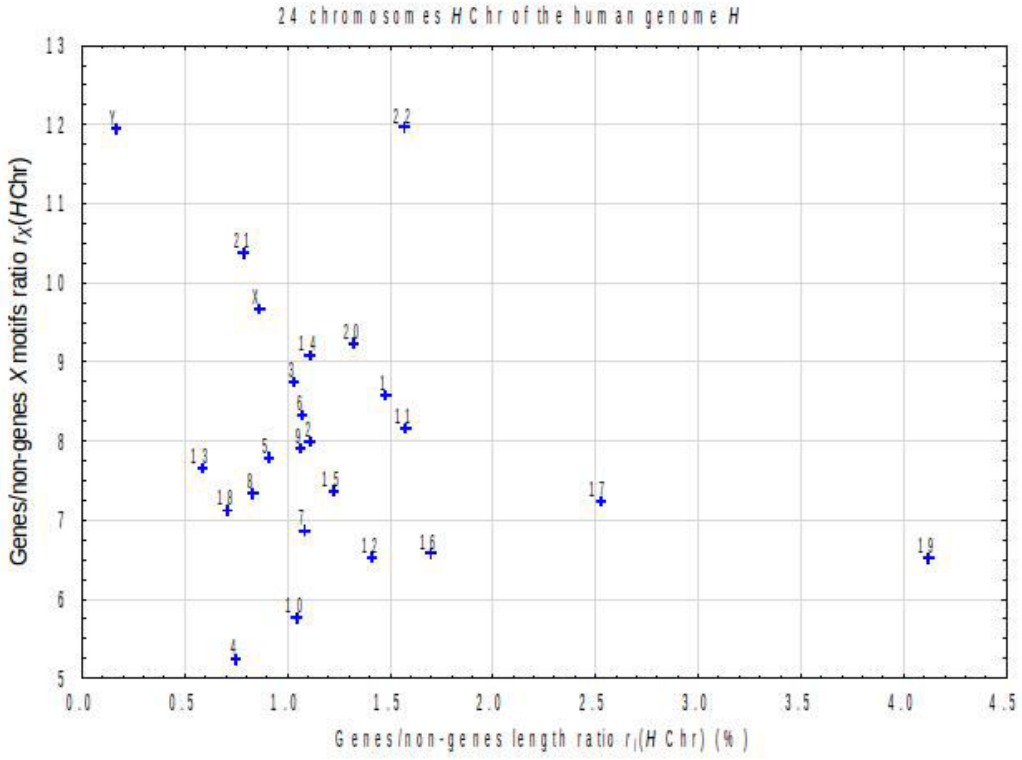

Figure 4.18: Base ratio $r(\mathcal{G}_{Chr})$ (Equation 3.7 in %) of coding/non-coding regions and base ratio $r_{m(X)}(\mathcal{H}_{Chr})$ (Equation 3.9) of X motifs of length $l \geq 10$ trinucleotides and cardinality (composition) $Card \geq 5$ in coding/non-coding regions of the 23 chromosomes $\mathcal{H}_{Chr}$ in the human genome $\mathcal{G} = \mathcal{H} =$ *Homo sapiens*. There is no correlation between $r(\mathcal{G}_{Chr})$ and $r_{m(X)}(\mathcal{H}_{Chr})$.

4.3.4 Largest X motifs in *Homo sapiens*

Table 4.29 shows the largest X motifs $m_{\mathcal{H}_{Chr}}(X)$ with cardinality (composition) $Card \geq 10$ trinucleotides (Equation 3.4) and expectation $E < 1$ (Equation 3.6) in the chromosomes $\mathcal{H}_{Chr}$ of the human genome $\mathcal{G} = \mathcal{H} =$*Homo sapiens*. The largest X motif $m_{\mathcal{H}_{13}}(X)$ is found in a non-coding region of the human chromosome $Chr = 13$. it has a length of $l = 36$ trinucleotides and an expectation $E\big[N\big(m_{Solmo_{15}}(X)\big)\big] = 7.5 \times 10^{-11}$ (Equation 3.6).

4.3.5 X motifs in coding regions versus non-coding regions in eukaryotic genomes

The maximal C^3 self-complementary trinucleotide circular code X is a well-known coding property of genes. Indeed, it is observed in genes of bacteria, eukaryotic, plasmids and viruses (Arquès and Michel, 1996; Michel, 2015).

Table 4.30 gives the base ratio $r(\mathcal{G})$ (Equation 3.7 in %) of coding/non-coding

Table 4.30: Base ratio $r(\mathcal{G})$ (Equation 3.7 in %) of coding/non-coding regions and base ratio $r_{m(X)}(\mathcal{G})$ (Equation 3.9) of X motifs $m(X)$ of length $l \geq 10$ trinucleotides and cardinality (composition) $Card \geq 5$ trinucleotides (Equation 3.8) in the 138 eukaryotic genomes $\mathcal{G}$.

Genome $\mathcal{G}$	$r(\mathcal{G})$ (%)	$r_{m(X)}(\mathcal{G})$	*Genome* $\mathcal{G}$	$r(\mathcal{G})$ (%)	$r_{m(X)}(\mathcal{G})$	*Genome* $\mathcal{G}$	$r(\mathcal{G})$ (%)	$r_{m(X)}(\mathcal{G})$
Anolis carolinensis	1.6	5.3	*Esox lucius*	5.5	12.2	*Ovis aries*	1.3	20.0
Anopheles gambiae	8.6	15.6	*Felis catus*	1.4	19.4	*Pan paniscus*	1.1	9.4
Apis mellifera	8.2	3.5	*Ficedula albicollis*	2.5	19.6	*Pan troglodytes*	1.1	8.8
Arabidopsisthaliana	38.6	5.4	*Fragaria vesca*	18.5	5.1	*Papio anubis*	1.3	7.5
Aspergillus fumigatus	93.7	8.7	*Gallus gallus*	2.9	16.6	*Phaeodactylum tricornutum*	114.7	3.0
Babesia bigemina	196.0	5.2	*Glycine max*	6.9	4.5	*Phaseolus vulgaris*	7.2	4.1
Babesia bovis	213.3	9.1	*Gorilla gorilla*	1.2	9.4	*Plasmodium cynomolgi*	69.9	5.6
Babesia microti	263.9	5.4	*Gossypium raimondii*	6.3	6.0	*Plasmodium falciparum*	111.1	79.3
Beta vulgaris	7.0	4.0	*Homo sapiens*	1.2	8.4	*Plasmodium knowlesi*	90.1	15.6
Bombus terrestris	8.1	3.8	*Kazachstania africana*	239.2	8.4	*Plasmodium vivax*	93.1	16.6
Bos taurus	1.3	21.9	*Kluyveromyces lactis*	223.9	9.8	*Poecilia reticulata*	5.9	16.5
Brachypo diumdistachyon	14.0	6.6	*Komagataella phaffii*	358.7	6.4	*Pongo abelii*	1.1	9.0
Brassica napus	13.9	4.8	*Lachancea thermotolerans*	260.5	16.4	*Populus trichocarpa*	13.2	2.8
Brassica oleracea	12.9	5.0	*Leishmania braziliensis*	94.8	9.3	*Prunus mume*	17.3	5.6
Brassica rapa	23.1	6.4	*Leishmania donovani*	82.2	6.8	*Rattus norvegicus*	1.4	10.2
Caenorhabditis briggsae	28.9	6.5	*Leishmania infantum*	95.0	8.8	*Saccharomyces cerevisiae*	257.2	15.2
Caenorhabditis elegans	36.1	6.4	*Leishmania major*	91.6	8.2	*Salmo salar*	3.3	11.7
Callithrix jacchus	1.2	8.8	*Leishmania mexicana*	96.5	7.8	*Scheffersomyces stipitis*	125.3	3.3
Camelina sativa	19.7	4.3	*Leishmania panamensis*	90.1	8.1	*Schizosaccharomyces pombe*	131.4	6.0
Candida dubliniensis	156.4	3.4	*Lepisosteus oculatus*	3.7	21.9	*Sesamum indicum*	15.1	5.0
Candida glabrata	179.8	9.3	*Macaca fascicularis*	1.2	7.6	*Setaria italica*	9.8	7.4
Candida orthopsilosis	202.1	6.9	*Macaca mulatta*	1.2	7.6	*Solanum lycopersicum*	4.4	4.1
Canis lupus	1.5	13.0	*Magnaporthe oryzae*	70.0	11.8	*Solanum pennellii*	3.9	3.3
Capra hircus	1.2	20.7	*Malus domestica*	7.4	7.1	*Sorghum bicolor*	6.0	6.7
Chlorocebus sabaeus	1.3	7.3	*Medicago truncatula*	14.2	4.6	*Sus scrofa*	1.2	26.4
Chrysemys picta	1.3	8.8	*Meleagris gallopavo*	2.7	14.6	*Taeniopygia guttata*	2.4	18.9
Cicer arietinum	9.0	4.5	*Micromonas sp.*	228.4	2.4	*Takifugu rubripes*	11.3	10.4
Ciona intestinalis	24.8	6.6	*Microtus ochrogaster*	1.5	15.8	*Tetrapisispora blattae*	165.4	7.7
Citrus sinensis	13.0	3.6	*Monodelphis domestica*	1.0	4.5	*Tetrapisispora phaffii*	197.6	9.5
Cryptococcus gattii	124.6	5.4	*Mus musculus*	1.3	12.5	*Thalassiosira pseudonana*	119.0	1.5
Cryptococcus neoformans	115.2	6.6	*Myceliophthora thermophila*	57.4	18.5	*Theileria annulata*	266.6	8.0
Cryptosporidium parvum	298.9	4.6	*Nasonia vitripennis*	13.6	11.4	*Theileria equi*	223.3	3.1
Cucumis sativus	15.2	6.5	*Naumovozyma castellii*	286.4	8.6	*Theileria orientalis*	216.1	2.7
Cyanidioschyzon merolae	81.5	3.7	*Naumovozyma dairenensis*	175.6	6.9	*Theileria parva*	215.3	5.1
Cynoglossus semilaevis	9.0	7.2	*Neospora caninum*	44.8	11.8	*Theobroma cacao*	11.6	6.7
Danio rerio	3.4	7.0	*Neurospora crassa*	58.1	11.5	*Thielavia terrestris*	58.4	14.6
Debaryomyces hansenii	288.2	7.7	*Nomascus leucogenys*	1.2	8.5	*Torulaspora delbrueckii*	367.6	6.6
Dictyostelium discoideum	161.8	13.3	*Ogataea parapolymorpha*	545.6	7.6	*Tribolium castaneum*	11.4	1.2
Drosophila melanogaster	17.4	11.1	*Oreochromis niloticus*	5.7	14.8	*Trypanosoma brucei*	150.1	2.1
Drosophila pseudoobscura	23.9	7.6	*Ornithorhynchus anatinus*	1.1	3.2	*Ustilago maydis*	156.3	8.6
Drosophila simulans	14.7	13.6	*Oryctolagus cuniculus*	1.1	19.2	*Vigna radiata*	9.8	4.9
Drosophila yakuba	20.3	10.5	*Oryza brachyantha*	12.8	12.4	*Vitis vinifera*	8.0	6.0
Elaeis guineensis	4.3	11.2	*Oryza sativa*	8.7	8.1	*Yarrowia lipolytica*	85.2	14.9
Equus caballus	1.4	18.1	*Oryzias latipes*	4.9	18.2	*Zea mays*	2.2	4.0
Eremothecium cymbalariae	202.6	3.1	*Ostreococcus lucimarinus*	231.1	1.3	*Zygosaccharomyces rouxii*	319.5	3.6
Eremothecium gossypii	335.6	15.8	*Ostreococcus tauri*	437.3	1.6	*Zymoseptoria tritici*	56.8	4.0
						Mean	79.2	9.3
						Median	15.2	7.6

regions and base ratio $r_{m(X)}(\mathcal{G})$ (Equation 3.9) of X motifs $m(X)$ of length $l \geq 10$ trinucleotides and cardinality (composition) $Card \geq 5$ trinucleotides (Equation 3.8) in coding/non-coding regions of the 138 eukaryotic genomes $\mathcal{G}$.

The lowest value $r(\mathcal{G})$ of coding/non-coding regions is observed in the genome of $\mathcal{G}$ =*Monodelphis domestica* with $r(\mathcal{G}) = 1.0\%$ ($r_{m(X)}(\mathcal{G}) = 4.5$). The highest value is found in the genome of $\mathcal{G}$ =*Ogataea parapolymorpha* with $r(\mathcal{G}) = 545.6\%$ ($r_{m(X)}(\mathcal{G}) = 7.6$). The mean value is $\overline{r}(\mathcal{G}) = 79.2\%$ and the median value $\widetilde{r}(\mathcal{G}) = 15.2\%$.

The lowest value $r_{m(X)}(\mathcal{G})$ of coding/non-coding regions is observed in the genome of $\mathcal{G}$ =*Tribolium castaneum* with $r_{m(X)}(\mathcal{G}) = 1.2$ ($r(\mathcal{G}) = 11.4\%$). The highest value is found in the genome of $\mathcal{G}$ =*Plasmodium falciparum* with $r_{m(X)}(\mathcal{G}) = 79.3$ ($r(\mathcal{G}) = 111.1\%$). The mean value is $\overline{r}(\mathcal{G}) = 9.3$ and the median value $\widetilde{r}(\mathcal{G}) = 7.6$.

Figure 4.17 gives a graphical representation of Table 4.30. No correlation was found between $r(\mathcal{G})$ and $r_{m(X)}(\mathcal{G})$, with sample correlation coefficient of -0.12.

Thus, as expected according to previous works, the X motifs occur preferentially in genes of genomes with a factor of about 8 ($\widetilde{r}_{m(X)}(\mathcal{G}) = 7.6 < 8 < \overline{r}_{m(X)}(\mathcal{G}) = 9.3$). Furthermore, this circular code property is verified whatever the base content of coding regions in the genomes, with sample correlation coefficient of $r = -0.12$.

Table 4.31: Base ratio $r(\mathcal{H}_{Chr})$ (Equation 3.7 in %) of coding/non-coding regions and base ratio $r_{m(X)}(\mathcal{G})$ (Equation 3.9) of X motifs $m(X)$ of length $l \geq 10$ trinucleotides and cardinality (composition) $Card \geq 5$ trinucleotides (Equation 3.8) in the 24 chromosomes $\mathcal{H}_{Chr}$ of the human genome $\mathcal{G} = \mathcal{H}$ = *Homo sapiens*.

$\mathcal{H}_{Chr}$	$r(\mathcal{H}_{Chr})$ (%)	$r_{m(X)}(\mathcal{G})$	$\mathcal{H}_{Chr}$	$r(\mathcal{H}_{Chr})$ (%)	$r_{m(X)}(\mathcal{G})$	$\mathcal{H}_{Chr}$	$r(\mathcal{H}_{Chr})$ (%)	$r_{m(X)}(\mathcal{G})$
1	1.5	8.6	*9*	1.1	7.9	*17*	2.5	7.2
2	1.1	8.0	*10*	1.1	5.8	*18*	0.7	7.1
3	1.0	8.7	*11*	1.6	8.2	*19*	4.1	6.5
4	0.8	5.2	*12*	1.4	6.5	*20*	1.3	9.2
5	0.9	7.8	*13*	0.6	7.7	*21*	0.8	10.4
6	1.1	8.3	*14*	1.1	9.1	*22*	1.6	12.0
7	1.1	6.9	*15*	1.2	7.4	*X*	0.9	9.7
8	0.8	7.3	*16*	1.7	6.6	*Y*	0.2	11.9
						Mean	1.3	8.1
						Median	1.1	7.8

4.3.6 X motifs in coding regions versus non-coding regions in *Homo sapiens*

Table 4.31 gives the base ratio $r(\mathcal{H}_{Chr})$ (Equation 3.7 in %) of coding/non-coding region and the base ratio $r_{m(X)}(\mathcal{H}_{Chr})$ (Equation 3.9) of X motifs of length $l \geq 10$

trinucleotides and cardinality (composition) $Card \geq 5$ trinucleotides (Equation 3.8) in coding/non-coding regions of the 24 chromosomes $\mathcal{H}_{Chr}$ in the human genome $\mathcal{G} = \mathcal{H} = \textit{Homo sapiens}$.

The lowest value $r(\mathcal{H}_{Chr})$ of coding/non-coding regions is observed in chromosome $Chr = Y$ with $r(\mathcal{H}_Y) = 0.2\%$ ($r_{m(X)}(\mathcal{H}_Y) = 11.9$). While the highest value is found in chromosome $Chr = 19$ with $r(\mathcal{H}_{19}) = 4.1\%$ ($r_{m(X)}(\mathcal{H}_{19}) = 6.5$). The mean value is $\overline{r}(\mathcal{H}_{Chr}) = 1.3\%$ and the median value $\widetilde{r}(\mathcal{H}_{Chr}) = 1.1\%$.

Remark 4.1. These two values $\overline{r}(\mathcal{H}_{Chr}) = 1.3\%$ and $\widetilde{r}(\mathcal{H}_{Chr}) = 1.1\%$. are very close to $r(\mathcal{H}) = 1.2\%$ (Table 4.30).

The lowest value $r_{m(X)}(\mathcal{H}_{Chr})$ of X motifs in coding/non-coding regions is observed in chromosome $Chr = 4$ with $r_{m(X)}(\mathcal{H}_4) = 5.2$ ($r(\mathcal{H}_4) = 0.8\%$). While the highest value is found in chromosome $Chr = 22$ with $r_{m(X)}(\mathcal{H}_{22}) = 12.0$ ($r(\mathcal{H}_{22}) = 1.6\%$). The mean value is $r_{m(X)}(\mathcal{H}_{Chr}) = 8.1$ and the median value $\overline{r}_{m(X)}(\mathcal{H}_{Chr}) = 7.8$.

Remark 4.2. These two values $\overline{r}_{m(X)}(\mathcal{H}_{Chr}) = 8.1$ and $\widetilde{r}_{m(X)}(\mathcal{H}_{Chr}) = 7.8$. are very close to $r_{m(X)}(\mathcal{H}) = 8.4$ (Table 4.30).

Table 4.18 gives a graphical representation of Table 4.31. There is no correlation found between $r(\mathcal{H}_{Chr})$ and $r_{m(X)}(\mathcal{H}_{Chr})$, with sample correlation coefficient of $r = -0.26$.

As in the general case, the X motifs occur preferentially in coding regions of human chromosomes with a factor of about 8 ($\widetilde{r}_{m(X)}(\mathcal{H}_{Chr}) = 7.8 < 8 < \overline{r}_{m(X)}(\mathcal{H}_{Chr}) = 8.1$). Furthermore, this circular code property is also verified whatever the base content of genes in human chromosomes, with sample correlation coefficient of $r = -0.26$.

4.4 Analysis of unitary circular code motifs in eukaryotic genomes

As was shown in sections 2.3.2, 2.3.2 and 2.3.2, the unitary circular code is closely related to simple repeats. The following results were generated using the algorithm described in section 3.5.4.1 with the eukaryotic genomes data (Table 3.2)

4.4.1 Occurrence of repeated dinucleotides in eukaryotic genomes

The repeated dinucleotides (Section 2.3.2) are generated from the unitary circular codes of dinucleotides (Section 3.5.1). Figures 4.19 and 4.20 give the occurrence number $N(di^+)$ (Equation 3.13) and the base number $B(di^+)$ (Equation 3.14) of all the repeated dinucleotides di^n of length $l = 2n \geq 30$ nucleotides ($n \geq 15$) in the genomes of eukaryotes. The results in the two Figures 4.19 and 4.20 are altogether consistent. The repeats $(AC)^+$ and $(GT)^+ = (\mathcal{C}(AC))^+$ have the highest occurrences in the eukaryotic genomes. Then, the repeat $(AT)^+$ (note that $\mathcal{C}(AT) = AT$) has a lower occurrence. The repeats $(AG)^+$ and $(CT)^+ = (\mathcal{C}(AG))^+$ have occurrences lower than $(AT)^+$. The repeat $(CG)^+$ (note that $\mathcal{C}(CG) = CG$) is almost absent. A repeated dinucleotide di^+ and its complementary repeated dinucleotide $(\mathcal{C}(di))^+$ have the same occurrences in the eukaryotic genomes: $N((AC)^+) \approx N((GT)^+) \approx 659400$, $B((AC)^+) \approx B((GT)^+) \approx 28800$ kb, $N((AG)^+) \approx N((CT)^+) \approx 299900$ and $B((AG)^+) \approx B((CT)^+) \approx 13500$ kb (Figures 4.19, 4.20). This property is related to the complementary property of the DNA double helix.

4.4.2 Occurrence of repeated trinucleotides in eukaryotic genomes

The repeated trinucleotides (Section 2.3.2) are generated from the unitary circular codes of trinucleotides (Section 3.5.2). Figures 4.21 and 4.22 give the occurrence number $N(tri^+)$ (Equation 3.13) and the base number $B(tri^+)$ (Equation 3.14) of all the repeated trinucleotides tri^n of length $l = 3n \geq 30$ nucleotides ($n \geq 10$) in the genomes of eukaryotes. Again, the results in the two Figures 4.21 and 4.22 are altogether consistent. The repeats $(AAT)^+$ and $(ATT)^+ = (\mathcal{C}(AAT))^+$ have the highest occurrences in the eukaryotic genomes. Then, the following repeats are observed by decreasing order of occurrence: $(AAG)^+$ and $(CTT)^+ = (C(AAG))^+$, $(AAC)^+$ and $(GTT)^+ = (C(AAC))^+$, $(ATC)^+$ and $(ATG)^+ = (\mathcal{C}(\mathcal{P}^2(ATC)))^+$ (i.e. $(ATG)^+$ and $(GAT)^+ = (\mathcal{C}(ATC))^+$ belong the same equivalence class by the circular permutation map $\mathcal{P}$), $(AGG)^+$ and $(CCT)^+ = (\mathcal{C}(AGG))^+$, $(AGC)^+$ and $(CTG)^+ = (\mathcal{C}(\mathcal{P}^2(AGC)))^+$ (i.e. $(CTG)^+$ and $(GCT)^+ = (\mathcal{C}(AGC))^+$ belong the same equivalence class), $(ACT)^+$ and $(AGT)^+ = (\mathcal{C}(ACT))^+$, and $(ACC)^+$ and $(GGT)^+ = (\mathcal{C}(ACC))^+$. The repeats $(ACG)^+$ and $(CGT)^+ = (\mathcal{C}(ACG))^+$, and $(CCG)^+$ and $(CGG)^+ = (\mathcal{C}(CCG))^+$

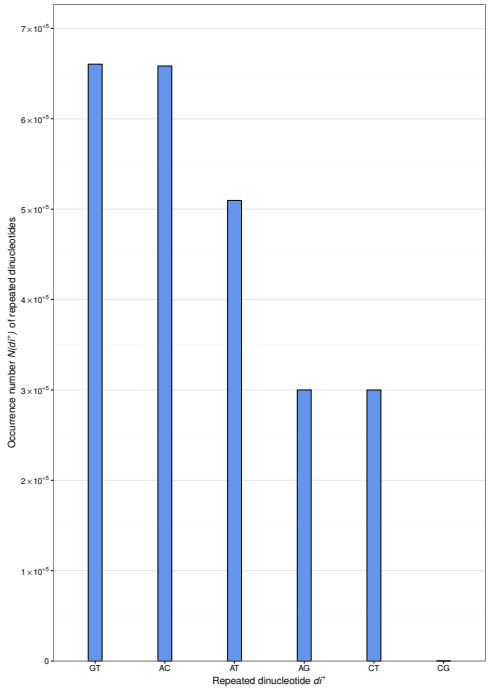

Figure 4.19: Occurrence number $N(di^+)$ (Equation 3.13) (descending order) of all the repeated dinucleotides di^n of length $l = 2n \geq 30$ nucleotides ($n \geq 15$) in the eukaryotic genomes. The repeated dinucleotides are generated from the unitary circular codes of dinucleotides.

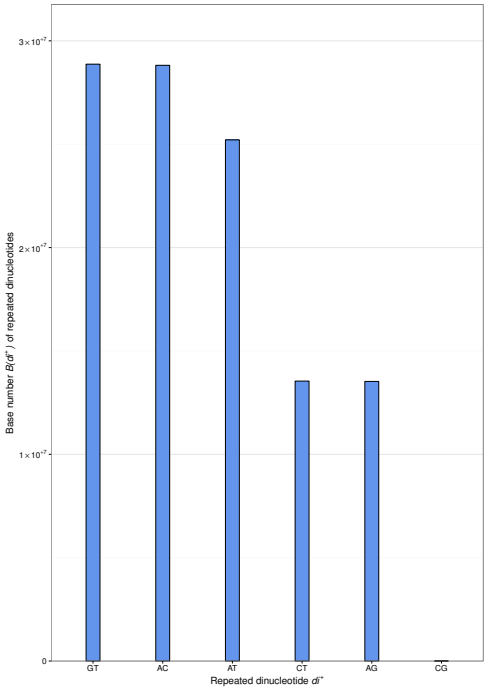

Figure 4.20: Base number $B(di^+)$ (Equation 3.14) (descending order) of all the repeated dinucleotides di^n of length $l = 2n \geq 30$ nucleotides ($n \geq 15$) in the eukaryotic genomes. The repeated dinucleotides are generated from the unitary circular codes of dinucleotides.

are almost absent.

A repeated trinucleotide tri^+ and its complementary repeated trinucleotide $(\mathcal{C}(tri))^+$ have the same occurrences in the eukaryotic genomes: $N((AAT)^+) \approx N((ATT)^+) \approx 95700$, $B((AAT)^+) \approx B((ATT)^+) \approx 4500$ kb, $N((AAG)^+) \approx N((CTT)^+) \approx 24500$ and $B((AAG)^+) \approx B((CTT)^+) \approx 1600$ kb, etc. (Figures 4.21 and 4.22). This property is again related to the complementary property of the DNA double helix. This result is also confirmed by the correlation matrix of the base number $B(tri^+)$ of all the repeated trinucleotides tri^+ (Table 4.33). The highest correlation is always observed between the repeated trinucleotides tri^+ and $(\mathcal{C}(tri))^+$ in the eukaryotic genomes. There is no significant correlation between a repeated trinucleotide tri^+ and the size of genomes as well as the count of A, C, G and T and GC content of genomes.

A second property is identified with the repeated trinucleotides tri^+ and $(\mathcal{C}(tri))^+$. Indeed, the repeated trinucleotides tri^+ and $(\mathcal{C}(tri))^+$ have increasing occurrences in the eukaryotic genomes conversely to their number of hydrogen

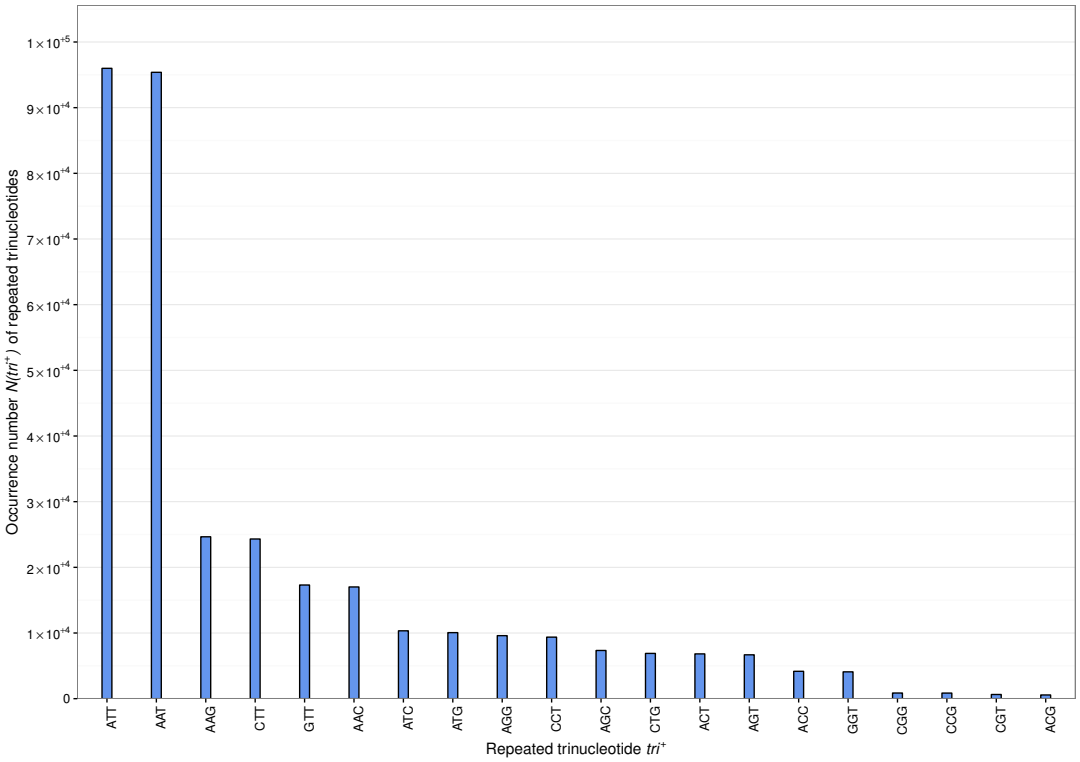

Figure 4.21: Occurrence number $N(tri^+)$ (Equation 3.13) (descending order) of all the repeated trinucleotides tri^n of length $l = 3n \geq 30$ nucleotides ($n \geq 10$) in the eukaryotic genomes. The repeated trinucleotides are generated from the unitary circular codes of trinucleotides.

bonds (two hydrogen bonds between A and T $= \mathcal{C}$(A) and three hydrogen bonds between C and G $= \mathcal{C}$(C)), from the highest occurrences for the two repeats $(AAT)^+$ and $(ATT)^+$ with a total of six hydrogen bonds to the lowest occurrences for the two repeats $(CCG)^+$ and $(CGG)^+$ with a total of nine hydrogen bonds.

4.4.3 Occurrence of repeated tetranucleotides in eukaryotic genomes

The repeated tetranucleotides (Section 3.5.3) are generated from the unitary circular codes of tetranucleotides (Section 2.3.2). Figures 4.23 and 4.24 give the occurrence number $N(tetra^+)$ (Equation 3.13) and the base number $B(tetra^+)$ (Equation 3.14) of the repeated tetranucleotides $tetra^n$ of length $l = 4n \geq 28$ nucleotides ($n \geq 7$) in the eukaryotic genomes. The results in the two Figures 4.23 and 4.24 are consistent and identify two classes of repeated tetranucleotides with higher occurrences. The 1st class with the highest occurrences in the eukaryotic genomes contains eight repeated tetranucleotides, by descending

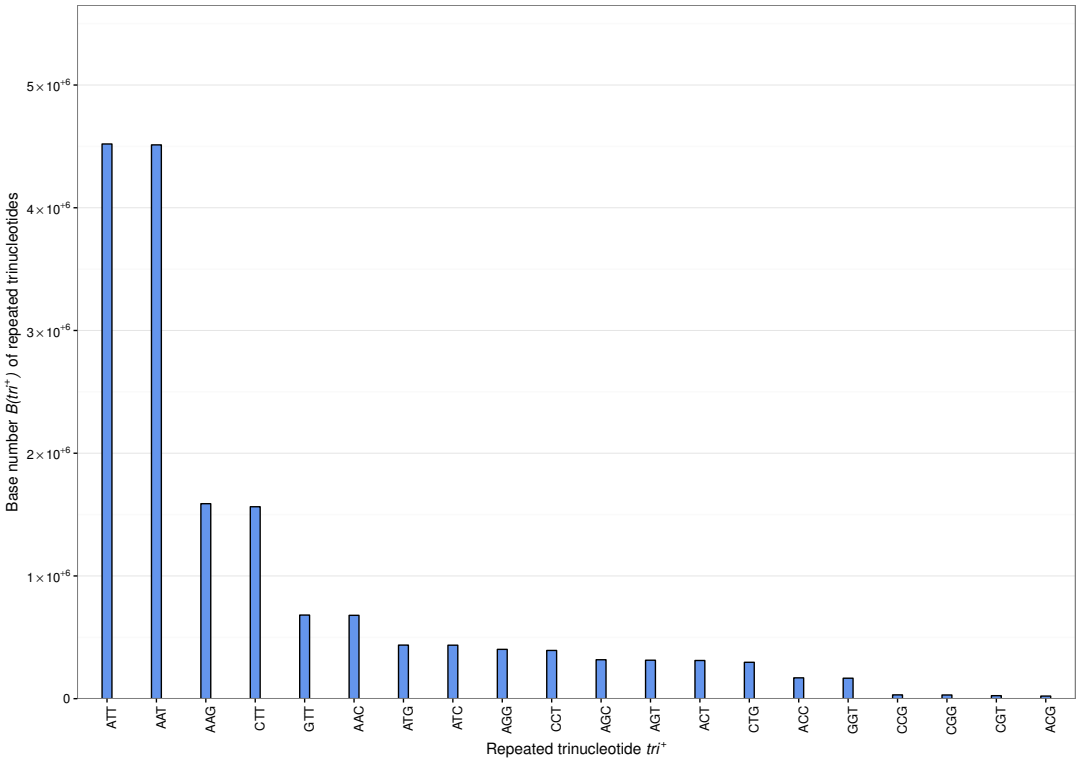

Figure 4.22: Base number $N(tri^{+})$ (Equation 3.14) (descending order) of all the repeated trinucleotides tri^{n} of length $l = 3n \geq 30$ nucleotides ($n \geq 10$) in the eukaryotic genomes. The repeated trinucleotides are generated from the unitary circular codes of trinucleotides.

order: $(AAAT)^{+}$ and $(ATTT)^{+} = (\mathcal{C}(AAAT))^{+}$, $(AAAG)^{+}$ and $(CTTT)^{+} = (\mathcal{C}(AAAG))^{+}$, $(AGAT)^{+}$ and $(ATCT)^{+} = (\mathcal{C}(AGAT))^{+}$, and $(AAGG)^{+}$ and $(CCTT)^{+} = (\mathcal{C}(AAAG))^{+}$ (Figure 4.23). Note that this repeated tetranucleotide order is different in Figure 4.24. The 2nd class with higher occurrences in the eukaryotic genomes contains 12 repeated tetranucleotides, by descending order: $(ATCC)^{+}$ and $(ATGG)^{+} = (\mathcal{C}(\mathcal{P}^{2}(ATCC)))$, $(AAAC)^{+}$ and $(GTTT)^{+} = (\mathcal{C}(AAAC))^{+}$, $(ACAG)^{+}$ and $(CTGT)^{+} = (\mathcal{C}(ACAG))^{+}$, $(ACAT)^{+}$ and $(ATGT)^{+} = (\mathcal{C}(ACAT))^{+}$, $(AATG)^{+}$ and $(ATTC)^{+} = (\mathcal{C}(\mathcal{P}^{3}(AATG)))$, and $(AGGG)^{+}$ and $(CCCT)^{+} = (\mathcal{C}(AGGG))^{+}$ (Figure 4.23). The repeats $(CCGG)^{+}$ (note that $\mathcal{C}(CCGG) = CCGG$), and $(CCCG)^{+}$ and $(CGGG)^{+} = (\mathcal{C}(CCCG))^{+}$ are almost absent (results not shown). A repeated tetranucleotide $tetra^{+}$ and its complementary repeated tetranucleotide $(\mathcal{C}(tri))^{+}$ also have the same occurrences in the eukaryotic genomes. The repeated tetranucleotides $tetra^{+}$ and $(\mathcal{C}(tetra))^{+}$ also have increasing occurrences conversely to their number of hydrogen bonds, from the highest occurrences for the two repeats $(AAAT)^{+}$ and $(ATTT)^{+}$ with a total of eight hydrogen bonds to the lowest occurrences for the two repeats

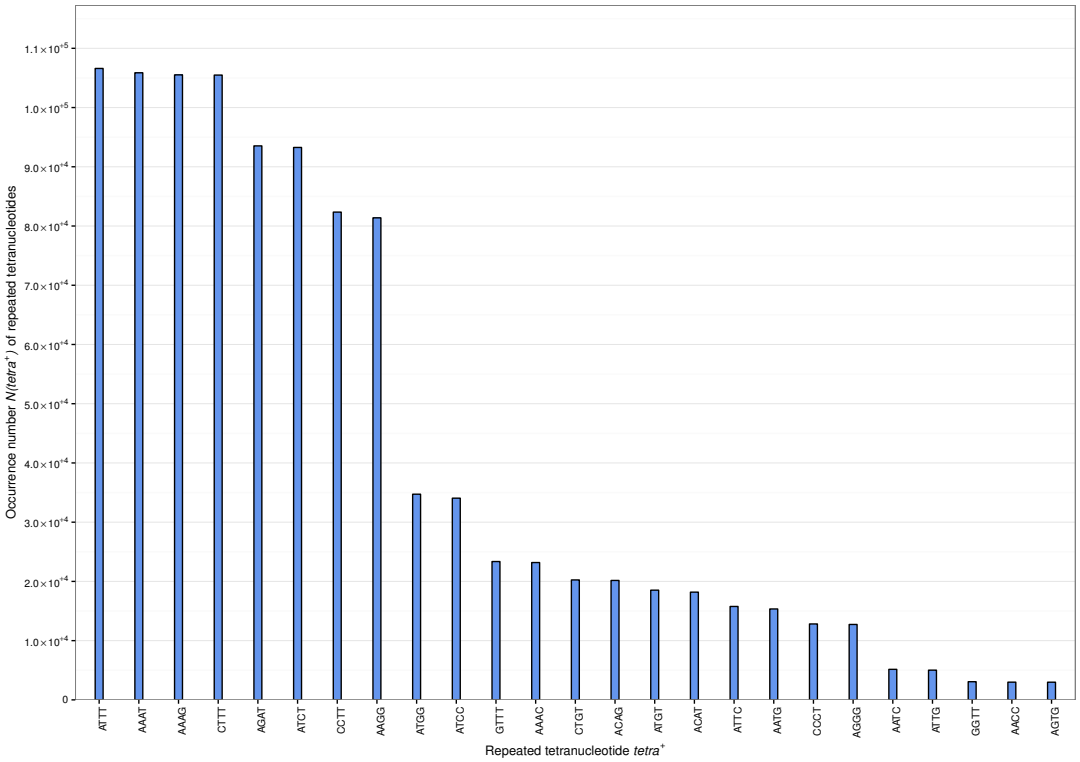

Figure 4.23: Occurrence number $N(tetra^+)$ (Equation 3.13) (descending order, showing first 25) of the repeated tetranucleotides $tetra^n$ of length $l = 4n \geq 28$ nucleotides ($n \geq 10$) in the eukaryotic genomes. The repeated tetranucleotides are generated from the unitary circular codes of tetranucleotides.

$(CCCG)^+$ and $(CGGG)^+$ with a total of 12 hydrogen bonds.

4.4.4 Largest repeated motifs in eukaryotic genomes

Table 4.32 shows the largest nucleotide lengths $l = 2n$ for the six repeated dinucleotides di^n, $l = 3n$ for the 20 repeated trinucleotides tri^n and $l = 4n$ for the 10 largest repeated tetranucleotides tri^n in the eukaryotic genomes. The largest repeated dinucleotide $(AT)^n$ of length $l = 11254$ nucleotides is observed in the chromosome 1 of *Medicago truncatula*. The largest repeated trinucleotide $(ATT)^n$ of length $l = 19275$ nucleotides is found in the chromosome 9 of *Citrus sinensis*. The largest repeated tetranucleotide $(ATCC)^n$ of length $l = 6952$ nucleotides is present in the chromosome 11 of *Solanum pennellii*.

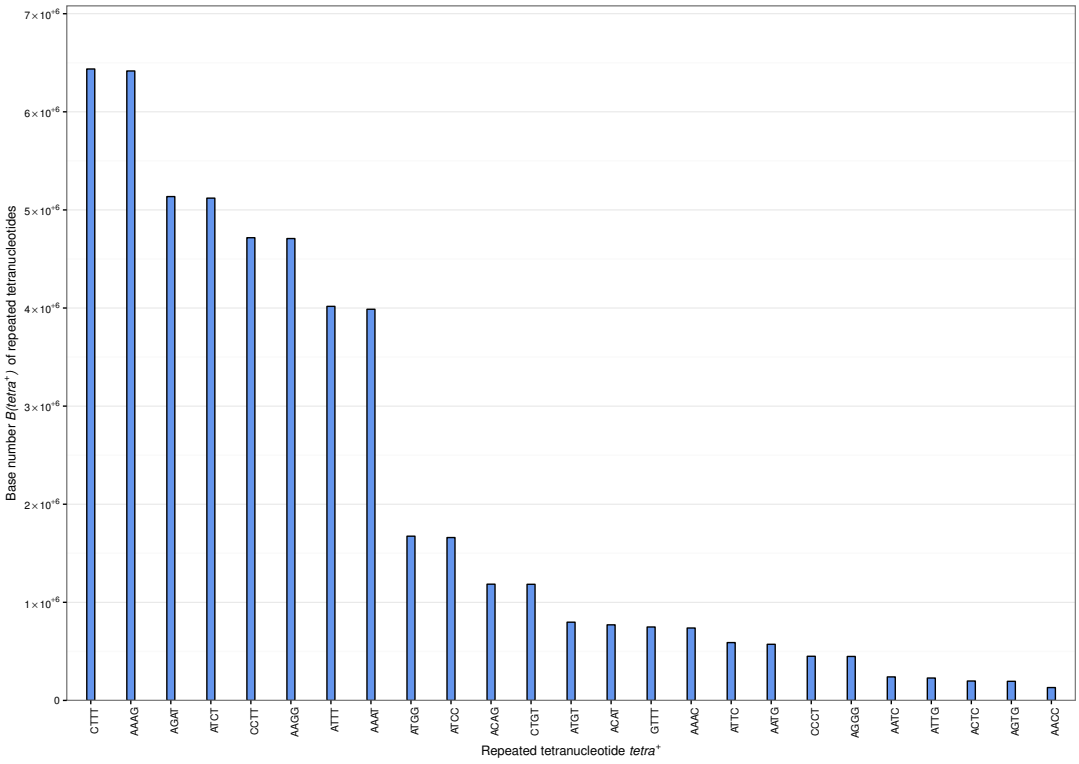

Figure 4.24: Base number $N(tetra^{+})$ (Equation 3.14) (descending order, showing first 25) of the repeated tetranucleotides $tetra^{n}$ of length $l = 4n \geq 28$ nucleotides ($n \geq 7$) in the eukaryotic genomes. The repeated tetranucleotides are generated from the unitary circular codes of tetranucleotides.

Table 4.32: Largest repeated motifs in nucleotides for the 6 repeated dinucleotides di^{n}, for the 20 repeated trinucleotides tri^{n} and for the 10 largest repeated tetranucleotides $tetra^{n}$ in the eukaryotic genomes. The 1st and 5th columns indicate the unitary circular code (UCC) motif, the 2nd and 6th columns mention the genome $\mathcal{G}$ while its chromosome number $\mathcal{G}_{Chr}$ are found in 3rd and 7th columns, respectively, and the 4th and 8th columns gives the length in nucleotides of the UCC motif.

UCC motif	*Genome* $\mathcal{G}$	$\mathcal{G}_{chr}$	Length (in bases)	UCC motif	*Genome* $\mathcal{G}$	$\mathcal{G}_{chr}$	Length (in bases)
$(AC)^n$	*Citrus sinensis*	7	4500	$(CCG)^n$	*Oryza brachyantha*	9	555
$(AG)^n$	*Citrus sinensis*	3	7648	$(CCT)^n$	*Ficedula albicollis*	14	1065
$(AT)^n$	*Medicago truncatula*	1	11254	$(CGG)^n$	*Oryza brachyantha*	7	210
$(CG)^n$	*Cucumis sativus*	7	432	$(CGT)^n$	*Solanum pennellii*	8	1815
$(CT)^n$	*Citrus sinensis*	5	7232	$(CTG)^n$	*Ficedula albicollis*	12	723
$(GT)^n$	*Beta vulgaris subsp. vulgaris*	5	2680	$(CTT)^n$	*Cicer arietinum*	Ca6	4263
$(AAC)^n$	*Solanum pennellii*	9	8655	$(GGT)^n$	*Homo sapiens*	2	630
$(AAG)^n$	*Solanum pennellii*	12	10536	$(GTT)^n$	*Cicer arietinum*	Ca8	4239
$(AAT)^n$	*Citrus sinensis*	4	12951	$(ATCC)^n$	*Solanum pennellii*	11	6952
$(ACC)^n$	*Oryza brachyantha*	11	363	$(ATCT)^n$	*Solanum pennellii*	6	5492
$(ACG)^n$	*Bombus terrestris*	B04	144	$(ATGG)^n$	*Solanum pennellii*	9	5076
$(ACT)^n$	*Solanum pennellii*	12	1728	$(AGAT)^n$	*Solanum pennellii*	10	4904
$(AGC)^n$	*Ficedula albicollis*	21	3555	$(AAAG)^n$	*Ficedula albicollis*	1	4780
$(AGG)^n$	*Ficedula albicollis*	6	822	$(ATTT)^n$	*Cynoglossus semilaevis*	5	4268
$(AGT)^n$	*Zea mays*	10	1926	$(CTTT)^n$	*Ficedula albicollis*	1	4200
$(ATC)^n$	*Camelina sativa*	5	2145	$(AAGG)^n$	*Ficedula albicollis*	15	4048
$(ATG)^n$	*Citrus sinensis*	9	2076	$(CCTT)^n$	*Ficedula albicollis*	1	3668
$(ATT)^n$	*Citrus sinensis*	9	19275	$(AGTG)^n$	*Cicer arietinum*	Ca2	3036

Table 4.33: Correlation matrix of the base number $B(tri^+)$ (Equation 3.14 and Figure 4.24) of all the repeated trinucleotides tri^n of length $l = 3n \geq 30$ nucleotides ($n \geq 10$) in the eukaryotic genomes. The highest correlation (in bold) is always between a repeated trinucleotide tri^+ and its complementary repeated trinucleotide $(\mathcal{C}(tri))^+$ (note that $(ATC)^+$ and $(ATG)^+ = (\mathcal{C}(\mathcal{P}^2(ATC)))^+$, and $(AGC)^+$ and $(CTG)^+ = (\mathcal{C}(\mathcal{P}^2(AGC)))^+$, details in Section 4.4.2). There is no significant correlation between a repeated trinucleotide tri^+ and the size of genomes as well as the count of A, C, G, T and GC content of genomes.

	Size	A count	C count	G count	T count	GC content	AAC	AAG	AAT	ACC	ACG	ACT	AGC	AGG	AGT	ATC	ATG	ATT	CCG	CCT	CGG	CGT	CTG	CTT	GGT	GTT
Size	1	0.11	-0.11	-0.11	0.11	-0.11	0.45	0.43	0.17	0.3	0.08	0.38	0.41	0.35	0.38	0.37	0.38	0.17	0.56	0.35	0.5	0.07	0.4	0.43	0.32	0.46
A count	0.11	1	-1	-1	1	-1	0.14	0.07	0.17	0	-0.19	0.05	-0.01	0.02	0.05	0.08	0.1	0.16	-0.03	0.02	-0.03	-0.14	0	0.06	0.01	0.12
C count	-0.11	-1	1	1	-1	1	-0.14	-0.07	-0.17	0	0.19	-0.05	0.01	-0.02	-0.05	-0.08	-0.1	-0.16	0.02	-0.02	0.03	0.14	0	-0.06	-0.01	-0.12
G count	-0.11	-1	1	1	-1	1	-0.14	-0.07	-0.17	0	0.19	-0.05	0.01	-0.02	-0.05	-0.08	-0.1	-0.16	0.03	-0.02	0.03	0.14	0	-0.06	0	-0.12
T count	0.11	1	-1	-1	1	-1	0.14	0.07	0.17	0	-0.19	0.05	-0.01	0.02	0.05	0.08	0.1	0.16	-0.02	0.02	-0.03	-0.14	0	0.06	0	0.12
GC content	-0.11	-1	1	1	-1	1	-0.14	-0.07	-0.17	0	0.19	-0.05	0.01	-0.02	-0.05	-0.08	-0.1	-0.16	0.03	-0.02	0.03	0.14	0	-0.06	0	-0.12
AAC	0.45	0.14	-0.14	-0.14	0.14	-0.14	1	0.57	0.75	0.55	0.19	0.38	0.29	0.58	0.38	0.85	0.86	0.75	0.3	0.57	0.23	0.21	0.29	0.56	0.56	**0.98**
AAG	0.43	0.07	-0.07	-0.07	0.07	-0.07	0.57	1	0.18	0.75	0.29	0.56	0.36	0.92	0.58	0.57	0.59	0.19	0.5	0.92	0.43	0.25	0.37	**1**	0.77	0.57
AAT	0.17	0.17	-0.17	-0.17	0.17	-0.17	0.75	0.18	1	0.11	0	0.27	0.04	0.2	0.27	0.77	0.76	**1**	0.04	0.2	0.02	0.01	0.04	0.18	0.1	0.77
ACC	0.3	0	0	0	0	0	0.55	0.75	0.11	1	0.44	0.43	0.49	0.72	0.43	0.53	0.56	0.11	0.38	0.7	0.26	0.38	0.5	0.74	**1**	0.55
ACG	0.08	-0.19	0.19	0.19	-0.19	0.19	0.19	0.29	0	0.44	1	0.28	0.18	0.28	0.27	0.21	0.21	0	0.19	0.28	0.14	**0.93**	0.18	0.28	0.44	0.18
ACT	0.38	0.05	-0.05	-0.05	0.05	-0.05	0.38	0.56	0.27	0.43	0.28	1	0.25	0.42	**0.99**	0.51	0.52	0.27	0.25	0.42	0.19	0.21	0.25	0.57	0.44	0.4
AGC	0.41	-0.01	0.01	0.01	-0.01	0.01	0.29	0.36	0.04	0.49	0.18	0.25	1	0.37	0.25	0.3	0.32	0.04	0.21	0.36	0.15	0.15	**1**	0.37	0.49	0.29
AGG	0.35	0.02	-0.02	-0.02	0.02	-0.02	0.58	0.92	0.2	0.72	0.28	0.42	0.37	1	0.43	0.5	0.51	0.2	0.39	**1**	0.32	0.22	0.38	0.93	0.73	0.58
AGT	0.38	0.05	-0.05	-0.05	0.05	-0.05	0.38	0.58	0.27	0.43	0.27	**0.99**	0.25	0.43	1	0.52	0.53	0.27	0.25	0.43	0.18	0.21	0.25	0.58	0.44	0.39
ATC	0.37	0.08	-0.08	-0.08	0.08	-0.08	0.85	0.57	0.77	0.53	0.21	0.51	0.3	0.5	0.52	1	**0.99**	0.77	0.28	0.49	0.21	0.19	0.31	0.56	0.54	0.87
ATG	0.38	0.1	-0.1	-0.1	0.1	-0.1	0.86	0.59	0.76	0.56	0.21	0.52	0.32	0.51	0.53	**0.99**	1	0.76	0.28	0.51	0.21	0.22	0.32	0.58	0.56	0.88
ATT	0.17	0.16	-0.16	-0.16	0.16	-0.16	0.75	0.19	**1**	0.11	0	0.27	0.04	0.2	0.27	0.77	0.76	1	0.04	0.2	0.02	0.01	0.04	0.18	0.11	0.78
CCG	0.56	-0.03	0.02	0.03	-0.02	0.03	0.3	0.5	0.04	0.38	0.19	0.25	0.21	0.39	0.25	0.28	0.28	0.04	1	0.39	**0.95**	0.17	0.22	0.5	0.39	0.31
CCT	0.35	0.02	-0.02	-0.02	0.02	-0.02	0.57	0.92	0.2	0.7	0.28	0.42	0.36	**1**	0.43	0.49	0.51	0.2	0.39	1	0.32	0.22	0.37	0.93	0.72	0.58
CGG	0.5	-0.03	0.03	0.03	-0.03	0.03	0.23	0.43	0.02	0.26	0.14	0.19	0.15	0.32	0.18	0.21	0.21	0.02	**0.95**	0.32	1	0.13	0.16	0.42	0.27	0.24
CGT	0.07	-0.14	0.14	0.14	-0.14	0.14	0.21	0.25	0.01	0.38	0.93	0.21	0.15	0.22	0.21	0.19	0.22	0.01	0.17	0.22	0.13	1	0.16	0.24	0.38	0.16
CTG	0.4	0	0	0	0	0	0.29	0.37	0.04	0.5	0.18	0.25	**1**	0.38	0.25	0.31	0.32	0.04	0.22	0.37	0.16	0.16	1	0.37	0.5	0.3
CTT	0.43	0.06	-0.06	-0.06	0.06	-0.06	0.56	**1**	0.18	0.74	0.28	0.57	0.37	0.93	0.58	0.56	0.58	0.18	0.5	0.93	0.42	0.24	0.37	1	0.76	0.56
GGT	0.32	0.01	-0.01	0	0	0	0.56	0.77	0.1	**1**	0.44	0.44	0.49	0.73	0.44	0.54	0.56	0.11	0.39	0.72	0.27	0.38	0.5	0.76	1	0.55
GTT	0.46	0.12	-0.12	-0.12	0.12	-0.12	**0.98**	0.57	0.77	0.55	0.18	0.4	0.29	0.58	0.39	0.87	0.88	0.78	0.31	0.58	0.24	0.16	0.3	0.56	0.55	1

4.4.5 Scarcity of repeated trinucleotides in eukaryotic genomes

Table 4.34: Scarcity of repeated trinucleotides (Tri^+ motifs) in the large genomes of eukaryotes. The 1st and 5th columns mention the 59 eukaryotic genomes $\mathcal{G}$ of large sizes $N(\mathcal{G}) > 300000$ kb, the 2nd 6th, 3th and 7th, and 4th and 8th provide the ratios $r(Di^+, \mathcal{G})$ (‰), $r(Tri^+, \mathcal{G})$ (‰) and $r(Tetra^+, \mathcal{G})$ (‰), respectively, (Equation 3.16) giving the proportion of the total base numbers $B(Di^+, \mathcal{G})$, $B(Tri^+, \mathcal{G})$ and $B(Tetra^+, \mathcal{G})$, respectively, (Equation 3.14) of all the repeated dinucleotides Di^+ (Section 2.3.2), all the repeated trinucleotides Tri^+ (Section 2.3.2) and all the repeated tetranucleotides $Tetra^+$ (Section 2.3.2), respectively, in the large eukaryotic genomes $\mathcal{G}$. The means $\overline{r}(Di^+)$, $\overline{r}(Tri^+)$ and $\overline{r}(Tetra^+)$ (Equation 3.17) and the medians $\tilde{r}(Di^+)$, $\tilde{r}(Tri^+)$ and $\tilde{r}(Tetra^+)$ of the ratios $r(Di^+, \mathcal{G})$, $r(Tri^+, \mathcal{G})$ and $r(Tetra^+, \mathcal{G})$ in the genomes of eukaryotes $\mathbb{E}$ lead to Equation 4.1.

Genome $\mathcal{G}$	$r(Di^+, \mathcal{G})$	$r(Tri^+, \mathcal{G})$	$r(Tetra^+, \mathcal{G})$	*Genome* $\mathcal{G}$	$r(Di^+, \mathcal{G})$	$r(Tri^+, \mathcal{G})$	$r(Tetra^+, \mathcal{G})$
Anolis carolinensis	0.814	2.602	0.56	*Microtus ochrogaster*	3.528	0.319	1.483
Beta vulgaris	0.827	0.668	0.05	*Monodelphis domestica*	2.447	0.256	2.066
Bos taurus	0.45	0.022	0.008	*Mus musculus*	5.061	0.812	2.51
Brassica napus	0.332	0.072	0.013	*Nomascus leucogenys*	0.541	0.066	0.494
Brassica oleracea	0.459	0.093	0.011	*Oreochromis niloticus*	1.684	0.125	0.292
Callithrix jacchus	0.791	0.033	0.391	*Ornithorhynchus anatinus*	0.223	0.09	0.262
Camelina sativa	0.707	0.202	0.009	*Oryctolagus cuniculus*	1.086	0.044	0.378
Canis lupus familiaris	1.119	0.207	1.734	*Oryza sativa Japonica*	0.859	0.133	0.069
Capra hircus	0.458	0.041	0.027	*Oryzias latipes*	0.188	0.061	0.827
Chlorocebus sabaeus	0.524	0.106	0.87	*Ovis aries*	0.506	0.06	0.033
Chrysemys picta bellii	0.565	0.007	0.091	*Pan paniscus*	0.503	0.07	0.355
Cicer arietinum	0.948	1.285	0.17	*Pan troglodytes*	0.501	0.072	0.371
Cynoglossus semilaevis	1.297	0.613	0.52	*Papio anubis*	0.462	0.082	0.774
Danio rerio	8.289	0.987	3.884	*Phaseolus vulgaris*	0.571	0.134	0.005
Elaeis guineensis	0.88	0.096	0.054	*Poecilia reticulata*	1.296	0.413	1.182
Equus caballus	0.18	0.012	0.057	*Pongo abelii*	0.432	0.068	0.32
Esox lucius	1.028	0.022	0.056	*Populus trichocarpa*	1.289	0.21	0.018
Felis catus	2.624	0.152	1.027	*Rattus norvegicus*	5.972	0.57	1.423
Ficedula albicollis	0.201	0.389	1.783	*Salmo salar*	6.069	0.089	1.268
Gallus gallus	0.07	0.029	0.323	*Setaria italica*	0.225	0.033	0.022
Glycine max	2.034	0.26	0.013	*Solanum lycopersicum*	0.878	0.176	0.034
Gorilla gorilla gorilla	0.448	0.055	0.315	*Solanum pennellii*	0.634	0.331	0.135
Gossypium raimondii	0.24	0.199	0.044	*Sorghum bicolor*	0.577	0.204	0.058
Homo sapiens	0.713	0.086	0.502	*Sus scrofa*	0.779	0.048	0.455
Lepisosteus oculatus	0.051	0.325	0.028	*Taeniopygia guttata*	0.124	0.079	0.214
Macaca fascicularis	0.612	0.109	0.979	*Theobroma cacao*	0.562	0.048	0.011
Macaca mulatta	0.595	0.093	0.778	*Vigna radiata*	3.389	0.251	0.023
Malus domestica	0.9	0.056	0.023	*Vitis vinifera*	0.987	0.284	0.035
Medicago truncatula	1.206	0.156	0.014	*Zea mays*	0.17	0.038	0.008
Meleagris gallopavo	0.088	0.035	0.162	**Means** $\overline{r}$	1.203	0.24	0.502
				Medians $\tilde{r}$	0.634	0.096	0.214

Table 4.34 shows the ratios $r(Di^+, \mathcal{G})$, $r(Tri^+, \mathcal{G})$ and $r(Tetra^+, \mathcal{G})$ giving the proportion of the total base numbers $B(Di^+, \mathcal{G})$, $B(Tri^+, \mathcal{G})$ and $B(Tetra^+, \mathcal{G})$ of all the repeated dinucleotides Di^+ (Section 2.3.2), all the repeated trinucleotides Tri^+ (Section 2.3.2) and all the repeated tetranucleotides $Tetra^+$ (Section 2.3.2) in the 59 large eukaryotic genomes $\mathcal{G}$ (sizes $N(\mathcal{G}) > 300000$ kb). Interestingly, the means $\overline{r}(D^+)$, $\overline{r}(Tri^+)$ and $\overline{r}(Tetra^+)$ (Equation 3.17) and the medians $\tilde{r}(Di^+)$, $\tilde{r}(Tri^+)$ and $\tilde{r}(Tetra^+)$ of the ratios $r(Di^+, \mathcal{G})$, $r(Tri^+, \mathcal{G})$ and $r(Tetra^+, \mathcal{G})$, re-

spectively, in the genomes of eukaryotes $\mathbb{E}$ leads both to the same following result

$$\begin{cases} \overline{r}(Di^{+}) > \overline{r}(Tetra^{+}) > \overline{r}(Tri^{+}) \\ \tilde{r}(Di^{+}) > \tilde{r}(Tetra^{+}) > \tilde{r}(Tri^{+}) \end{cases} \tag{4.1}$$

These inequalities are evaluated by two statistical tests: a paired sample Student's t-test (parametric statistical hypothesis test assuming a normal distribution of the population) and a Wilcoxon signed-rank W-test (non-parametric statistical hypothesis test). The comparisons of the means $\overline{r}(Di^{+})$ and $\overline{r}(Tetra^{+})$, and the means $\overline{r}(Tetra^{+})$ and $\overline{r}(Tri^{+})$ with the t-test have significant p-values equal to 3×10^{-5} and 7×10^{-3}, respectively. The comparisons of the distribution of Di^{+} and $Tetra+$, and the distribution of $Tetra^{+}$ and Tri^{+} with the Wilcoxon test also have significant p-values equal to 10^{-6} and 9×10^{-3}, respectively. Thus, the total base number of Di^{+} is greater than the total base number of $Tetra^{+}$ which is greater than the total base number of Tri^{+} . In other words, there is a scarcity of repeated trinucleotides in the large eukaryotic genomes compared to the repeated dinucleotides and the repeated tetranucleotides. For the eukaryotic genomes $\mathcal{G}$ of small sizes $N(\mathcal{G}) < 300000$ kb, the analysis has the same statistical trend. However, it is not conclusive and should be investigated in the future with the increase of genome data.

4.5 Identical trinucleotide pairs of the X circular code in eukaryotic gene sequences

Unitary circular codes (UCC) of dinucleotides, trinucleotides and tetranucleotides are associated with the repeated dinucleotides (Di^{+} motifs), the repeated trinucleotides (t^{+} motifs) and the repeated tetranucleotides ($Tetra^{+}$ motifs) which are identified in the genomes of eukaryotes. Furthermore, there is a scarcity of t^{+} motifs in the large genomes of eukaryotes compared to the Di^{+} and $Tetra^{+}$ motifs (Section 4.4.5). Otherwise, a circular code X is observed in genes of bacteria, eukaryotes, plasmids and viruses (Arquès and Michel, 1996; Michel, 2015)). The problem investigated here is whether the unitary circular codes of trinucleotides in genomes may have some traces in the trinucleotide circular code X in genes.

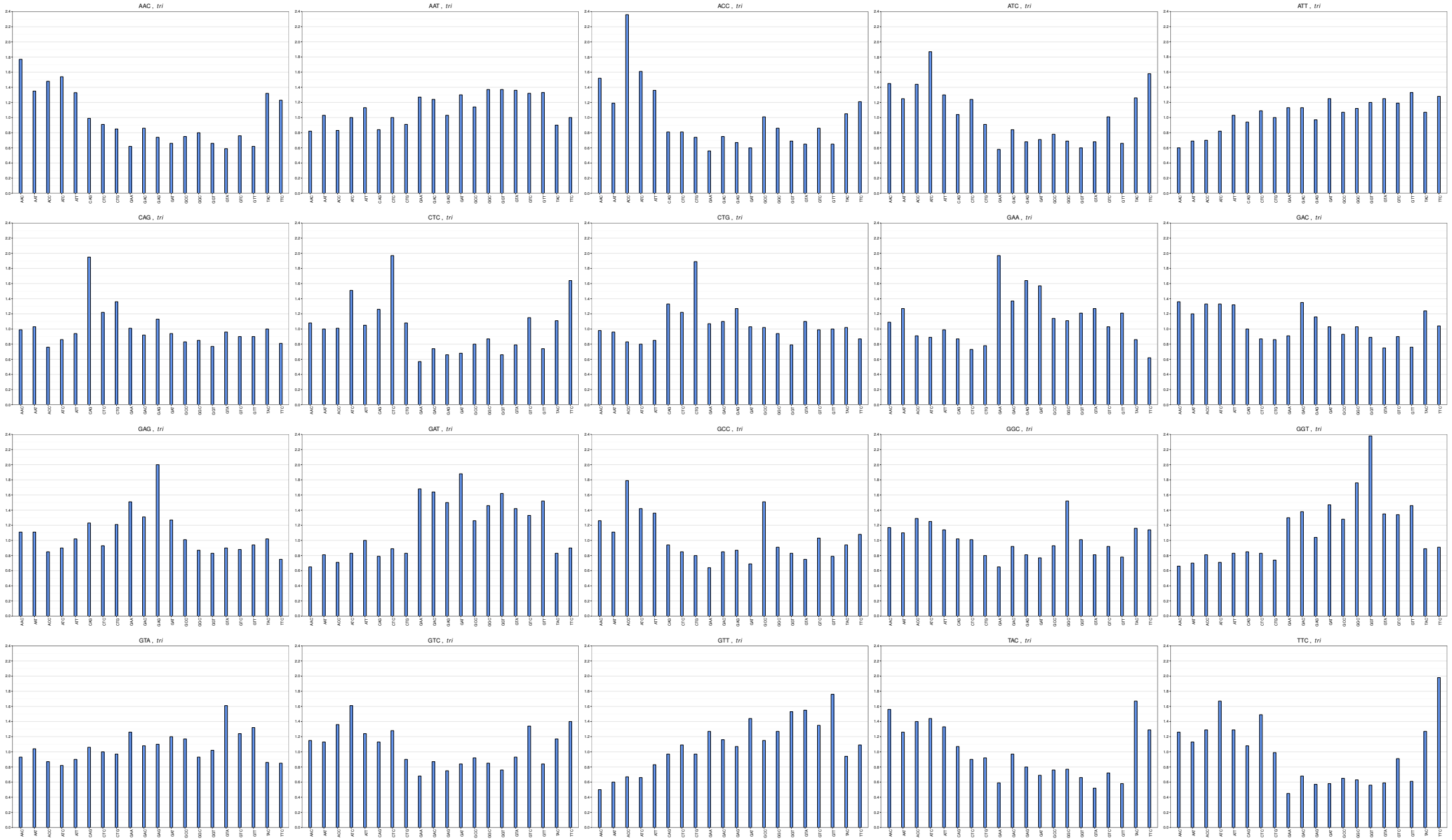

Figure 4.25: Identical trinucleotide pairs of the maximal C^3 self-complementary trinucleotide circular code X preferentially used in the gene sequences of eukaryotes identified by the median $\tilde{r}(t_1t_2)$ (Equation 3.27) of the observed/theoretical ratios $r(t_1t_2, \mathcal{G})$ (Equation 3.26) of the trinucleotide pairs $t_1t_2 \in X^2$. Each figure gives in ordinate the median $\tilde{r}(t_1t_2)$ of a given trinucleotide $t_1 \in X$ (in label) by varying the 20 trinucleotides $t_2 \in X$ in abscissa.

Figure 4.25 identifies a new property of the circular code X. Indeed, by varying the 20 trinucleotides $t_2 \in X$ for a given trinucleotide $t_1 \in X$, the medians $\tilde{r}(t_1t_2)$ (Equation 3.27) of the observed/theoretical ratios $r(t_1t_2, \mathcal{G})$ (Equation 3.26) of the trinucleotide pairs $t_1t_2 \in X^2$ identify 14 trinucleotide pairs such that the values of $\tilde{r}(t_1t_2)$ are maximal when the trinucleotide $t_2 = t_1$. These 14 trinucleotide pairs tt with an identical trinucleotide t are described according to t as follows:

$$\begin{aligned} t \in X' = \{&AAC, ACC, ATC, CAG, CTC, CTG, GAA, \\ &GAG, GAT, GGT, GTA, GTT, TAC, TTC\} \end{aligned} \tag{4.2}$$

where X' is a subset of X. There $\tilde{r}(tt)$ values are significantly greater than 1 (see Remark 2): $\tilde{r}(AACAAC) = 1.70$, $\tilde{r}(ACCACC) = 2.26$, $\tilde{r}(ATCATC) = 1.93$, $\tilde{r}(CAGCAG) = 1.79$, $\tilde{r}(CTCCTC) = 1.90$, $\tilde{r}(CTGCTG) = 1.71$, $\tilde{r}(GAAGAA) = 2.03$, $\tilde{r}(GAGGAG) = 1.98$, $\tilde{r}(GATGAT) = 1.93$, $\tilde{r}(GGTGGT) = 2.25$, $\tilde{r}(GTAGTA) = 1.57$, $\tilde{r}(GTTGTT) = 1.69$, $\tilde{r}(TACTAC) = 1.65$ and $\tilde{r}(TTCTTC) = 1.99$. The trinucleotide pair GGCGGC has the highest value ($\tilde{r}(GGCGGC) = 1.33$) still close to 1. The exceptions are the five trinucleotide pairs $t_1t_2 \in \{AATGTC, ATTGTT, GACATT, GCCACC, GTCATC\}$ with $\tilde{r}(t_1t_2)$ values close to 1 (except one case): $\tilde{r}(AATGTC) = 1.40$, $\tilde{r}(ATTGTT) = 1.32$, $\tilde{r}(GACATT) = 1.39$, $\tilde{r}(GCCACC) = 1.83$ and $\tilde{r}(GTCATC) = 1.59$.

Surprisingly, the trinucleotide set X' is also self-complementary, i.e. $X' = \mathcal{C}(X')$. All these results are retrieved with the two other ratios $r(t_1t_2)$ (Equation 3.26) and $\overline{r}(t_1t_2)$ (Equation 3.27) (results not shown). Thus, with a few exceptions related to values of $\tilde{r}(t_1t_2)$, $r(t_1t_2)$ and $\overline{r}(t_1t_2)$ close to 1, identical trinucleotide pairs of the circular code X are preferentially used in the genes of eukaryotes.

4.6 Summary

New properties of the circular code theory is identified here. We were able to show strong evidence in favour of the X circular code. The first was the presence of the conserved G530, A1492 and A1493 nucleotides of the ribosomal rRNA in X circular code motifs. These nucleotides were proven to have a function in the translation process by distinguishing cognate from non-cognate tRNAs by anticodon-codon interactions with the mRNA codon (Wilson, 2014). Our study in the ribosome was conducted in an expansive manner, we started in the decoding center, then we moved to conserved X circular code motifs near the decoding center. Afterwards, we showed an exhaustive study of X circular code motifs in tRNA sequences for which we presented a coverage of X motifs in tRNA sequences.

The study on the eukaryotic genomes showed us that the X_0 circular code is preferential when compared to X_1 and X_2, 23 bijective circular code transformations and 30 random generated codes by analysing the occurrence of the large motifs retrieved from the mentioned codes. The results presented distinguish the X circular code. These studies were done on the genomes of 138 organisms with complete chromosomes, while giving a detailed study for the *Homo sapiens* genome. We also saw the preferential presence of X circular code motifs in coding regions when compared to non-coding regions in eukaryotic genomes.

We were able to find interesting facts about the scarcity of trinucleotide simple repeats when compared to its dinucleotide and tetranucleotide counter parts. While showing there is no significant correlation between a repeated trinucleotide and the size of genomes as well as the count of A, C, G, T and GC content of genomes. We found that there is a preference for identical trinucleotide pairs in the gene sequences of eukaryotic genomes when it comes to X circular code trinucleotides.

In the next chapter we will conclude our work with a summary of our finding and some hypothesises that can be drawn from these results while also proposing some theories and questions raised.

5

Conclusion

The results here obtained in this thesis bring several new contributions to the circular code theory. This is the first time that X circular code motifs are studied in a biological context. Thus, these motifs have the circular code property (Definition 2.5), allowing retrieval of the reading frame, the property C^3 (Definition 2.10) allowing retrieval of the two shifted frames and the complementary property (Definition 2.9) allowing pairing between X circular code motifs. X circular code motifs were found in the sequences of ribosomes (mRNA, tRNA and rRNA) and in complete sequences of eukaryotic chromosomes. The results strengthen the concept of translation code based on the circular code proposed in Michel, 2012.

The circular code theory contributed to the analysis of the ribosome decoding center, in particular to its primary structure which is related to the mathematical property of circular code. The universally conserved nucleotides A1492 and A1493 in all studied rRNAs of bacteria, archaea, nuclear eukaryotes, and chloroplasts belongs to X motifs (m_{AA}, Table 4.1). The conserved nucleotide G530 in rRNAs of bacteria and archaea also belongs to X motifs (m_G, Table 4.2). The development of a tool (Section 3.4.1) associated with the global multiple sequence alignment allows to identify the X motifs m_G in nuclear and chloroplast rRNAs (Table 4.2) as it was not previously identified experimentally. Furthermore, it reveals a new X motif (m) which is universally conserved in the seven studied organisms (Table 4.3). Finally, the three X motifs m_{AA}, m_G, and m belong to the ribosome decoding center in all studied rRNAs of bacteria, archaea, nuclear eukaryotes, and chloroplasts (Figures 4.1, 4.2, 4.3, 4.4, 4.5, 4.6 and 4.7).

Several biological considerations can be stressed from these results. The function of the ribosome decoding center which has been attributed to a very few nucleotides only, precisely the nucleotides A1492, A1493 and G530, may

be related to motifs containing at least two successive trinucleotides up to a maximum of five successive trinucleotides (Tables 4.1 and 4.2).

Near the ribosome decoding center seven X circular code motifs $PrRNAXm$ which are conserved in 16S rRNAs of bacteria and archaea are identified (Figures 4.8, 4.9 and 4.10), in particular the large X motif $PrRNAXm_1$ $GC, GGT, AAT, ACC, GGC, GGC, C$ of 18 nucleotides in *P. furiosus*, the large common X motif $PrRNAXm_3$ $G, AAY, R_1CC, GR_2T, GGC, GAA, GGC$ of 19 nucleotides in *E. coli* ($Y = T$, $R_1 = A$ and $R_2 = G$) and *T. thermophilus* ($Y = C$, $R_1 = G$ and $R_2 = A$), the large X motif $PrRNAXm_4$ $TA, GAT, ACC, CTG, GTA, GTC, CA$ of 19 nucleotides in *E. coli*, the large common X motif $PrRNAXm_6$ $T, TAC, GRC, CWG, GGC, KAC, AC$ of 18 nucleotides in *E. coli* ($R = A$, $W = A$ and $K = T$) and *T. thermophilus* ($R = G$, $W = T$ and $K = G$).

Four X circular code motifs $ErRNAXm$ which are conserved in 18S rRNAs of *S. cerevisiae*, *T. aestivum* and *H. sapiens* are found near the ribosome decoding center (Figures 4.11, 4.12, 4.13), in particular the large common X motif $ErRNAXm_2$ $G, NTC, GAA, GAY, GAT, CAG, AT$ of 18 nucleotides in *S. cerevisiae*, *T. aestivum* and *H. sapiens* and the large X motif $ErRNAXm_4$ $TC, TTC, AAC, GAG, GAA, TTC, CT$ of 19 nucleotides in *S. cerevisiae*.

These X circular code motifs from the rRNA could hold some function in the positioning of the tRNA during the translation process by interacting with the X motif from the tRNA sequence.

The final step of the ribosomal study was to expand it to include the tRNA sequences of *E. coli*, *T. thermophilus* and *P. furiosus* which showed several new features for the structure of tRNAs. The high coverage of X motifs in the 5' and 3' regions of these tRNAs (88% and 71%, respectively; with the exception of the 3' regions of tRNAs of Asp, Glu, His and SeC; Table 4.27) means that tRNAs may be constructed by a concatenation of X motifs. This hypothesis is strengthened by the fact that four very large X motifs of length greater than or equal to 20 nucleotides are found in tRNAs having in average 79 nucleotides (Table 4.27): $Ala - tRNAXm_1$ $GC, CTC, AAT, GGC, ATT, GAG, GAG, GTC, A$ of 24 nucleotides in *T. thermophilus*, $Glu - tRNAXm_2$ $TA, GAG, GCC, CAG, GAC, ACC, GCC, CT$ of 22 nucleotides in *E. coli*, $Ser - tRNAXm_3$ $TG, GTC, GAA, GGC, GGC, ACC, CTG, CT$ of 22 nucleotides in *T. thermophilus* and $Tyr - tRNAXm_2$ $T, GCC, GTC, ATC, GAC, TTC, GAA, GGT, T$ of 23 nucleotides in *E. coli*, and 14 large X motifs of lengths 16–19 nucleotides: $Arg - tRNAXm_1$, $Arg - tRNAXm_3$, $Arg - tRNAXm_4$, $Arg - tRNAXm_6$, $Asn - tRNAXm_3$, $Asp - tRNAXm_2$, $Ile - tRNAXm_2$, $Leu - tRNAXm_1$, $Ser - tRNAXm_2$, $Ser - tRNAXm_4$, $Ser - tRNAXm_5$, $Ser - tRNAXm_7$, $Thr - tRNAXm_4$ and $Val - tRNAXm_3$. Remember that X motifs of lengths equal to 9 nucleotides retrieve the reading frame with a probability of 99.9% and X motifs of lengths

greater than or equal to 12 nucleotides retrieve, by definition, the reading frame with a probability of 100% (Table 3 and Fig. 4 in (Michel, 2012)). We also note that the coverage and the length of the X motifs could be greater if substitutions in X motifs were considered.

New properties of this circular code theory are also identified here with statistical studies of X large motifs in genomes of eukaryotes. This study shines light on non-gene regions, that were not examined previously, as well as gene regions. It has also been proposed that the circular code X, which is associated with the regular RNA transcription, may use its bijective transformation codes for coding nucleotide exchanging RNA transcription particularly in mitochondria (Michel and Seligmann, 2014). The large X motifs (having lengths $l \geq 15$ trinucleotides and cardinalities (composition) $Card \geq 10$ trinucleotides, Equation 3.4) have the highest occurrence in genomes of eukaryotes compared to (i) its 23 large bijective motifs from the bijective transformation circular codes, (ii) its two large permuted motifs $m(X_1)$ and $m(X_2)$ from the permuted circular codes $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$, and (iii) large random motifs $m(R)$ from random codes R (Section 4.3.1). The largest X motifs identified in genomes are presented (Section 4.3.2 Table 4.28), e.g. an X motif in a non-gene region of the genome *Solanum pennellii* with a length of 155 trinucleotides (465 nucleotides) and an expectation $E = 10^{-71}$ (Equation 3.6), two X motifs in non-gene regions of the genome *Salmo salar* with lengths of 118 trinucleotides (354 nucleotides) and an expectation $E = 10^{-52}$, etc. Large X motifs are also found in the human genome (Section 4.3.4 and Table 4.29). The largest X motif occurs in a non-gene region of the human chromosome 13 with a length of 36 trinucleotides and an expectation $E = 10^{-11}$.

X motifs in non-gene regions of genomes are possibly evolutionary relics of primitive genes using the circular code for translation. However, the mean value $\overline{r}_{m(X)}(\mathcal{G})$ and the median value $\tilde{r}_{m(X)}(\mathcal{G})$ giving the proportion of X motifs (having lengths $l \geq 10$ trinucleotides and cardinalities $Card \geq 5$ trinucleotides in genes/non-genes of the 138 complete eukaryotic genomes $\mathcal{G}$ are close to 8 ($\overline{r}_{m(X)}(\mathcal{G}) = 9.3 \approx \tilde{r}_{m(X)}(\mathcal{G}) = 7.6 \approx 8$, Section 4.3.5 and Table 4.30). This factor of 8 is retrieved for the X motifs in genes/non-genes of the 24 human chromosomes $\mathcal{H}_{Chr}$ ($\overline{r}_{m(X)}(\mathcal{H}_{Chr}) = 8.1 \approx \tilde{r}_{m(X)}(\mathcal{H}_{Chr}) = 7.8 \approx 8$, Section 4.3.6 and Table 4.31).

Therefore, the X motifs is found in non-coding regions but occur preferentially in coding region. This property is true whatever the base content of genes in the genomes as there is no correlation between the base proportion of genes/non-genes in genomes and the base proportion of X motifs in genes/non-genes of genomes (Figures 4.14, 4.16). From a biological point of view, this property may be explained by the fact that mutations (substitution, insertion and deletion of nucleotides) are more frequent in non-gene regions compared to genes. Finally, the statistical analysis developed here is based on the search of exact X motifs. X motifs with a few mutations in genomes of eukaryotes

should also be investigated in future.

The origin of this trinucleotide circular code X in genes is an open problem since its discovery in 1996. We show in our work that the circular code concept in low-complexity DNA regions exists with the unitary circular codes (UCC) of dinucleotides, trinucleotides and tetranucleotides generating UCC motifs of repeated dinucleotides (Di^+ motifs), repeated trinucleotides (Tri^+ motifs) and repeated tetranucleotides ($Tetra^+$ motifs) in eukaryotic genomes. Precisely, 12 UCC codes of dinucleotides are all strong comma-free, and four of them $\{AT\}$, $\{CG\}$, $\{GC\}$ and $\{TA\}$ are in addition self-complementary. 48 UCC codes of trinucleotides are strong comma-free and 12 UCC codes of trinucleotides are comma-free. 180 UCC codes of tetranucleotides are strong comma-free, 60 UCC codes of tetranucleotides are comma-free and 12 strong comma-free $\{AATT\}$, $\{ACGT\}$, $\{AGCT\}$, $\{CATG\}$, $\{CCGG\}$, $\{CTAG\}$, $\{GATC\}$, $\{GGCC\}$, $\{GTAC\}$, $\{TCGA\}$, $\{TGCA\}$ and $\{TTAA\}$ are in addition self-complementary. Thus, the Di^+, Tri^+ and $Tetra^+$ motifs allow to retrieve, to main and to synchronize a frame modulo 2, modulo 3 and modulo 4, respectively, in non-coding regions. Furthermore, the C^2, C^3 and C^4 properties allow to retrieve, maintain and synchronize the shifted frames and the self-complementary property allows DNA pairing in non coding-regions. An UCC motif and its complementary UCC motif have the same distribution in eukaryotic genomes, both from their occurrence number and their total base number. This property is observed with the Di^+, Tri^+ and $Tetra^+$ motifs. In addition for the Tri^+ and $Tetra^+$ motifs, an UCC motif and its complementary UCC motif have increasing occurrences conversely to their number of hydrogen bonds. For the Di^+ motifs, the repeat $(CG)^+$ has indeed the lowest occurrence but the repeat $(AT)^+$ does not have the highest occurrence. The largest nucleotide lengths of Di^+, Tri^+ and $Tetra^+$ motifs in the studied eukaryotic genomes are given in Table 4.32.

Surprisingly, a scarcity of repeated trinucleotides (Tri^+ motifs) in the large eukaryotic genomes is observed compared to the the Di^+ and $Tetra^+$ motifs. This statistical result is found with the mean and the median and confirmed by two statistical tests (a paired sample Student's t-test and a Wilcoxon signed-rank W-test). Thus, the unitary circular codes of trinucleotides associated to the repeated trinucleotides in eukaryotic genomes may have been involved in the formation of the trinucleotide circular code X in genes. This is mainly due to the fact that unitary circular code motifs are a circular code of cardinality equal to 1, one-point mutation would make this motif a circular code motif of cardinality. So on and so forth, we can easily have a circular code of cardinality 5 by having that same number as point mutations. A class of circular code motifs we found abundantly in the eukaryotic genomes.

A consequence of such an assumption would be the persistence of some statistical properties of repeated trinucleotides in the circular code X. Unexpectedly, identical trinucleotide pairs of the circular code X are preferentially used in the

eukaryotic genes. Indeed, 14 trinucleotides among 20 of the circular code X are preferentially followed by itself in the eukaryotic genes. This statistical result is observed with three ratios. Thus, the unitary circular codes of trinucleotides may have been involved in the formation of the trinucleotide circular code X. For the first time since 20 years, the circular code theory in genes is extended here to genomes. Circular code could be a mathematical structure of genes as well as genomes.

Publications in peer-reviewed journals

Karim El Soufi and Christian J. Michel (2014). "Circular code motifs in the ribosome decoding center". In: *Computational Biology and Chemistry* 52, pp. 9–17. ISSN: 1476-9271. URL: http://www.sciencedirect.com/science/article/pii/S1476927114000802

Karim El Soufi and Christian J. Michel (2015). "Circular code motifs near the ribosome decoding center". In: *Computational Biology and Chemistry* 59, Part A, pp. 158–176. ISSN: 1476-9271. URL: http://www.sciencedirect.com/science/article/pii/S1476927115300335

Karim El Soufi and Christian J. Michel (2016). "Circular code motifs in genomes of eukaryotes". In: *Journal of Theoretical Biology* 408, pp. 198–212. ISSN: 0022-5193. URL: http://www.sciencedirect.com/science/article/pii/S0022519316302053

Karim El Soufi and Christian J. Michel (2017). "Unitary circular code motifs in genomes of eukaryotes". In: *BioSystems*

Bibliography

Anger, Andreas M. et al. (2013). "Structures of the human and Drosophila 80S ribosome". In: *Nature* 497.7447, pp. 80–85. ISSN: 0028-0836. URL: http://dx.doi.org/10.1038/nature12104.

Armache, Jean-Paul et al. (2010a). "Cryo-EM structure and rRNA model of a translating eukaryotic 80S ribosome at 5.5-Å resolution". In: *Proceedings of the National Academy of Sciences* 107.46, pp. 19748–19753. URL: http://www.pnas.org/content/107/46/19748.abstract.

Armache, Jean-Paul et al. (2010b). "Localization of eukaryote-specific ribosomal proteins in a 5.5-Å cryo-EM map of the 80S eukaryotic ribosome". In: *Proceedings of the National Academy of Sciences* 107.46, pp. 19754–19759. URL: http://www.pnas.org/content/107/46/19754.abstract.

Armache, Jean-Paul et al. (2013). "Promiscuous behaviour of archaeal ribosomal proteins: Implications for eukaryotic ribosome evolution". In: *Nucleic Acids Research* 41.2, pp. 1284–1293. URL: http://nar.oxfordjournals.org/content/41/2/1284.abstract.

Arquès, D. G. and C. J. Michel (1996). "A complementary circular code in the protein coding genes." eng. In: *J Theor Biol* 182.1, pp. 45–58. DOI: 10.1006/jtbi.1996.0142. URL: http://dx.doi.org/10.1006/jtbi.1996.0142.

Brilot, Axel F. et al. (2013). "Structure of the ribosome with elongation factor G trapped in the pretranslocation state". In: *Proceedings of the National Academy of Sciences* 110.52, pp. 20994–20999. URL: http://www.pnas.org/content/110/52/20994.abstract.

Bulygin, Konstantin et al. (2009). "Sites of 18S rRNA contacting mRNA 3 and 5 of the P site codon in human ribosome: A cross-linking study with mRNAs carrying 4-thiouridines at specific positions". In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1789.3, pp. 167–174. ISSN: 1874-9399. URL: http://www.sciencedirect.com/science/article/pii/S1874939908002691.

Chenna, Ramu et al. (2003). "Multiple sequence alignment with the Clustal series of programs". In: *Nucleic Acids Research* 31.13, pp. 3497–3500. URL: http://nar.oxfordjournals.org/content/31/13/3497.abstract.

Crick, F H C, J S Griffith, and L E Orgel (1957). "Codes without commas". In: *Proceedings of the National Academy of Sciences of the United States of America* 43.5, pp. 416–421. ISSN: 1091-6490. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC528468/.

Demeshkina, Natalia et al. (2012). "A new understanding of the decoding principle on the ribosome". In: *Nature* 484.7393, pp. 256–259. ISSN: 0028-0836. URL: http://dx.doi.org/10.1038/nature10913.

El Soufi, Karim and Christian J. Michel (2016). "Circular code motifs in genomes of eukaryotes". In: *Journal of Theoretical Biology* 408, pp. 198–212. ISSN: 0022-5193. URL: http://www.sciencedirect.com/science/article/pii/S0022519316302053.

El Soufi, Karim and Christian J. Michel (2014). "Circular code motifs in the ribosome decoding center". In: *Computational Biology and Chemistry* 52, pp. 9–17. ISSN: 1476-9271. URL: http://www.sciencedirect.com/science/article/pii/S1476927114000802.

El Soufi, Karim and Christian J. Michel (2015). "Circular code motifs near the ribosome decoding center". In: *Computational Biology and Chemistry* 59, Part A, pp. 158–176. ISSN: 1476-9271. URL: http://www.sciencedirect.com/science/article/pii/S1476927115300335.

El Soufi, Karim and Christian J. Michel (2017). "Unitary circular code motifs in genomes of eukaryotes". In: *BioSystems*.

Fan-Minogue, Hua and David M Bedwell (2007). "Eukaryotic ribosomal RNA determinants of aminoglycoside resistance and their role in translational fidelity". In: *RNA* 14.1, pp. 148–157. ISSN: 1469-9001. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2151042/.

Fimmel, Elena, Alberto Danielli, and Lutz Strüngmann (2013). "On dichotomic classes and bijections of the genetic code". In: *Journal of Theoretical Biology* 336, pp. 221–230. ISSN: 0022-5193. URL: http://www.sciencedirect.com/science/article/pii/S0022519313003500.

Fimmel, Elena, Christian J. Michel, and Lutz Strüngmann (2016). "n-Nucleotide circular codes in graph theory". In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 374.2063. ISSN: 1364-503X. DOI: 10.1098/rsta.2015.0058. eprint: http://rsta.royalsocietypublishing.org/content/374/2063/20150058.full.pdf. URL: http://rsta.royalsocietypublishing.org/content/374/2063/20150058.

Fimmel, Elena et al. (2014). "Circular codes, symmetries and transformations". In: *Journal of Mathematical Biology* 70.7, pp. 1623–1644. ISSN: 1432-1416. DOI: 10.1007/s00285-014-0806-7. URL: http://dx.doi.org/10.1007/s00285-014-0806-7.

Gamow, G. (1954). "Possible Relation between Deoxyribonucleic Acid and Protein Structures". In: *Nature* 173.4398, pp. 318–318. URL: http://dx.doi.org/10.1038/173318a0.

Gogala, Marko et al. (2014). "Structures of the Sec61 complex engaged in nascent peptide translocation or membrane insertion". In: *Nature* 506.7486, pp. 107–110. ISSN: 0028-0836. URL: http://dx.doi.org/10.1038/nature12950.

Hanson, Robert (2010). "Jmol - a paradigm shift in crystallographic visualization". In: *J. Appl. Cryst.* 43.5, pp. 1250–1260. ISSN: 0021-8898. URL: https://doi.org/10.1107/S0021889810030256.

Herráez, Angel (2006). "Biomolecules in the computer: Jmol to the rescue". In: *Biochem. Mol. Biol. Educ.* 34.4, pp. 255–261. ISSN: 1539-3429. URL: http://dx.doi.org/10.1002/bmb.2006.494034042644.

Higgins, Desmond G, Julie D Thompson, and Toby J Gibson (1996). "Using CLUSTAL for multiple sequence alignments". In: *Computer Methods for Macromolecular Sequence Analysis*. Vol. Volume 266. Academic Press, pp. 383–402. URL: http://www.sciencedirect.com/science/article/pii/S0076687996660248.

Jeanmougin, F et al. (1998). "Multiple sequence alignment with Clustal X". In: *Trends in biochemical sciences* 23.10, pp. 403–405. ISSN: 0968-0004. URL: http://dx.doi.org/10.1016/S0968-0004(98)01285-7.

Jenner, Lasse B et al. (2010). "Structural aspects of messenger RNA reading frame maintenance by the ribosome". In: *Nat Struct Mol Biol* 17.5, pp. 555–560. ISSN: 1545-9993. URL: http://dx.doi.org/10.1038/nsmb.1790.

Larkin, M.A. et al. (2007). "Clustal W and Clustal X version 2.0". In: *Bioinformatics* 23.21, pp. 2947–2948. URL: http://bioinformatics.oxfordjournals.org/content/23/21/2947.abstract.

Michel, Christian J. (2014). "A genetic scale of reading frame coding". In: *Journal of Theoretical Biology* 355, pp. 83–94. ISSN: 0022-5193. URL: http://www.sciencedirect.com/science/article/pii/S0022519314001684.

Michel, Christian J. (2012). "Circular code motifs in transfer and 16S ribosomal RNAs: A possible translation code in genes". In: *Computational Biology and Chemistry* 37, pp. 24–37. ISSN: 1476-9271. URL: http://www.sciencedirect.com/science/article/pii/S147692711100096X.

Michel, Christian J. (2015). "The maximal C3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses". In: *Journal of Theoretical Biology* 380, pp. 156–177. ISSN: 0022-5193. URL: http://www.sciencedirect.com/science/article/pii/S002251931500171X.

Michel, Christian J. and Giuseppe Pirillo (2010). "Identification of all trinucleotide circular codes". In: *Computational Biology and Chemistry* 34.2, pp. 122–125. ISSN: 1476-9271. URL: http://www.sciencedirect.com/science/article/pii/S1476927110000204.

Michel, Christian J., Giuseppe Pirillo, and Mario A. Pirillo (2012). "A classification of 20-trinucleotide circular codes". In: *Information and Computation* 212, pp. 55 –63. ISSN: 0890-5401. DOI: http://dx.doi.org/10.1016/j.ic.2011.12.001. URL: http://www.sciencedirect.com/science/article/pii/S0890540111001702.

Michel, Christian J. and Hervé Seligmann (2014). "Bijective transformation circular codes and nucleotide exchanging RNA transcription". In: *Biosystems* 118, pp. 39–50. ISSN: 0303-2647. URL: http://www.sciencedirect.com/science/article/pii/S0303264714000215.

Moazed, D. and H. F. Noller (1990). "Binding of tRNA to the ribosomal A and P sites protects two distinct sets of nucleotides in 16 S rRNA." eng. In: *J Mol Biol* 211.1, pp. 135–145. DOI: 10.1016/0022-2836(90)90016-F. URL: http://dx.doi.org/10.1016/0022-2836(90)90016-F.

Nirenberg, Marshall W. and J. Heinrich Matthaei (1961). "The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides". In: *Proceedings of the National Academy of Sciences* 47.10, pp. 1588–1602. URL: http://www.pnas.org/content/47/10/1588.short.

Powers, T. and H. F. Noller (1994). "Selective perturbation of G530 of 16 S rRNA by translational miscoding agents and a streptomycin-dependence mutation in protein S12." eng. In: *J Mol Biol* 235.1, pp. 156–172.

Reuveni, Shlomi et al. (2011). "Genome-Scale Analysis of Translation Elongation with a Ribosome Flow Model". In: *PLoS Computational Biology* 7.9, e1002127–. ISSN: 1553-7358. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3164701/.

Seligmann, Hervé (2013a). "Polymerization of non-complementary RNA: Systematic symmetric nucleotide exchanges mainly involving uracil produce mitochondrial RNA transcripts coding for cryptic overlapping genes". In:

Biosystems 111.3, pp. 156–174. ISSN: 0303-2647. URL: http://www.sciencedirect.com/science/article/pii/S0303264713000269.

Seligmann, Hervé (2013b). "Systematic asymmetric nucleotide exchanges produce human mitochondrial RNAs cryptically encoding for overlapping protein coding genes". In: *Journal of Theoretical Biology* 324, pp. 1–20. ISSN: 0022-5193. URL: http://www.sciencedirect.com/science/article/pii/S0022519313000490.

Sharma, Manjuli R. et al. (2007). "Cryo-EM study of the spinach chloroplast ribosome reveals the structural and functional roles of plastid-specific ribosomal proteins". In: *Proceedings of the National Academy of Sciences* 104.49, pp. 19315–19320. URL: http://www.pnas.org/content/104/49/19315.abstract.

Thompson, JD et al. (1997). "The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools". In: *Nucleic acids research* 25.24, pp. 4876–4882. ISSN: 0305-1048. URL: http://dx.doi.org/10.1093/nar/25.24.4876.

Watson, J. D. and F. H. C. Crick (1953). "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid". In: *Nature* 171.4356, pp. 737–738. URL: http://dx.doi.org/10.1038/171737a0.

Willighagen, Egon L (2001). "Processing CML conventions in Java". In: *Internet Journal of Chemistry* 4.4, pp. 1099–8292.

Wilson, Daniel N. (2014). "Ribosome-targeting antibiotics and mechanisms of bacterial resistance". In: *Nat Rev Micro* 12.1, pp. 35–48. ISSN: 1740-1526. URL: http://dx.doi.org/10.1038/nrmicro3155.

Yoshizawa, Satoko, Dominique Fourmy, and Joseph D. Puglisi (1999). "Recognition of the Codon-Anticodon Helix by Ribosomal RNA". In: *Science* 285.5434, pp. 1722–1725. URL: http://science.sciencemag.org/content/285/5434/1722.abstract.

Karim EL SOUFI

Study of circular code motifs in nucleic acid sequences

Le travail effectué dans cette thèse présente une nouvelle approche de la théorie du code circulaire dans les gènes qui a été initiée en 1996. Cette approche consiste à analyser les motifs construits a partir de ce code circulaire, ces motifs particulier sont appelés motifs de code circulaire. Ainsi, nous avons développé des algorithmes de recherche pour localiser les motifs de code circulaire dans les séquences d'acides nucleiques afin de leur trouver une signification bioinformatique. En effet, le code circulaire X identifie dans les gènes est un ensemble de trinucleotides qui a la propriété de retrouver, synchroniser et maintenir la phase de lecture. Nous avons commence notre analyse avec le centre de décodage du ribosome (ARNr) qui est une région majeure dans le processus de traduction des gènes aux protéines. Puis, nous avons étendu les résultats obtenus avec le ribosome aux ARN de transfert (ARNt) pour étudier les interactions ARNr-ARNt. Enfin, nous avons généralisé la recherche de motifs de code circulaire X dans l'ADN aux chromosomes d'eucaryotes complets.

The work done in this thesis presents a new direction for circular code identified in 1996 by analysing the motifs constructed from circular code. These particular motifs are called circular code motifs. We applied search algorithms to locate circular code motifs in nucleic acid sequences in order to find biological significance. In fact, the circular code X, which was found in gene sequences, is a set of trinucleotides that have the property of reading frame retrieval, synchronization and maintenance. We started our study in the ribosomal decoding centre (rRNA), an important region involved in the process of translating genes into proteins. Afterwards, we expanded our scope to study the interaction of rRNA through the X circular code. Finally, we search for the X circular code motifs in the complete DNA sequences of chromosomes of the eukaryotic genomes. This study introduced new properties to the circular code theory.